РОССИЙСКАЯ АКАДЕМИЯ НАУК
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР ИМ. А. А. ДОРОДНИЦЫНА РАН
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ «ЛЭТИ»
при поддержке
РОССИЙСКОГО ФОНДА ФУНДАМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ
КОМПАНИИ FORECSYS

Математические методы распознавания образов

ММРО-13

Ленинградская область, г. Зеленогорск,
30 сентября — 6 октября 2007

Доклады 13-й Всероссийской конференции,
посвящённой 15-летию РФФИ

Москва, 2007

УДК 004.85+004.89+004.93+519.2+519.25+519.7
ББК 22.1:32.973.26-018.2
М34

М34 **Математические методы распознавания образов.** 13-я Всероссийская конференция: Сборник докладов. — М.: МАКС Пресс, 2007. 668 с.
ISBN 978-5-317-02060-6

В сборнике представлены доклады 13-й Всероссийской конференции «Математические методы распознавания образов», посвящённой 15-летию РФФИ, проводимой Вычислительным центром им. А. А. Дородницына Российской академии наук при финансовой и организационной поддержке РФФИ (грант № 07-07-06050) и компании Forecsys.

Конференция регулярно проводится один раз в два года, начиная с 1983 г., и является самым представительным российским научным форумом в области распознавания образов и анализа изображений, интеллектуального анализа данных, машинного обучения, обработки сигналов, математических методов прогнозирования.

УДК 004.85+004.89+004.93+519.2+519.25+519.7
ББК 22.1:32.973.26-018.2

ISBN 978-5-317-02060-6

Оргкомитет

Председатель: Журавлев Юрий Иванович, *академик РАН*
Зам. председателя: Матросов Виктор Леонидович, *чл.-корр. РАН*
Ученый секретарь: Воронцов Константин Вячеславович, *к.ф.-м.н.*
Члены: Граничин Олег Николаевич, *д.ф.-м.н.*
Донской Владимир Иосифович, *д.ф.м.н.*
Дедус Флоренц Федорович, *д.т.н.*
Немирко Анатолий Павлович, *д.ф.м.н.*
Устинин Михаил Николаевич, *д.ф.-м.н.*
Вальков Антон Сергеевич, *к.ф.-м.н.*
Инякин Андрей Сергеевич, *к.ф.-м.н.*
Песков Николай Владимирович, *к.ф.-м.н.*

Программный комитет

Председатель: Рудаков Константин Владимирович, *чл.-корр. РАН*
Зам. председателя: Дюкова Елена Всеволодовна, *д.ф.-м.н.*
Ученый секретарь: Чехович Юрий Викторович, *к.ф.-м.н.*
Члены: Микаэлян Андрей Леонович, *академик РАН*
Жижченко Алексей Борисович, *чл.-корр. РАН*
Сойфер Виктор Александрович, *чл.-корр. РАН*
Местецкий Леонид Моисеевич, *д.т.н.*
Моттль Вадим Вячеславович, *д.ф.м.н.*
Пытьев Юрий Петрович, *д.ф.м.н.*
Рязанов Владимир Васильевич, *д.ф.м.н.*
Рейер Иван Александрович, *к.т.н.*

Технический комитет

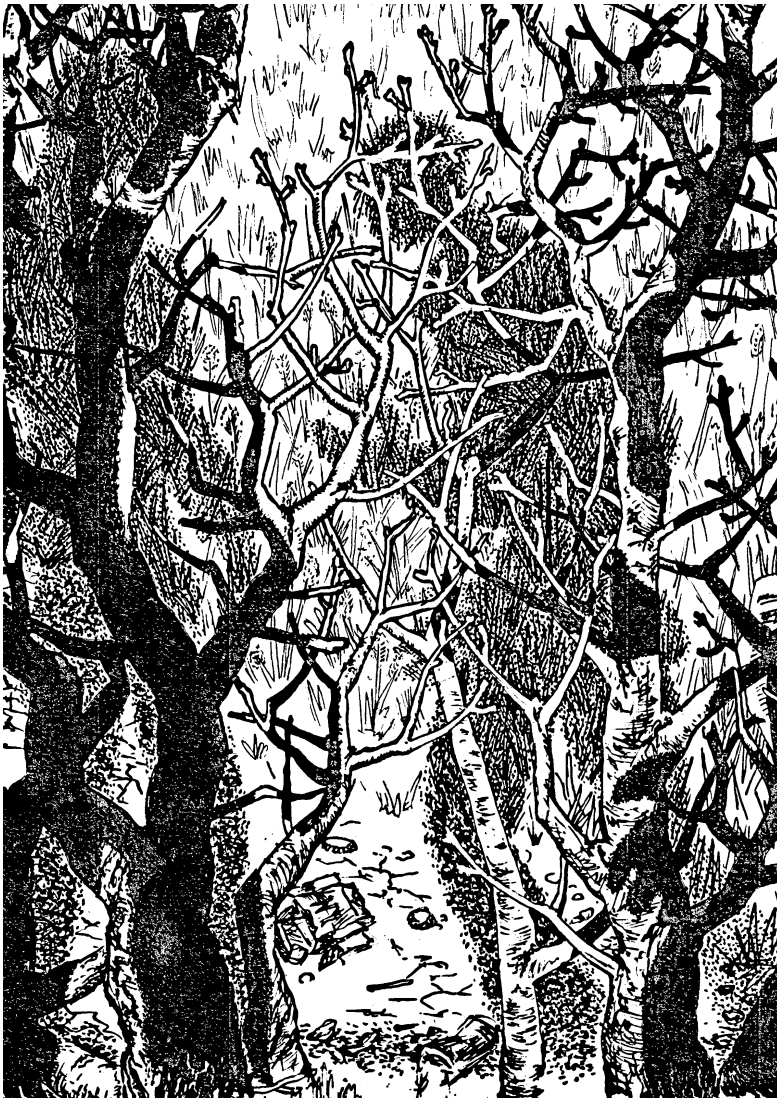
Председатель: Громов Андрей Николаевич
Члены: Гуз Иван Сергеевич
Ефимов Александр Николаевич
Ивахненко Андрей Александрович
Каневский Даниил Юрьевич
Лисица Андрей Валерьевич
Назарова Мария Николаевна
Никитов Глеб Владимирович
Пустовойтов Никита Юрьевич

Краткое оглавление

Фундаментальные основы распознавания и прогнозирования

Код раздела: TF (Theory and Fundamentals)

- Статистические основы обучения по прецедентам.
- Дискретно-логические основы обучения по прецедентам.
- Алгебраический подход к проблеме распознавания.
- Проблема обобщающей способности.
- Теория возможности и неопределённые нечёткие модели.
- Устойчивость обучения.
- Байесовский вывод.
- Теория многокритериального выбора в задачах принятия решений.
- Теоретические проблемы распознавания и прогнозирования.



О неморсовости гауссовой смеси

Апраушева Н. Н., Сорокин С. В.
plat@ccas.ru, www2007@ccas.ru
Москва, Вычислительный центр РАН

Широкое использование гауссовых смесей при решении задач классификации вызывает необходимость определения числа их мод, на чём основан модальный анализ [1, 2]. При этом в некоторых публикациях, например [2], среди свойств гауссовой смеси отмечается их морсовость. Но наши исследования функции плотности вероятности двухкомпонентной гауссовой смеси показали, что она имеет вырожденные критические точки (ВКТ). Получено уравнение геометрического места ВКТ, которое является уравнением границы областей её унимодальности и бимодальности.

Исследовалась смесь нормальных распределений с плотностью вероятности

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \sum_{i=1}^{k} \pi_i f_i(x), \quad f_i(x) = \exp\left(-\frac{1}{2\sigma^2}\,(x-\mu_i)^2\right),$$

где $x \in R,\ k = 2,\ \sigma^2$ — дисперсия компонент смеси, μ_i — математическое ожидание i-й компоненты, π_i — её априорная вероятность, $\pi_i \in (0,\ 1)$, $\sum_{i=1}^{k} \pi_i = 1$.

Гладкая функция является функцией Морса, если все её критические точки (КТ) невырождены [3].

Вырожденные критические точки функции $f(x)$ являются решениями системы уравнений $f'_x(x) = 0$, $f''_{xx}(x) = 0$. Если функция $f(x)$ имеет вырожденную критическую точку, то при небольшом изменении параметров распределения наблюдается неустойчивость, т. е. меняется число её критических точек [3]. Поиск ВКТ функции $f(x)$ проводился на основе результатов [4]. Для определённости положим $\mu_1 < \mu_2$.

Теорема 1. *При $k = 2$ и $\rho \leqslant 2$ функция $f(x)$ унимодальна, где ρ — расстояние Махаланобиса, $\rho = (\mu_2 - \mu_1)\sigma^{-1}$.*

Теорема 2. *При $k = 2$ и $\rho > 2,\ \pi_1 = \pi_2$ функция $f(x)$ бимодальна и точка $x_c = \frac{1}{2}(\mu_1 + \mu_2)$ является точкой её минимума.*

Из теорем 1 и 2 имеем: для $k = 2$ и $\rho = 2,\ \pi_1 = \pi_2$ точка $x_{c_1} = \frac{1}{2}(\mu_1 + \mu_2)$ является вырожденной модой функции $f(x)$, что проверяется подстановкой значений параметров $\pi_1 = \pi_2,\ \mu_2 = \mu_1 + 2\sigma,\ x_{c_1} = \mu_1 + \sigma$ в выражения $f'_x(x)$ и $f''_{xx}(x)$. Вырожденная критическая точка минимума смеси x_{c_2} была найдена для $k = 3,\ \mu_1 = -2.446,\ \mu_2 = 0,\ \mu_3 = 2.446,\ \pi_1 = 0.4,\ \pi_2 = 0.2,\ x_{c_2} = 0$.

Теорема 3. *При $k = 2$ и $\rho > 2$, $\pi_1 \neq \pi_2$ функция $f(x)$ унимодальна, если*

$$\left|\ln(\pi_1\pi_2^{-1})\right| \geqslant \tfrac{1}{2}\rho + 2\ln\left(\tfrac{1}{2}\left(\rho + \sqrt{\rho^2 - 4}\right)\right). \tag{1}$$

Эксперименты показали, что при выполнении неравенства, противоположного неравенству (1), при фиксированном значении ρ и различных значениях π_1, π_2 для числа критических точек m функции $f(x)$ имеют место неравенства $1 \leqslant m \leqslant 3$. Если $m = 2$, то одна из КТ функции $f(x)$, точка перегиба, является вырожденной.

Для $k = 2$ при различных значениях параметров ρ, π_1, π_2 экспериментальным путём было найдено 15 критических точек перегиба, для каждой из которых вычислялись значения параметров ρ и $\psi_1(\rho) = |\ln(\pi_1\pi_2)^{-1}|$; при тех же значениях ρ вычислялись значения функции $\psi_2(\rho)$, представляющей собой правую часть неравенства (1), и функции $\psi = \psi_2 - \psi_1$. На Рис. 1 представлены графики функций ψ_1, ψ_2, ψ. Для получения 15 критических точек перегиба фиксировалось значение μ_1 и варьировались значения μ_2, π_1, π_2.

Аппроксимировав функцию ψ параболой, $\psi = a\rho^2 + b\rho + c$ и вычислив неизвестные коэффициенты a, b, c методом наименьших квадратов, получили уравнение регрессии

$$\tilde{\psi} = -0.327\rho^2 + 3.867\rho - 3.738, \tag{2}$$

средняя абсолютная погрешность $\Delta = 0.101$, средняя относительная погрешность $\delta = 0.024$, средне-квадратическое отклонение $\tilde{\sigma}_1 = 0.124$.

На основе введенных обозначений ψ_1, ψ_2, ψ и формул (1), (2) для функции $\tilde{\psi}_1$, $\tilde{\psi}_1 = \psi_2 - \tilde{\psi}$, имеем выражение

$$\left|\ln\left(\pi_1\pi_2^{-1}\right)\right| = 0.827\rho^2 + 2\ln[2^{-1}(\rho + \sqrt{\rho^2 - 4})] - 3.867\rho + 3.738, \tag{3}$$

уравнение (3) является аппроксимационным уравнением границы унимодальности и бимодальности функции $f(x)$.

Оценим доверительный коридор для функции $\psi(\rho)$, используя неравенство Чебышева

$$\mathsf{P}\left\{|\psi - \tilde{\psi}| \leqslant t\tilde{\sigma}_1\right\} \geqslant 1 - t^{-2},$$

при $t = 3$, $\tilde{\sigma}_1 = 0.124$ имеем

$$\mathsf{P}\left\{\tilde{\psi} - 0.372 < \psi < \tilde{\psi} + 0.372\right\} > 0.888. \tag{4}$$

Поскольку $\tilde{\psi}_1 = \psi_2 - \tilde{\psi}$, то для искомой функции ψ_1 на основании (3), (4) имеем

$$\mathsf{P}\left\{d + 3.366 < \psi_1 < d + 4.110\right\} > 0.888. \tag{5}$$

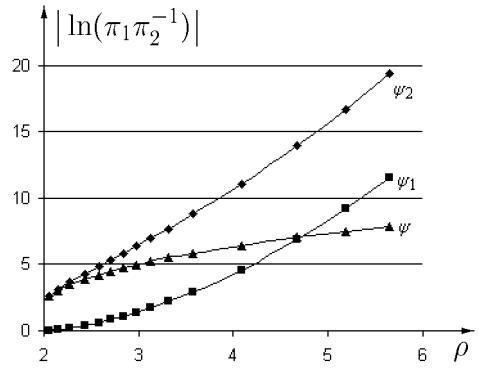


Рис. 1. Графики функций $\psi_1(\rho)$, $\psi_2(\rho)$ и $\psi(\rho)$.

Выражение (5) с вероятностью более 0.888 даёт доверительный коридор для всех ВКТ перегиба функции $f(x)$, в этом коридоре функция $f(x)$ может быть как унимодальной, так и бимодальной.

Таким образом, при $\rho > 2$ и $\pi_1 \neq \pi_2$ с вероятностью более 0.888 функция $f(x)$ унимодальна, если выполняется неравенство

$$\left|\ln(\pi_1\pi_2^{-1})\right| > d + 4.110,$$

и бимодальна, если выполняется неравенство

$$\left|\ln(\pi_1\pi_2^{-1})\right| < d + 3.366.$$

Итак, для исследуемой гауссовой смеси получены достаточные условия её унимодальности и бимодальности.

Литература

[1] *Дюран Б., Оделл П.* Кластерный анализ. — М.: Статистика, 1975.

[2] *Carreira-Perpiñán M.A., Williams C.* On the Number of Modes of a Gaussian Mixture. Inform. Res. Report EDI-INF-RR-0159. School of Inf. Univ. of Edinburg, 2003.

[3] *Арнольд В. И., Варченко А. Н., Гусейн-Заде С. М.* Особенности дифференцируемых отображений. — М.: Наука, 1982. — 304 с.

[4] *Апраушева Н. Н., Сорокин С. В.* Об унимодальности простейшей гауссовой смеси // ЖВМиМФ, 2004. — Т. 44, № 5. — С. 838–846.

Конструктная компьютеризация силлогистики

Брусенцов Н. П., Владимирова Ю. С.
ramil@cs.msu.su, julia_vladi@mtu-net.ru
Москва, МГУ им. М. В. Ломоносова, факультет ВМиК

Аристотелева силлогистика непарадоксальна и не вписывается в исчисления «классической» логики, потому что в основе ее лежит принцип сосуществования противоположностей [1, с. 87–92], предотвращающий возникновение химер, в частности, именуемых парадоксами и провозглашаемых законами, но не соответствующих реальности, подобно положенному в основу формальной логики хризиппову закону исключенного третьего, отвергающего сосуществование противоположностей и заблокировавшего развитие диалектической логики Аристотеля.

Воссозданная на основе принципа сосуществования силлогистика — вовсе не «узкая система, неприменимая ко всем видам рассуждений», как это «установлено» Яном Лукасевичем, а наоборот, совершенно безупречная, адекватная, диалектическая логика, однако трёхзначная, и потому в двухзначных исчислениях не отобразимая. Булева алгебра допускает трёхзначность только в элементарных конъюнкциях и дизъюнкциях. Так, конъюнкции xyz, xyz', xy различаются тем, что термин z первой присущ, второй антиприсущ, а в третьей присущность его несущественна: $xy(z \vee z') \equiv xy$, в ней z умалчивается. К сожалению, для членов ДНФ и КНФ подобное не предусмотрено: умалчивание означает исключенность члена, а несущественность неотобразима, третье невозможно. Естественней (и единообразней) сохранить принятое в элементарных выражениях — умалчивать несущественное и ввести функтор исключения, допустим, «минус». Теперь непарадоксальную импликацию (полноценное следование) $x \Rightarrow y$ можно выразить трехчленом $xy \vee -xy' \vee x'y'$, тогда как материальная импликация $x \to y$ будет: $xy \vee -xy' \vee x'y \vee x'y'$. Это экстенциональный (объемный) вариант подчинения логики высказываний принципу сосуществования противоположностей.

Суждения силлогистики истолковываются интенсинально — в них речь не о высказываниях и предикатах, а о существовании или несуществовании «вещей», охарактеризованных совокупностями их существенных особенностей. Выражающая отношение следования $x \Rightarrow y$ общеутвердительная посылка $\mathsf{A}xy$ («Все x суть y») представима конъюнкцией трех дизъюнктов:

$$\mathsf{A}xy \equiv \mathsf{V}xy\,\mathsf{V}'xy'\,\mathsf{V}x'y',$$

где $\mathsf{V}xy$ — существование xy-вещей, $\mathsf{V}'xy'$ — несуществование xy'-вещей, $\mathsf{V}x'y'$ — существование $x'y'$-вещей. Умалчивание существования или

несуществования xy-вещей означает несуществовенность его для представленного отношения.

В минимальной форме $\mathsf{A}xy \equiv \mathsf{V}x\,\mathsf{V}'xy'\,\mathsf{V}y'$, а в универсуме Аристотеля УА [1, с. 91], необходимо подчиненном принципу сосуществования противоположностей $\mathsf{V}x\,\mathsf{V}x'\,\mathsf{V}y\,\mathsf{V}y'$, т. е. предполагающем, что x-, x'-, y-, y'-вещи необходимо существуют, $\mathsf{V}'xy'$ истолковывается как $\mathsf{V}x\,\mathsf{V}'xy'\,\mathsf{V}y'$ — несовместимость x с y', что равносильно $x \Rightarrow y$. Таким образом, дизъюнкт $\mathsf{V}'xy'$, означающий в модальной логике парадоксальную «строгую импликацию» Льюиса [3], в универсуме Аристотеля обретает смысл полноценного содержательного следования.

Общеотрицательная посылка $\mathsf{E}xy$ «Все x суть y» получается из $\mathsf{A}xy$ инверсией y: $\mathsf{E}xy \equiv \mathsf{A}xy' \equiv \mathsf{V}'xy\,\mathsf{V}x\,\mathsf{V}y$. Частноутвердительная посылка $\mathrm{I}xy$ «Некоторые x суть y», «Существуют xy», несовместимая с общеотрицательной, в УА равнозначна ее инверсии: $\mathrm{I}xy \equiv \mathrm{inv}(\mathsf{V}'xy) \equiv \mathsf{V}xy$, что, с учетом $\mathsf{V}x\,\mathsf{V}x'\,\mathsf{V}y\,\mathsf{V}y'$, означает $\mathrm{I}xy \equiv \mathsf{V}xy\,\mathsf{V}x'\,\mathsf{V}y'$. Соответственно, частноотрицательная посылка $\mathsf{O}xy$ как инверсия общеутвердительной $\mathsf{A}xy$ будет $\mathsf{O}xy \equiv \mathsf{V}xy'\,\mathsf{V}x'\,\mathsf{V}y \equiv \mathrm{I}xy'$. Таким образом, при использовании инверсии терминов в силлогистике достаточно двух функторов — A и I. При этом восполнимы упущенные традиционной теорией отношения: $\mathsf{A}x'y$, $\mathsf{A}x'y'$, $\mathrm{I}x'y$, $\mathrm{I}x'y'$.

Наша цель — компьютеризация восполненной и упорядоченной посредством принципа сосуществования противоположностей аристотелевой силлогистики путём конструктного кодирования [4] её суждений и программной реализации умозаключений (модусов). Предполагается принятое в УА истолкование выражений.

Вывод из пары общих посылок общего заключения реализуется склеиванием их (элиминацией среднего термина). Например, модус Barbara:

$$\begin{aligned}\mathsf{A}xy\,\mathsf{A}yz \equiv \mathsf{V}'xy'\,\mathsf{V}'yz' \equiv \mathsf{V}'(xy' \vee yz') \equiv \mathsf{V}'(xy'z \vee xy'z' \vee xyz' \vee x'yz') \equiv \\ \equiv \mathsf{V}'(xy' \vee xz' \vee yz') \equiv \mathsf{V}'xy'\,\mathsf{V}'xz'\,\mathsf{V}'yz' \Rightarrow \mathsf{V}'xz' \equiv \mathsf{A}xz.\end{aligned}$$

В случае неосуществимости склеивания общего заключения нет, но из пары общих посылок непременно есть частное заключение, для получения которого вместо одной из общих употребляются подчиненные ей частные, с одной из которых заключение необходимо будет. Например, из $\mathsf{A}xy'\,\mathsf{A}yz$ общего заключения нет. Посылке $\mathsf{A}xy'$ подчинены $\mathrm{I}xy' \equiv \mathsf{V}xy'$ и $\mathrm{I}x'y \equiv \mathsf{V}x'y$.

$$\mathrm{I}xy'\,\mathsf{A}yz \equiv \mathsf{V}xy'\,\mathsf{V}'yz' \equiv \mathsf{V}xy'\big(\mathsf{V}'xyz' \vee \mathsf{V}'x'yz'\big) \text{ — нет заключения.}$$

$$\begin{aligned}\mathrm{I}x'y\,\mathsf{A}yz \equiv \mathsf{V}x'y\,\mathsf{V}'yz' \equiv \mathsf{V}x'y\big(\mathsf{V}'xyz' \vee \mathsf{V}'x'yz'\big) \Rightarrow \\ \Rightarrow \mathsf{V}x'y\,\mathsf{V}'x'yz' \Rightarrow \mathsf{V}x'yz \Rightarrow \mathsf{V}x'z \equiv \mathrm{I}x'z.\end{aligned}$$

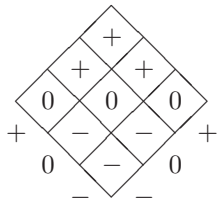


Рис. 1. Таблица операции $\oplus$.

Таким образом, из пары с общим (средним) термином, включающей общую и частную посылки, заключение возможно, но не необходимо. Из пары частных посылок заключение, как известно, невозможно.

Компьютеризация категорической силлогистики просто и экономно реализуется при помощи четырехтритных конструктов [4], кодирующих трехтерминные дизъюнкты существования и несуществования. Значение первого (головного) трита указывает тип дизъюнкта: «+» — существование, представляющее частную посылку, «−» — несуществование, общая посылка. Последующие триты сопоставлены терминам x, y, z, указывая их статусы. Например, $\mathsf{V}xy'z \equiv (++-+)$, $\mathsf{V}xy \equiv (+++0)$, $\mathsf{V}x'z \equiv (+-0+)$, $\mathsf{V}'xy' \equiv (-+-0)$, $\mathsf{V}'xy'z \equiv (-+-+)$, $\mathsf{V}'xz' \equiv (-+0-)$.

Общее заключение из пары общих посылок, если оно существует, достигается склеиванием (потритным «логическим сложением» $\oplus$, Рис. 1) соответствующих конструктов. Например,

$$\begin{aligned}\mathsf{A}xy'\,\mathsf{A}y'z' \equiv \mathsf{V}'xy\,\mathsf{V}'y'z \equiv (-++0) \oplus (-0-+) \equiv \\ \equiv (-+0+) \equiv \mathsf{V}'xz \equiv \mathsf{A}xz'.\end{aligned}$$

Частное заключение получается склеиванием конструкта, кодирующего частную посылку с инверсией представляющего общую. Например,

$$\begin{aligned}\mathrm{I}x'y'\,\mathsf{A}y'z' \equiv (+--0) \oplus \mathrm{inv}(-0-+) \equiv \\ \equiv (+--0) \oplus (+0+-) \Rightarrow (+-0-) \equiv \mathrm{I}x'z'.\end{aligned}$$

Если склеивания (элиминации среднего термина) нет, то и заключения не существует.

Интеллект реализованной в диалоговой системе структурированного программирования ДССП программы силлогистического вывода значительно превзошел то, что достигнуто «невооруженными» умами людей. Число правильных модусов, составляющее в традиционной логике 19, а в математической логике сокращенное до 15, оказалось равным 128.

Литература

[1] *Брусенцов Н. П.* Искусство достоверного рассуждения. — М.: Фонд «Новое тысячелетие». — 1998.

[2] *Брусенцов Н. П.* Трехзначная интерпретация силлогистики Аристотеля // Историко-математические исследования. Вторая серия. Вып. 8 (43). — М.: «Янус-К», 2003. — С. 317–327.

[3] *Слинин Н. И.* Современная модальная логика. — Л.: Изд-во Ленинградского ун-та, 1976. — С. 8.

[4] *Брусенцов Н. П., Владимирова Ю. С.* Троичная компьютеризация логики // ММРО-12. — М.: МАКС-Пресс, 2005. — С. 40–42.

Об ускорении процессов обучения и принятия решений

Вайнцвайг М. Н.

wainzwei@iitp.ru

Москва, Институт проблем передачи информации РАН

Мышление обычно рассматривается как адаптивный механизм организации поведения человека и высших животных, обеспечивающий им в широком классе изменений условий внешней среды возможность автономного существования и воспроизведения.

Трудности моделирования работы этого механизма связаны со следующими обстоятельствами.

Давно пришли к тому, что организация поведения проходит в рамках процесса постановки и достижения целей на основе хранящихся в памяти законов и правил, позволяющих посредством анализа, логического вывода и других преобразований информации, поступающей от органов чувств, принимать решения, т. е. находить неизвестные пути к целям, формировать действия, предсказывать изменения и пр.

Поскольку не все законы заранее известны, то необходимым этапом организации поведения становится обучение, т. е. основанный на наблюдениях, пробах, ошибках и обобщении поиск законов.

Следует отметить, что процедуры принятия решений и обучения по существу являются процедурами поиска экстремума (принятие решений — поиск оптимальных путей к целям, обучение — поиск оптимальных функций аппроксимации результатов наблюдения).

Известны два основных способа поиска экстремума: переборы и градиентный спуск, поэтому подходы к моделированию мыслительных процессов базировались на использование каждого их этих способов.

Первый подход основан на сложившемся в математике представлении об универсальном классе функций, с которыми сталкивается человек, как о классе, описываемом в терминах рекурсивных функций, машин

Тьюринга, продукций Поста, нормальных алгоритмов и пр., послуживших прототипами современных языков программирования. Поскольку для таких функций существует понятие универсальной функции, позволяющей организовать их перебор, то функции мышления пытались реализовать в рамках таких переборов.

На этой основе в мире возник широкий поток работ: «универсальный решатель проблем» GPS, распознающие системы, экспертные системы, игровые программы, системы доказательства теорем, решения задач и пр.

Вскоре, однако, выяснилось, что такие переборы в их непосредственном виде реализуемы лишь для достаточно простых и частных предметных областей. В общем же случае в силу NP-сложности переборы оказываются практически не реализуемыми. Это касается как процесса обучения, где объем перебора гипотез растет быстрее, чем экспоненциально с числом связываемых ими характеристик, так и логического вывода, где объем перебора также быстрее, чем экспоненциально растет с числом применимых на каждом шаге законов (аксиом) и правил вывода.

Поскольку заранее неизвестно, какие характеристики ситуаций должны связываться законами, то при обучении, как правило, приходится ориентироваться на максимально широкий их набор, из-за чего перебор гипотез становится практически нереализуем.

Альтернативой языково-переборного подхода является нейросетевой (или коннекционистский) подход, где и обучение, и принятие решений определяется сетью параллельно работающих функциональных элементов (нейронов), реализующих простые (как правило, непрерывные) функции. Конкретный вид этих функций определяется набором параметров (весов синапсов), значения которых формируются в процессе адаптации, т. е. непрерывной подстройки к нужным значениям выхода.

К этому направлению относятся персептрон, сети Хопфилда, ассоциативная память Кохонена, процедура backpropagation, адаптивные критики.

Казалось бы, за счет распараллеливания вычислений и использования только непрерывных функций здесь можно было бы получить большой выигрыш в скорости обработки. Однако, даже при относительно небольшой сложности нейронных сетей (размерности пространства синапсов), обучение моделей, как правило, оказывается слишком долгим, и требующим слишком большого числа показов примеров. Как и при первом подходе, оно здесь становится возможным лишь в достаточно простых случаях, когда сетью нейронов связывается относительно небольшое число характеристик. В сложных же случаях приходится разбивать

характеристики на небольшие группы (простые подзадачи), что, как правило, делается «вручную».

Таким образом, независимо от подхода, основные трудности моделирования мыслительных процессов связаны с проблемой NP-сложности и состоят в поиске методов автоматического сокращения: для обучения — числа связываемых гипотезами характеристик; для принятия решений — числа законов, применяемых на каждом шаге логического вывода.

Наша работа в основном и направлена на преодоление этих трудностей. Ее цель — построение возможно более полной и способной работать в реальном времени модели рекурсивного развития интеллекта [1–4], обеспечивающей возможность организации в реальном мире все более сложного поведения.

В основе модели лежит работа ассоциативной памяти с использованием внутреннего языка описания ситуаций. На множестве понятий языка — переменных определенных типов (объектных, числовых, логических) вводится метрика, где расстояния между понятиями определяется близостью соответствующих им событий в пространстве времени и последовательностью связывающих их законов. Переход из одной ситуации в другую определяется последовательностью событий — изменений значений понятий.

Процесс развития основывается на следующей рекурсивной схеме.

1. Наиболее близкие между собой в метрике понятий контрастные события, последовательно обращая на себя внимание, связываются гипотезами-отношениями, имеющими вид ассоциативных операторов, например, адаптивных сетей Кохонена или Хопфилда.
2. При последовательной адаптации гипотез строятся законы, позволяющие по известным значениям одних понятий находить неизвестные значения других понятий.
3. Соответствующие законам отношения становятся новыми понятиями, которые, в свою очередь, могут связываться гипотезами и законами.
4. Законы модифицируют метрику, уменьшая расстояние между понятиями, что открывает возможность построения новых гипотез.

Так происходит постепенное пополнение внутреннего языка новыми понятиями, гипотезами и законами. Метрика в пространстве понятий по существу оценивает сложность достижения одних ситуаций из других и используется не только при обучении, но и при поиске оптимальных путей к целям.

Литература

[1] *Вайнцвайг М. Н., Полякова М. П.* Формирование понятий и законов на основе анализа динамики зрительных картин // Труды 2-й международной

конференции «Проблемы управления и моделирования в сложных системах». Самара, 2000. — С. 166–170.

[2] *Вайнцвайг М. Н., Полякова М. П.* Архитектура и функции механизма мышления IEEE AIS'03, CAD-2003 (труды конференции) том.1, М.: Физматлит, 2003. — С. 208–213.

[3] *Вайнцвайг М. Н., Полякова М. П.* Архитектура системы представления зрительных динамических сцен // Математические методы распознавания образов, ММРО-11, Москва, 2003. — С. 261–263.

[4] *Вайнцвайг М. Н., Полякова М. П.* О моделировании мышления // От моделей поведения к искусственному интеллекту. — М.: УРСС, 2006. — С. 280–286.

Устойчивость обучения метода релевантных векторов

Васильев О. М., Ветров Д. П., Кропотов Д. А.
ovasiliev@inbox.ru, vetrovd@yandex.ru, dkropotov@yandex.ru
Москва, ВМиК МГУ, ВЦ РАН

Рассматривается проблема оценивания качества обучения метода релевантных векторов в классической постановке задачи классификации с двумя классами. Получена оценка отклонения эмпирического риска от риска на скользящем контроле для метода релевантных векторов.

Определения и инструментарий

В задаче классификации восстанавливаемая величина y принимает значения из множества ответов $Y = \{-1, 1\}$, а независимая величина $\mathbf{x}$ принимает значения из множества объектов $X = \mathbb{R}^n$. Задана прецедентная информация (выборка) $S = \{\mathbf{z}_1 = (\mathbf{x}_1, y_1), \ldots, \mathbf{z}_m = (\mathbf{x}_m, y_m)\}$, $(\mathbf{z}_1, \ldots, \mathbf{z}_m) \in Z^m = (X \times Y)^m$. Рассмотрим также выборки $S^i = S \setminus \{\mathbf{z}_i\}$, $i = 1, \ldots, m$, получаемые удалением из S одного наблюдения. *Алгоритмическим оператором* называется отображение из X в $\mathbb{R}$, выбранное из некоторого параметризованного семейства. В случаях, если необходимо подчеркнуть роль параметров $\mathbf{u}$ некоторого алгоритмического оператора g, будем использовать альтернативную запись $g = [\mathbf{u}]$. Алгоритмы классификации, рассматриваемые в работе, имеют вид $\mathrm{sign}(g)$, где g — некоторый алгоритмический оператор.

Методом обучения μ называется отображение, сопоставляющее произвольной выборке S' алгоритм классификации $\mu_{S'}$. Далее рассматривается единственный метод обучения, поэтому введём специальные обозначения для результатов его обучения (алгоритмов и соответствующих алгоритмических операторов) на выборках S и S^i, $i = 1, \ldots, m$. Будем обозначать $\mu_S = \mathrm{sign}(f) = \mathrm{sign}([\mathbf{w}])$ и $\mu_S^i = \mathrm{sign}(f^i) = \mathrm{sign}([\mathbf{w}^i])$ результат обучения метода μ на выборке S и S^i соответственно. Для задачи

классификации используем ценовую функцию $c\colon \mathbb{R}^2 \to \mathbb{R}$ вида

$$c(y,y') = \begin{cases} 1, & yy' \leqslant 0; \\ 1 - yy', & 0 \leqslant yy' \leqslant 1; \\ 0, & yy' \geqslant 1. \end{cases} \tag{1}$$

Эмпирическим риском алгоритма μ_S будем называть величину

$$R_{\mathrm{em}}(f) = \frac{1}{m}\sum_{i=1}^{m} c\big(f(\mathbf{x}_i), y_i\big),$$

Риском на скользящем контроле будем называть величину

$$R_{\mathrm{lo}}(f) = \frac{1}{m}\sum_{i=1}^{m} c\big(f^i(\mathbf{x}_i), y_i\big).$$

Определение 1. *Метод обучения для задачи классификации будем называть β-устойчивым в обучении, если для всех $S \in Z^m$, всех $(\mathbf{x}, y) \in S$ и всех $i = 1, \ldots, m$ выполняется условие $\big|[\mathbf{w}](\mathbf{x}) - [\mathbf{w}^i](\mathbf{x})\big| \leqslant \beta$.*

Метод релевантных векторов RVM [1] в своем наиболее распространенном варианте строит алгоритмы классификации в форме

$$y = \operatorname{sign}(g(\mathbf{x})) = \operatorname{sign}([\mathbf{u}](\mathbf{x})) = \operatorname{sign}\left(\sum_{i=1}^{m} u_i K(\mathbf{x}, \mathbf{x}_i)\right),$$

где $\mathbf{u} \in \mathbb{R}^m$, $K(\cdot,\cdot)$ — некоторая функция, называемая *ядром*.

Алгоритмический оператор $f = [\mathbf{w}]$ (или $f^i = [\mathbf{w}^i]$), получаемый этим методом обучения по выборке S (или S^i), доставляет минимум по параметру $g = [\mathbf{u}]$, соответственно, функционалам

$$\frac{1}{m}\sum_{k=1}^{m} \log\big(1 + e^{-y_k g(\mathbf{x}_k)}\big) + u^{\mathrm{T}}\Lambda u, \quad \frac{1}{m}\sum_{k\neq i} \log\big(1 + e^{-y_k g(\mathbf{x}_k)}\big) + u^{\mathrm{T}}\Lambda u,$$

где $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$, $\lambda_i \geqslant 0$, $i = 1, \ldots, m$ — автоматически вычисляемые коэффициенты регуляризации.

Определение 2 (Дивергенция Брегмана [2]). *Пусть $F\colon \mathbb{R}^m \to \mathbb{R}$ — выпуклая функция. Тогда, если $\nabla F(\mathbf{g})$ — градиент F в точке $\mathbf{g}$, то $F(\mathbf{g}') \geqslant F(\mathbf{g}) + \langle \mathbf{g}' - \mathbf{g}, \nabla F(\mathbf{g})\rangle$. Дивергенцией точек $\mathbf{g}$ и $\mathbf{g}'$ называется величина $d_F(\mathbf{g}', \mathbf{g}) \triangleq F(\mathbf{g}) - F(\mathbf{g}') - \langle \mathbf{g} - \mathbf{g}', \nabla F(\mathbf{g}')\rangle \geqslant 0$.*

Лемма 1 (О дивергенциях [2]). *Пусть $N\big([u]\big) = u^T\Lambda u$. Тогда*

$$d_N(f, f^i) + d_N(f^i, f) \leqslant \frac{1}{m}\big|f(\mathbf{x}_i) - f^i(\mathbf{x}_i)\big|.$$

Применим лемму 1 для RVM.

Лемма 2. *Для RVM справедлива оценка суммы дивергенций:*

$$d_N(f, f^i) + d_N(f^i, f) = 2\big\|\Lambda^{\frac{1}{2}}(\mathbf{w} - \mathbf{w}^i)\big\|^2.$$

Следовательно, по лемме о дивергенциях,

$$\big\|\Lambda^{\frac{1}{2}}(\mathbf{w} - \mathbf{w}^i)\big\|^2 \leqslant \frac{1}{2m}\big|f(\mathbf{x}_i) - f^i(\mathbf{x}_i)\big|.$$

Обозначим $m \times m$-матрицу $\big(K(\mathbf{x}_i, \mathbf{x}_j)\big)_{i,j=1}^m$ через $\hat{K}$.

Лемма 3. *Справедлива следующая оценка отклонения алгоритмических операторов RVM:*

$$\big|f(\mathbf{x}_j) - f^i(\mathbf{x}_j)\big| \leqslant \big\|\hat{K}^{\mathrm{T}}\Lambda^{-\frac{1}{2}}\big\| \cdot \big\|\Lambda^{\frac{1}{2}}(\mathbf{w} - \mathbf{w}^i)\big\|.$$

Теорема 4. *Метод релевантных векторов является β-устойчивым в обучении с показателем $\beta = \frac{1}{2m}\big\|\hat{K}^{\mathrm{T}}\Lambda^{-\frac{1}{2}}\big\|^2$. При этом $\big|R_{\mathrm{lo}}(f) - R_{\mathrm{em}}(f)\big| \leqslant \beta$.*

Работа поддержана РФФИ, проекты № 07-01-00211, № 05-01-00332.

Литература

[1] *Tipping M. E.* Sparse Bayesian Learning and the Relevance Vector Machines // Journal of Machine Learning Research. — 2001. — Vol. 1, № 5. — P. 211–244.

[2] *Bousquet O., Elisseeff A.* Stability and Generalization // Journal of Machine Learning Research. — 2002. — Vol. 2, № 3. — P. 499–526.

Свойства расстояния и меры опровержимости на высказываниях экспертов как формулах многозначных логик

Викентьев А. А., Новиков Д. В.

vikent@math.nsc.ru

Новосибирск, Институт Математики СО РАН,
Новосибирский государственный университет

В настоящее время появляется все больший интерес к построению решающих функций на основе анализа экспертной информации, заданной в виде вероятностных логических высказываний нескольких экспертов. В данной работе предложено записывать высказывания экспертов в виде формул n-значной логики Лукасевича. В произвольном случае найдено правильное обобщение расстояния между такими формулами и меры опровержимости таких формул, что позволяет более тонко (по сравнению

с двузначной логикой) решать прикладные задачи. В частности, значение истинности на модели может служить и субъективной вероятностью формулы. Ясно, что различные такие высказывания экспертов (и соответствующие им формулы) несут в себе разное количество информации, а, значит, возникает вопрос о ранжировании высказываний экспертов и сравнении их по информативности (далее — мере опровержимости при подтверждении высказывания). Для решения этих задач в работе будут введены и исследованы функция расстояния (см. [1]) между двумя такими формулами и мера опровержимости формул.

Определения основных понятий

Определение 1. *Множество элементарных высказываний $S^n(\varphi)$, используемых при написании формулы многозначной логики φ, назовем носителем формулы φ.*

Определение 2. *Назовем носителем совокупности знаний $S^n(\Sigma)$ объединение носителей формул, входящих во множество формул Σ, т. е. $S^n(\Sigma) = \bigcup_{\varphi\in\Sigma} S^n(\varphi)$.*

Определение 3. *Назовем множеством возможных значений носителя совокупности формул (знаний) с указанием всевозможных их значений истинности $Q_n(\Sigma) = \left\{\varphi_{\frac{k}{n-1}} \;\middle|\; \varphi \in S^n(\Sigma),\ k = 0, \ldots, n-1\right\}$.*

Далее нас интересуют значения истинности, отличные от нуля, $k > 0$.

Определение 4. *Моделью M назовем любое подмножество $Q_n(\Sigma)$ такое, что M не содержит одновременно $\varphi_{\frac{k}{n-1}}$ и $\varphi_{\frac{l}{n-1}}$ при любых $k \neq l$ и $\varphi \in Q(\Sigma)$.*

Множество всех моделей будем обозначать $P^n(S(\Sigma))$.

Для упрощения записи верхний индекс в формулах, означающий n-значность логики, будем опускать.

Лемма 1. $|P(S(\Sigma))| = n^{|S(\Sigma)|}$.

Введем обозначение для множества моделей формулы с фиксированным для нее значением истинности:

$$\mathrm{Mod}_{S(\Sigma)}(A)_{\frac{k}{n-1}} = \left\{M \;\middle|\; M \in P(S(\Sigma)),\ M \models A_{\frac{k}{n-1}}\right\}.$$

Определение 5. *Расстоянием между формулами φ и ψ, такими, что $S(\varphi) \cup S(\psi) \subseteq S(\Sigma)$, на множестве $P(S(\Sigma))$ назовем величину*

$$\rho_{S(\Sigma)}(\varphi,\psi) = \frac{\left|\bigcup\limits_{k=1}^{n-1} \mathrm{Mod}_{S(\Sigma)}\left(\varphi_{\frac{k}{n-1}} \wedge \psi_0\right)\right| + \left|\bigcup\limits_{k=1}^{n-1} \mathrm{Mod}_{S(\Sigma)}\left(\varphi_0 \wedge \psi_{\frac{k}{n-1}}\right)\right|}{n^{|S(\Sigma)|}}.$$

Свойства расстояний и меры опровержимости

Лемма 2. *Для любых формул φ, ψ, таких, что $S(\varphi) \cup S(\psi) \subseteq S(\Sigma)$ справедливы следующие утверждения:*

1) $0 \leqslant \rho_{S(\Sigma)}(\varphi, \psi) \leqslant 1$;
2) $\rho_{S(\Sigma)}(\varphi, \psi) = \rho_{S(\Sigma)}(\psi, \varphi)$;
3) $\rho_{S(\Sigma)}(\varphi, \psi) = 0 \Leftrightarrow \varphi \equiv \psi$;
4) $\rho_{S(\Sigma)}(\varphi, \psi) = 1 \Leftrightarrow \bigcup\limits_{l=1}^{n-1} \bigcup\limits_{k=1}^{n-1} \Big(\mathrm{Mod}(\varphi)_{\frac{k}{n-1}} \uplus \mathrm{Mod}(\psi)_{\frac{l}{n-1}} \Big) = P(S(\Sigma))$, *где $\uplus$ — прямое объединение;*
5) $\rho_{S(\Sigma)}(\varphi, \psi) \leqslant \rho_{S(\Sigma)}(\varphi, \chi) + \rho_{S(\Sigma)}(\chi, \psi)$;
6) Если $\varphi^1 \equiv \varphi^2$ *, то* $\rho_{S(\Sigma)}(\varphi^1, \psi) = \rho_{S(\Sigma)}(\varphi^2, \psi)$;

Лемма 3 (О расширении). *Для любого $S(\Sigma_0)$ такого, что $S(\varphi) \cup \cup S(\psi) \subseteq S(\Sigma_0)$ и любого $S(\Sigma_1)$ такого, что $S(\Sigma_0) \subseteq S(\Sigma_1)$, имеет место равенство:*

$$\rho_{S(\Sigma_0)}(\varphi, \psi) = \rho_{S(\Sigma_1)}(\varphi, \psi).$$

Лемма о расширении позволяет ограничить носители моделей при подсчете расстояний.

Определение 6. *Мерой опровержимости $I_{S(\Sigma)}(\varphi)$ для формул из $\Phi(\Sigma) = \big\{ \varphi \;\big|\; S(\varphi) \subset S(\Sigma) \big\}$ назовем величины*

$$I_{S(\Sigma)}(\varphi) = \sum_{i=0}^{n-2} \alpha_i \frac{\big| \mathrm{Mod}_{S(\Sigma)}\big(\varphi_{\frac{l}{n-1}}\big) \big|}{n^{|S(\Sigma)|}},$$

где α_i удовлетворяет условиям: $0 \leqslant \alpha_i \leqslant 1$, $\alpha_i + \alpha_{n-1-i} = 1$, $\alpha_k \geqslant \alpha_i$, для всех $i = 0, \ldots, \frac{n-1}{2}$ и всех $k = 0, \ldots, i$.

Лемма 4 (свойства меры $I_{S(\Sigma)}$). *Для любых $\varphi, \psi \in \Phi(\Sigma)$*

1) $0 \leqslant I_{S(\Sigma)}(\varphi) \leqslant 1$;
2) $I_{S(\Sigma)}(\varphi) + I_{S(\Sigma)}(\neg\varphi) = 1$;
3) $I_{S(\Sigma)}(\varphi \wedge \psi) \geqslant \max\big\{ I_{S(\Sigma)}(\varphi), I_{S(\Sigma)}(\psi) \big\}$;
4) $I_{S(\Sigma)}(\varphi \vee \psi) \leqslant \min\big\{ I_{S(\Sigma)}(\varphi), I_{S(\Sigma)}(\psi) \big\}$;
5) $I_{S(\Sigma)}(\varphi \vee \psi) + I_{S(\Sigma)}(\varphi \wedge \psi) = I_{S(\Sigma)}(\varphi) + I_{S(\Sigma)}(\psi)$;
6) $I^3_{S(\Sigma)}(\varphi \wedge \psi) = \frac{1}{2}\big(I^3_{S(\Sigma)}(\varphi) + I^3_{S(\Sigma)}(\psi) + \rho^3_{S(\Sigma)}(\neg\varphi, \neg\psi) \big)$;
7) $I^3_{S(\Sigma)}(\varphi \vee \psi) = \frac{1}{2}\big(I^3_{S(\Sigma)}(\varphi) + I^3_{S(\Sigma)}(\psi) - \rho^3_{S(\Sigma)}(\neg\varphi, \neg\psi) \big)$.

Эта лемма доказывает общие свойства меры опровержимости, а для $n = 3$ указывает на справедливость гипотезы Г. С. Лбова, верную для $n = 2$, см. [1]. При $n > 3$ такой связи с расстоянием нет, но есть более сложные зависимости. Доказаны также и другие свойства расстояний

и меры опровержимости для частного случая $n = 3$, похожие на случай $n = 2$ (см., например, [1]). Все результаты использованы при написании программы вторым автором и апробированы на прикладной задаче при различных n. Подбор нужного значения n в конкретной задаче является частью процесса адаптации для введения расстояния и меры опровержимости для получения более тонких знаний. В случае $n = 2$ проведены теоретические исследования по следующим вопросам. При организации поиска логических закономерностей требуются расстояния между высказываниями экспертов и формулами в моделях в произвольный (текущий) момент времени с фиксированными знаниями. Планируем обработку сообщений экспертов в различные моменты (срезы) времени с возможностью того, что исходные гипотезы-предположения у экспертов, вообще говоря, могут изменяться. Значит, будет происходить адаптация во времени самой теории (по знаниям экспертов), и, соответственно этому, будем применять другие модели экспертов. Аппарат для обработки таких знаний подготовлен в работах Викентьева А. А., начатых со Лбовым Г. С. и Кореневой Л. Н. Сигнал о смене класса моделей (а, значит, и теории) будет исходить либо от самих экспертов (по их изменяющимся знаниям), либо при получении неправильных результатов инженером-разработчиком Базы Знаний при использовании старой теории.

Работа выполнена при поддержке РФФИ, проект № 07-01-00331а.

Литература

[1] *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: Изд-во Ин-та математики СО РАН, 1999.

[2] *Кейслер Г., Чэн Ч. Ч.* Теория моделей. — М.: Мир, 1977.

[3] *Карпенко А. С.* Логики Лукасевича и простые числа — М.: Наука, 2000.

Слабая вероятностная аксиоматика и надёжность эмпирических предсказаний

Воронцов К. В.

voron@ccas.ru

Москва, Вычислительный центр РАН

Задача эмпирического предсказания является одной из центральных в прикладной статистике и машинном обучении: получив выборку данных, необходимо предсказать определённые свойств аналогичных данных, которые станут известны позже, и оценить точность предсказания. В сообщении предлагается новая формализация постановки задачи, не требующая привлечения классической вероятностной аксиоматики.

Пусть задано множество объектов $\mathbb{X}$ и выборка $X^L = (x_1, \ldots, x_L) \subseteq \mathbb{X}$ длины L. Рассмотрим множество всех её разбиений на две подвыборки длины ℓ и k соответственно: $X^L = X_n^\ell \cup X_n^k,\ \ell + k = L$, где нижний индекс $n = 1, \ldots, N$ пробегает все $N = C_L^k$ разбиений.

Пусть задано множество R и функция $T\colon \mathbb{X}^* \times \mathbb{X}^* \to R$, где $\mathbb{X}^*$ — множество всех конечных выборок из $\mathbb{X}$.

Рассмотрим эксперимент, в котором с равной вероятностью реализуется одно из разбиений n, после чего наблюдателю сообщается выборка X_n^ℓ. Не зная *скрытой выборки* X_n^k, наблюдатель должен построить функцию $\hat{T}\colon \mathbb{X}^* \to R$, значение которой на *наблюдаемой выборке* $\hat{T}_n = \hat{T}(X_n^\ell)$ предсказывало бы значение $T_n = T(X_n^k, X_n^\ell)$, существенно зависящее от скрытой выборки X_n^k. Требуется также оценить надёжность предсказания, т. е. указать *оценочную функцию* $\eta(\varepsilon)$ такую, что

$$\mathsf{P}_n\big\{d(\hat{T}_n, T_n) > \varepsilon\big\} \leqslant \eta(\varepsilon), \tag{1}$$

где $d\colon R \times R \to \mathbb{R}$ — заданная функция, характеризующая величину отклонения $d(\hat{r}, r)$ предсказанного значения $\hat{r} \in R$ от неизвестного истинного значения $r \in R$. Параметр ε называется *точностью*, а величина $(1 - \eta(\varepsilon))$ — *надёжностью* предсказания. Если в (1) достигается равенство, то $\eta(\varepsilon)$ называется *точной оценкой*. Оценка $\eta(\varepsilon)$ может зависеть от ℓ и k, а также от вида функций T и $\hat{T}$. Если (1) выполняется при достаточно малых ε и η, то говорят, что в окрестности предсказываемого значения имеет место *концентрация вероятности* [5].

Заметим, что данная постановка задачи не опирается на классическую аксиоматику теории вероятностей. Здесь понятие вероятности является лишь синонимом доли разбиений: $\mathsf{P}_n\{\varphi(n)\} = \frac{1}{N}\sum_{n=1}^{N}\varphi(n)$ для произвольного предиката $\varphi\colon \{1, \ldots, N\} \to \{0, 1\}$, заданного на множестве разбиений выборки X^L. Тем не менее, мы предпочитаем пользоваться привычным термином *вероятность* и говорить, что задача эмпирического предсказания поставлена в *слабой вероятностной аксиоматике*.

Слабая аксиоматика ориентирована на задачи анализа данных, в которых все выборки конечные и все величины наблюдаемые, т. е. являются функциями конечных выборок. В классической колмогоровской аксиоматике вероятность события, функция распределения и матожидание случайной величины являются величинами ненаблюдаемыми. В задачах анализа данных слабая аксиоматика имеет ряд преимуществ.

1. Упрощается понятийный аппарат. Нет необходимости использовать теорию меры, предельный переход к бесконечной выборке, различные типы сходимости, и т. д. Однако это не мешает сформулировать и доказать аналоги многих фундаментальных утверждений теории вероятностей и математической статистики: закон больших чисел, сходимость

эмпирических распределений (критерий Смирнова), ранговые критерии, оценки Вапника-Червоненкиса [6], и т. д.

2. Сильная (колмогоровская) аксиоматика требует, чтобы на множестве объектов $\mathbb{X}$ существовала σ-аддитивная алгебра событий, объекты X^L выбирались случайно из фиксированной генеральной совокупности, и все рассматриваемые функции выборок были измеримы. Требования случайности, независимости и одинаковой распределённости могут быть проверены с помощью статистических тестов. Однако гипотезы σ-аддитивности и измеримости эмпирической проверке не поддаются [1]. Слабая аксиоматика обходится без этих гипотез. Фактически, в ней остаётся только гипотеза равновероятности разбиений, эквивалентная предположению о независимости выборки X^L. Об объектах вне выборки X^L вообще не делается никаких предположений.

3. Из оценки «слабого» функционала $\mathsf{P}_n\{\varphi(X_n^\ell, X_n^k)\} \leqslant \eta$ всегда можно получить оценку «сильного» функционала, взяв матожидание по выборке X^L от обеих частей неравенства:

$$\mathsf{E}_{X^L}\,\mathsf{P}_n\{\varphi(X_n^\ell, X_n^k)\} = \mathsf{P}_{X^L}\{\varphi(X^\ell, X^k)\} \leqslant \mathsf{E}_{X^L}\eta.$$

Если η не зависит от выборки, то оценка переносится непосредственно. Для оценок типа Вапника-Червоненкиса это было проделано в [3].

4. С другой стороны, «слабые» функционалы легко поддаются эмпирическому измерению. Для этого суммирование по всем разбиениям заменяется суммированием по некоторому подмножеству разбиений (в методе Монте-Карло — по случайному). Таким образом, слабая аксиоматика является единой отправной точкой как для теоретико-вероятностного, так и для экспериментального анализа надёжности эмпирических предсказаний. В теории машинного обучения становится предельно понятной связь между теоретическими оценками обобщающей способности и практическими методиками, основанными на скользящем контроле.

5. В современной вычислительной теории обучения [5] для получения верхних оценок надёжности используется математический аппарат функционального анализа и оценки концентрации вероятностной меры. Это мощная и красивая математическая теория, но, к сожалению, в ходе вывода оценок их точность теряется практически бесконтрольно на многочисленных промежуточных шагах. Проблема в том, что «асимптотичность» заложена как в самом понятии вероятности, так и в стремлении получить оценку в виде «изящной» формулы, даже ценой значительной потери её точности. Слабая аксиоматика во многих случаях приводит к точным, не асимптотическим, оценкам. Иногда это довольно сложные комбинаторные выражения, требующие значительных объёмов вычисле-

ний. Однако во многих случаях находятся эффективные алгоритмические решения вычислительных проблем.

Вышесказанное позволяет выдвинуть смелую гипотезу: *для исследования задач эмпирического предсказания достаточно слабой вероятностной аксиоматики.* Пока вопрос о границах её применимости остаётся открытым. Приведём два высказывания, подтолкнувших автора к проведению данного исследования.

А. Н. Колмогоров [4]: «. . . представляется важной задача освобождения всюду, где это возможно, от излишних вероятностных допущений. На независимой ценности чисто комбинаторного подхода к теории информации я неоднократно настаивал в своих лекциях».

Ю. К. Беляев [2]: «. . . возникло глубокое убеждение, что в теории выборочных методов можно получить содержательные аналоги большинства основных утверждений теории вероятностей и математической статистики, которые к настоящему времени найдены в предположении взаимной независимости результатов измерений».

Работа выполнена при поддержке РФФИ, проект №05-01-00877, и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

[1] *Алимов Ю. И.* Альтернатива методу математической статистики. — Знание, 1980.

[2] *Беляев Ю. К.* Вероятностные методы выборочного контроля. — М.: Наука, 1975.

[3] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / Под ред. О. Б. Лупанова. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[4] *Колмогоров А. Н.* Теория информации и теория алгоритмов / Под ред. Ю. В. Прохорова. — М.: Наука, 1987.

[5] *Lugosi G.* On concentration-of-measure inequalities. // Machine Learning Summer School, Australian National University, Canberra. — 2003.

[6] *Vapnik V.* Statistical Learning Theory. — Wiley, New York, 1998.

О теоретико-возможностном методе медицинской диагностики

Газарян В. А., Нагорный Ю. М., Пытьев Ю. П.
pytyev@phys.msu.ru
Москва, МГУ

Теоретико-возможностный метод используется для решения задач медицинской диагностики с помощью возможностного аналога алгоритма «Кора».

Функциональные нарушения в системе пищеварения активно изучаются на протяжении ряда последних лет. Проводимые совместно с врачами клиники НИИ питания РАМН исследования позволили значительно продвинуться на пути создания нового комплексного метода компьютерной оценки состояния больных с функциональными нарушениями. По обучающей выборке опросников больных — пациентов НИИ питания с верифицированным диагнозом — с помощью алгоритма типа «Кора» [1] был создан комплексный тест оценки самочувствия больного синдромом раздраженной толстой кишки (СРТК) и описана симптоматика СРТК в общем виде. Однако, как было показано в [2], детерминистские и статистические методы недостаточно эффективны при неформализованном характере данных и ограниченном размере обучающей выборки. Для уточнения диагностических критериев был привлечен теоретико-возможностный метод, что привело к созданию алгоритма, ранжирующего группы симптомов по их возможностям при определенных заболеваниях [2]. В данной работе диагностика заболеваний при нечетких данных проводится на основании решающего правила, *оптимальнoго с теоретико-возможностной точки зрения, т. е. минимизирующего возможность ошибки классификации.*

Оптимальное решение задачи классификации достигается при минимизации возможности (необходимости) потерь [3]. Согласно теореме о P-оптимизации, субъект χ следует отнести к классу q^*:

$$P(q^*, x) = \min_q P(q, x), \tag{1}$$

$$P(q, x) = \sup_k \min(\varphi^{\chi|\varkappa}(x|k), l(k, q)), \tag{2}$$

где $P(q, x)$ — возможность ошибки при отнесении субъекта $\chi = x$ к классу $\varkappa = q$, $\chi = \left(\chi^1, \ldots, \chi^n\right)$, $x = \left(x^1, \ldots, x^n\right)$, x^j — значение j-го признака (симптома), $\varphi^{\chi|\varkappa}(x|k)$ — условное распределение возможностей пациенту класса k обладать признаками (симптомами) x.

Условное распределение $\varphi^{\chi|\varkappa}(x|k)$ получим в результате стохастического моделирования условных возможностей определенных наборов

признаков при условии принадлежности субъекта выделенным классам путем применения *оптимизированного* алгоритма гранулирования [4]. Обучение возможностного аналога алгоритма классификации «Кора» проводится на основании ранжированных по возможностям групп признаков.

В каждом классе k определяется набор w_{s_k}, имеющий максимальную переходную возможность:

$$p(k|w_{s_k}) = \max_s p(k|w_s),\ s = 1, \ldots, S_k.$$

Далее в (2) значение $p(x|k)$ используется в качестве $\varphi^{\chi|\varkappa}(x|k)$.

Субъект x относится к тому классу q^*, которому соответствует минимальная возможность ошибки (1) (в котором есть набор $w_{s_{q^*}}$, имеющий минимальную возможность ошибки):

$$P(q^*, w_{s_{q^*}}) = \min_q P(q, w_{s_q}).$$

В докладе приведены результаты применения теоретико-возможностного метода к задачам диагностики функциональных нарушений системы пищеварения и острого аппендицита.

Литература

[1] *Газарян В. А., Иваницкая Н. В., Пытьев Ю. П., Шаховская А. К.* // Вестн. Моск. ун-та. Физ. Астрон. — 2003. — № 2. — С. 12.

[2] *Газарян В. А., Илюшин В. Л., Пытьев Ю. П., Шаховская А. К.* // Вестн. Моск. ун-та. Физ. Астрон. — 2005. — № 4. — С. 3.

[3] *Пытьев Ю. П.* Возможность. Элементы теории и применения. — М.: УРСС, 2000.

[4] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применения. — М.: Физматлит, 2007.

Решение задач распознавания с невыполненной гипотезой компактности

Гуров С. И., Потепалов Д. Н., Фатхутдинов И. Н.
sgur@cs.msu.ru, mmp@cs.msu.ru
Москва, МГУ им. М. В. Ломоносова, факультет ВМиК

Реализованные на сегодняшний день алгоритмы логического синтеза имеют разную эффективность на различных типах схем. Поэтому возникает задача определения наилучшего алгоритма синтеза исходя из тех или иных характеристик входного описания схемы.

В первой решённой задаче зафиксирован набор синтезирующих алгоритмов и выбран критерий качества синтеза. Для совокупности описаний схем (прецедентов) выясняется распределение их по классам: принадлежность данному классу означает, что синтезированная соответствующим алгоритмом схема имеет наилучшие характеристики по указанному критерию. Целью исследования являлась разработка и реализация метода определения областей компетентности алгоритмов, позволяющего для каждого нового описания схемы указать наилучший (в данном смысле) алгоритм её синтеза.

Поставленная задача решалась методами распознавания образов. Суть предложенного решения заключается в построении признакового пространства образов (описаний схем) с последующим решением в нём задачи классификации. Первый, почти не разработанный в теоретическом плане и трудоёмкий на практике этап состоит в: (1) фиксации набора первичных числовых признаков прецедентов; (2) построении достаточно представительной совокупности вторичных признаков как функций от первичных; (3) отборе наиболее информативных вторичных признаков. На втором этапе строится решающее правило, относящее описание схемы в построенном признаковом пространстве к тому или иному классу.

Во второй задаче требовалось определить, сумеет ли имеющийся алгоритм оптимизации схемы обработать её в заданное время. Эта задача решалась аналогично предыдущей.

Разработанные алгоритмы указанных задач оказались эффективными.

Анализ показывал, что схемы из одного и того же класса образовывали некомпактные области в пространствах их исходного описания, т.е. для них не выполнялась т.н. гипотеза компактности (ГК). Это характеризует задачи как чрезвычайно сложные в практическом и теоретическом аспектах, для которых обычные методы распознавания оказываются непригодными.

Гипотеза компактности в задачах распознавания

Наиболее общая неформальная (и, как представляется, наиболее адекватная сегодняшнему уровню понимания проблемы) формулировка ГК предложена ещё в классической монографии [1]: «*образам соответствуют компактные множества в пространстве выбранных свойств*»[1]. Предположение о выполнении ГК лежит в основе подавляющего большинства подходов к решению различных типов задач распознавания (а при решении задач кластеризации является определяющим). Действительно, при выполнении ГК полученные в рамках любого из традиционных методов алгоритмы классификации имеют, очевидно, невысокую сложность и, как следствие, легко реализуемы, не требуют больших вычислительных ресурсов по памяти, времени счёта, и т. д. Также важной характеристикой таких алгоритмов являются достаточно высокие оценки их надёжности — прямое следствием выполнения ГК.

Задачи, где ГК не имеет места, указанные «хорошие» свойства алгоритмов распознавания — при использовании обычных подходов — отнюдь не гарантируются. В силу этого на протяжении последних лет они находились вне круга интересов разработчиков: теория, способная дать направления решения таких задач отсутствовала. Здесь надо сказать, что попытки решения задач с невыполненной ГК делались ещё на заре развития теории и практики распознавания образов. В известной монографии М. М. Бонгарда [3] приведён демонстрационный пример решения задачи указанного типа. Однако этот подход на многие годы оставался невостребованным...

Для решения задач данного типа необходимо было понять, что же понимается под неформальным понятием «компактные образы», откуда берётся и как формируется пространство свойств, и т. д. В связи с этим в последнее время были предприняты некоторые попытки формализации ГК [6, 5]. Однако они либо носили частный характер, либо уводили проблему в общефилософскую плоскость [2]. Так что на сегодняшний день этот вопрос является открытым.

В данной работе приведены примеры практического решения двух задач с невыполненной ГК. Решения основывались на подходе М. М. Бонгарда. При этом мы не предлагаем своей формулировки ГК, посчитав имеющиеся у нас попытки её определения предварительными и оставив их для дальнейших публикаций по данной теме исследования.

[1] Данная формулировка, по сути, лишь несколько уточняет первоначальную, высказанную М. А. Айзерманом в первых работах по распознаванию образов конца 50-х годов XX в.

Решения задач. Результаты

Для решения задач использовался вышеупомянутый подход М. М. Бонгарда. На основе первичных признаков (до 10–12) объектов генерировались вторичные признаки (до десятков тысяч) как функции от первичных, из которых отбирались наиболее информативные. Окончательно задачи классификации (определения областей компетентности имеющихся алгоритмов) решались в сформированном признаковом пространстве. Ошибка классификации, определяемая методом скользящего контроля с одним исключаемым прецедентом, не превосходила нескольких процентов.

В задаче № 1 исследовались 19 практических алгоритмов синтеза схем и 120 описаний реальных схем (первые результаты см. в [4]). В задаче № 2 исследовались 185 схем (разбиение по классам: 175 + 10).

Программные модули, реализующие разработанные алгоритмы, интегрированы в открытую среду SIS [7], информационно совместимую со многими промышленными системами синтеза БИС, в частности используемой в фирме Intel Inc.

В развитие подхода видится целесообразным разработать способ уменьшения числа рассматриваемых вторичных признаков на основе исключения из рассмотрения тождественно равных, но структурно отличных признаков и установления нахождения причинно-следственной связи между информативностью двух различных вторичных признаков.

Работа выполнена при поддержке РФФИ, проект № 07-01-00211.

Литература

[1] *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — 384 с.

[2] *Белозерский Л. А.* Современный взгляд на гипотезу компактности // Штучний інтелект (Донецк). — 2005. — № 3. — С. 6–12.

[3] *Бонгард М. М.* Проблема узнавания. — М.: Наука, 1967.

[4] *Гранкин М. В., Гуров С. И., Фатхутдинов И. Н.* Определение областей компетентности алгоритмов при невыполненной гипотезе компактности // Штучний інтелект (Донецк). — 2006. — № 2. — С. 88–98.

[5] *Донской В. И.* О метрических свойствах кратчайших эмпирических закономерностей // Уч. записки ун-та им. В. И. Вернадского. — 2003. — № 2. — С. 143–147.

[6] *Загоруйко Н. Г., Елкина В. Н., Киприянова Т. П.* Пакет Прикладных Программ ОТЭКС для Обработки Таблиц Экпериментальных данных. Версия 4.0 — `www.math.nsc.ru/AP/oteks/Russian/`.

[7] SIS: A System for Sequential Circuit Synthesis // Dep. of Electrical Engineering and Computer Science Univ. of California, Berkeley. — 1992.

Алгоритмы идентификации изображений в случайной и нечёткой морфологии

Зубюк А. В.
zubuk@cmpd2.phys.msu.su
Москва, МГУ им. М. В. Ломоносова, физический факультет

В работе [1] рассмотрены основные понятия случайной и нечёткой морфологии и даны постановки ряда задач идентификации изображений, имеющих случайную или нечёткую форму. В настоящем докладе рассмотрены алгоритмы решения этих задач.

Математические методы морфологического анализа изображений предназначены для решения задач анализа и интерпретации изображений реальных сцен, полученных при неопределённых условиях, таких, например, как характер освещения, его спектральный состав, оптические характеристики объектов и т. п. Очевидно, что изображения одной и той же сцены, полученные при разных условиях наблюдения, могут существенно отличаться друг от друга. Это обстоятельство является источником ряда проблем, возникающих при решении указанных задач. В частности, невозможно путём попиксельного сравнения яркостей определить, являются ли два изображения изображениями одной и той же сцены. В то же время, все изображения одной и той же сцены имеют «сходные черты», позволяющие отличать изображения этой сцены от изображений других сцен. Такой «сходной чертой» в ряде случаев может являться геометрическая форма изображённых объектов. Таким образом, существует инвариант, не изменяющийся при изменении условий наблюдения. Методы морфологического анализа изображений ориентированы, прежде всего, на анализ формы изображённых объектов в терминах, инвариантных относительно условий получения изображений [2, 3].

Пусть моделью изображения является элемент евклидова пространства $\mathcal{R}_N$. Тогда всевозможные изменения условий его регистраций приведут к тому, что изображение одного и того же объекта будет изменяться в пределах некоторого множества V пространства $\mathcal{R}_N$. Это множество называется *формой изображения этого объекта* [2], т. к. оно отражает его характеристики, не зависящие от условий регистрации изображений.

Случайная и нечёткая формы изображений

Напомним кратко основные определения, которые были даны в [1].

Определение 1. *Случайной формой элементов пространства $\mathcal{R}_N$ назовём вероятностное пространство* $(\Omega, \mathcal{A}, \Pr)$, *где Ω — множество непересекающихся подмножеств евклидова пространства $\mathcal{R}_N$ (форм), образующих разбиение пространства $\mathcal{R}_N$, $\mathcal{A}$ — некоторая σ-алгебра подмножеств множества Ω,* а $\Pr$ — *заданная на ней вероятность.*

Каждому событию $A \in \mathcal{A}$ соответствует форма (подмножество) $V = \bigcup_{\omega \in A} \omega \subset \mathcal{R}_N$, вероятность которой есть $\Pr(A)$. По аналогии с определением 1 введём понятие нечёткой формы:

Определение 2. *Нечёткой формой элементов пространства $\mathcal{R}_N$ назовём пространство с возможностью* $(\Omega, \mathcal{A}, \mathrm{P})$, *где Ω — множество непересекающихся форм, образующих разбиение $\mathcal{R}_N$, $\mathcal{A}$ — некоторая σ-алгебра подмножеств Ω, а* P *— заданная на ней возможность.*

Задачи идентификации изображений, имеющих случайную или нечеткую форму

В докладе рассмотрены алгоритмы решения задач идентификации изображений, поставленных в [1]. Помимо этого, рассмотрены задачи, в которых известна некоторая априорная информация о предъявляемых изображениях. Такая информация заключается в следующем. Предъявляемое изображение является случайным (или нечётким) и для каждой формы $\omega \in \Omega$ известно его условное распределение внутри этой формы. Учёт этой информации делает гипотезы в минимаксной задаче проверки гипотез простыми. Приведём постановку такой задачи в случайной морфологии.

Пусть заданы две случайные формы: $F_1 = (\Omega, \mathcal{A}, \Pr_1)$ и $F_2 = (\Omega, \mathcal{A}, \Pr_2)$, где вероятности $\Pr_1$ и $\Pr_2$ заданы плотностями $\mathrm{pr}^{(1)}(\cdot)$ и $\mathrm{pr}^{(2)}(\cdot)$ соответственно, и предъявляемое для идентификации изображение ξ формируется по схеме $\xi = f + \nu$, где f и ν — случайные элементы пространства $\mathcal{R}_N$ (случайные элементы f, ω и ν независимы). Требуется по предъявленному изображению ξ определить, какую случайную форму (F_1 или F_2) имеет элемент f. Для этого можно воспользоваться рандомизированным критерием π, являющимся решением следующей минимаксной задачи:

$$\begin{cases} \max(\alpha_1, \alpha_2) \sim \min\limits_{\pi_1, \pi_2}; \\ \pi_1(x) + \pi_2(x) = 1, \ x \in \mathcal{R}_N; \\ \pi_i(x) \geqslant 0, \ x \in \mathcal{R}_N, \ i = 1, 2; \end{cases} \tag{1}$$

где $\alpha_i \stackrel{\mathrm{def}}{=} 1 - \int_{\mathcal{R}_N} \mathrm{pr}^t_\xi(x)\, \pi_i(x)\, dx$, $i = 1, 2$, $\mathrm{pr}^t_\xi(x)$ — распределение случайного элемента ξ, зависящее от распределений элементов ω, f и ν.

Алгоритмы решения поставленных задач и их свойства

В докладе рассмотрено применение алгоритмов типа случайного поиска (см [5, 6]) для решения минимаксных задач. В частности, для решения минимаксной задачи в теоретико-вероятностной постановке она

сводится к задаче поиска априорного распределения, выравнивающего частные ошибки в байесовской задаче проверки гипотез, где условными распределениями величины ξ являются её распределения, используемые в минимаксной задаче. Такое априорное распределение является наихудшим в смысле полной байесовской ошибки (см. [4]). Для нахождения «выравнивающего» распределения используются методы градиентного случайного поиска (см. [5]), в частности, алгоритм с парной пробой. При этом в процессе работы алгоритма не производится точное вычисление целевого функционала (разности частных ошибок). Вместо этого используется его оценка — разность частот событий, соответствующих частным ошибкам. Такой способ позволил увеличить скорость численного решения минимаксных задач, возникающих при идентификации изображений. В докладе также рассмотрено применение алгоритмов типа случайного поиска для решения минимаксных задач идентификации изображений, имеющих нечёткую форму.

Для всех разработанных алгоритмов доказаны теоремы о сходимости решающих правил, получаемых в результате их работы, к решениям поставленных задач.

Приведённые в докладе алгоритмы позволяют решать задачи анализа и интерпретации реальных сцен, упомянутые во введении.

Работа выполнена при поддержке РФФИ, проект № 05-01-00532-а.

Литература

[1] *Пытьев Ю. П., Зубюк А. В.* Случайная и нечёткая морфология (эмпирическое восстановление модели, идентификация) // Материалы IX Межд. конф. «Интеллектуальные системы и компьютерные науки». — 2006.

[2] *Пытьев Ю. П.* Морфологический анализ изображений // Докл. АН СССР. — 1983. — Т. 269, № 5. — С. 1061–1064.

[3] *Pyt'ev Yu. P.* Morphological Image Analysis // Pattern Recognition and Image Analysis. — 1993. — Vol. 3, № 1. — P. 19–28.

[4] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применения. — Физматлит, 2007.

[5] *Гладков Д. И.* Оптимизация систем неградиентным случайным поиском. — М: Энергоатомиздат, 1984.

[6] *Жиглявский А. А.* Математическая теория глобального случайного поиска — Л: Изд-во Ленингр. ун-та, 1985.

Верхние оценки переобученности и профили разнообразия логических закономерностей

Ивахненко А. А., Воронцов К. В.
andrej_iv@mail.ru, voron@ccas.ru
Москва, Вычислительный Центр РАН

Логические алгоритмы классификации представляют собой композиции элементарных классификаторов, называемых также закономерностями. Существуют два противоположных подхода к повышению качества (обобщающей способности) таких алгоритмов: либо увеличение числа закономерностей в композиции [1], либо повышение качества закономерностей. Качество алгоритма в обоих случаях может оказаться сопоставимым, однако при втором подходе получаются более простые, легко интерпретируемые алгоритмы. В данной работе понятие обобщающей способности, которое обычно определяется для алгоритмов, распространяется на случай закономерностей. В рамках комбинаторного подхода [2] выводятся сложностные оценки качества закономерностей. Предлагается методика эмпирического измерения завышенности получаемых оценок, основанная на скользящем контроле.

Основные определения

Рассмотрим стандартную постановку задачи классификации. Задано множество допустимых объектов X, конечное множество имён классов Y и обучающая выборка $X^\ell = (x_i, y_i)_{i=1}^\ell \subset X \times Y$. Предполагается, что $y_i = y^*(x_i)$, где $y^*\colon X \to Y$ — неизвестная целевая зависимость. Требуется построить алгоритм $a\colon X \to Y$, приближающий y^* на всём X.

Закономерностью называется предикат $\varphi_y\colon X \to \{0,1\}$, выделяющий достаточно много объектов класса y и достаточно мало объектов всех остальных классов. Предикат φ_y *выделяет* объект x, если $\varphi_y(x) = 1$.

Логические алгоритмы представляются в виде линейных композиций вида $a(x) = \arg\max\limits_{y\in Y} \sum_{t=1}^{T_y} w_y^t \varphi_y^t(x)$, где φ_y^t — закономерности, w_y^t — веса закономерностей, T_y — число закономерностей класса y.

Методом обучения называется отображение μ, которое по выборке X^ℓ строит набор закономерностей $\mu X^\ell \equiv \mu(X^\ell) = \left\{\varphi_y^t(x)\right\}_{y\in Y}^{t=1,T_y}$.

Частота ошибок закономерности φ_y на выборке $U \subset X$ есть

$$\nu(\varphi_y, U) = \frac{1}{|U|} \sum_{x\in U} \big[\varphi_y(x) \neq [y^*(x) = y]\big].$$

Переобученностью закономерности $\varphi_y \in \mu X^\ell$ при заданной контрольной выборке X^k называется разность частот её ошибок на контроле и на обучении $\delta(\varphi_y, X^\ell, X^k) = \nu(\varphi_y, X^k) - \nu(\varphi_y, X^\ell)$.

Рассмотрим множество всех разбиений полной выборки $X^L = X_n^\ell \cup X_n^k$ на две подвыборки — обучающую длины ℓ и контрольную длины k, где $\ell + k = L$, индекс n пробегает множество всех разбиений $N = \{1, \ldots, C_L^\ell\}$.

Введём функционал *полного скользящего контроля* $Q_\varepsilon(\mu, X^L)$ как долю переобученных закономерностей среди закономерностей, построенных методом μ по всевозможным подвыборкам $X_n^\ell \subset X^L$:

$$Q_\varepsilon(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} \frac{1}{|\mu X_n^\ell|} \sum_{\varphi \in \mu X_n^\ell} \big[\delta(\varphi, X_n^\ell, X_n^k) > \varepsilon\big].$$

где $\varepsilon \in [0, 1)$ — *порог переобученности*. Аналогичный функционал, его верхние оценки и связь с теорией Вапника-Червоненкиса рассматривались в [2] для алгоритмов классификации и регрессии.

Коэффициенты и профили разнообразия

Назовем предикаты $\varphi, \varphi' \colon X \to \{0, 1\}$ *неразличимыми* или эквивалентными на выборке X^L, если $\varphi(x) = \varphi'(x)$ для всех $x \in X^L$. *Коэффициентом разнообразия* (shatter coefficient) $\Delta(\Phi, X^L)$ множества предикатов Φ на выборке X^L называется максимальное число попарно неразличимых предикатов из Φ, оно же число классов эквивалентности на Φ. Коэффициент разнообразия характеризует сложность множества предикатов Φ относительно заданной выборки X^L.

Рассмотрим множество закономерностей, получаемых методом μ по всевозможным обучающим подвыборкам: $\Phi_L^\ell \equiv \Phi_L^\ell(\mu, X^L) = \bigcup_{n=1}^N \mu X_n^\ell$. Его коэффициент разнообразия $\Delta_L^\ell \equiv \Delta_L^\ell(\mu, X^L) = \Delta(\Phi_L^\ell, X^L)$ назовём *локальным коэффициентом разнообразия* метода μ на выборке X^L.

Разобьём множество Φ_L^ℓ на $L + 1$ подмножеств, состоящих из закономерностей с фиксированным числом ошибок m на полной выборке X^L: $\Phi_m \equiv \Phi_m(\mu, X^L) = \big\{\varphi \in \Phi_L^\ell \colon \nu(\varphi, X^L) = \frac{m}{L}\big\},\ m = 0, \ldots, L$.

Локальным профилем разнообразия метода μ на выборке X^L назовём последовательность коэффициентов разнообразия $D_m \equiv D_m(\mu, X^L) = \Delta(\Phi_m, X^L),\ m = 0, \ldots, L$. Очевидно, что $\Delta_L^\ell = D_0 + \cdots + D_L$.

Наряду с функционалом Q_ε определим функционал $Q_{\varepsilon,m}$ как долю переобученных закономерностей, допускающих m ошибок на X^L:

$$Q_{\varepsilon,m}(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} \frac{1}{|\mu X_n^\ell|} \sum_{\varphi \in \mu X_n^\ell} \big[\delta(\varphi, X_n^\ell, X_n^k) > \varepsilon\big] \big[\nu(\varphi, X^L) = \tfrac{m}{L}\big].$$

Теорема 1. *Для любых μ, X^L и порога переобученности $\varepsilon \in [0, 1)$*

$$Q_{\varepsilon,m}(\mu, X^L) \leqslant D_m H\tbinom{m\ s_1}{L\ \ell}, \quad m = 0, \ldots, L, \tag{1}$$

где $H\binom{m\ s_1}{L\ \ell} = \sum_{s=s_0}^{s_1} C_m^s C_{L-m}^{\ell-s} / C_L^\ell$ — хвост гипергеометрического распределения, $s_0 = \max\{0, m - k\}$ и $s_1 = \big\lfloor \frac{\ell}{L}(m - \varepsilon k) \big\rfloor$.

Задача	L	Глобальный	Локальный	Эффективный
crx	690	$2.8 \cdot 10^8$	$3.5 \cdot 10^4$	21 ± 11
german	1000	$5.2 \cdot 10^8$	$3.1 \cdot 10^4$	47 ± 38
hepatitis	155	$5.5 \cdot 10^6$	$1.8 \cdot 10^4$	58 ± 46
horse-colic	300	$1.9 \cdot 10^6$	$1.3 \cdot 10^4$	5 ± 3
hypothyroid	3163	$5.3 \cdot 10^8$	$2.2 \cdot 10^4$	43 ± 28
liver	345	$1.5 \cdot 10^7$	$2.9 \cdot 10^4$	12 ± 8
promoters	106	$4.4 \cdot 10^9$	$5.3 \cdot 10^4$	13 ± 4

Таблица 1. Коэффициенты разнообразия на 7 задачах классификации из репозитория UCI. Выборка разбивалась 20 раз случайным образом на равные части $\ell = k$ со стратификацией классов; $\varepsilon = 0.05$.

Теорема 2. *Для любых μ, X^L и порога переобученности $\varepsilon \in [0,1)$*

$$Q_\varepsilon(\mu, X^L) \leqslant \sum_{m=0}^{L} D_m H\tbinom{m\ s_1}{L\ \ell}.$$

Определим *эффективный локальный профиль разнообразия* $\widehat{D}_m$ как гипотетическое значение локального профиля D_m, при котором оценка (1) не является завышенной, т. е. неравенство обращается в равенство:

$$\widehat{D}_m = Q_{\varepsilon,m}(\mu, X^L) \;/\; H\tbinom{m\ s_1}{L\ \ell}, \quad m = 0, \ldots, L.$$

Эту величину легко измерить эмпирически, если в функционале $Q_{\varepsilon,m}$ заменить сумму по всем разбиениям N суммой по некоторому подмножеству $N' \subset N$ (в методе Монте-Карло N' — случайное подмножество).

Наконец, *эффективный локальный коэффициент разнообразия* определим как $\widehat{\Delta}_L^\ell = \widehat{D}_0 + \cdots + \widehat{D}_L$. Эта величина показывает, какое значение должен был бы принимать локальный коэффициент разнообразия, чтобы верхняя оценка не была завышенной. Данная методика измерения завышенности существенно уточняет методику, ранее предложенную в [3].

Измерение профилей разнообразия: эксперименты и выводы

Алгоритмы поиска закономерностей, основанные на непосредственном переборе предикатов, очень удобны для эмпирического исследования завышенности сложностных оценок, поскольку: (а) глобальный коэффициент разнообразия (*функция роста* по Вапнику) вычисляется по эффективной рекурсивной формуле [3]; (б) локальный коэффициент оценивается снизу числом различных закономерностей, найденных методом μ на подвыборках $\{X_n^\ell \colon n \in N'\}$. Результаты сравнения этих величин с эффективным локальным коэффициентом приведены в Таблице 1.

Эффективный локальный коэффициент всегда оказывался существенно меньшим длины выборки L. Это означает, что при фиксиро-

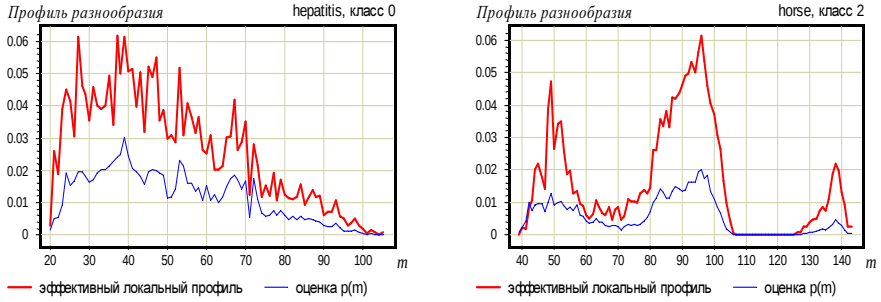


Рис. 1. Сравнение эффективного профиля $\widehat{D}_m$ и нормированного локального профиля $p(m)$ на двух задачах: hepatitis и horse.

ванных X^L, y^* и μ *ёмкость* (VC-dimension) эффективно используемого множества закономерностей никогда не превышает единицы.

Эмпирические нижние оценки локальных коэффициентов завышены, как минимум, на три порядка. Ни один из известных на сегодня подходов, включая наиболее точные [4], не способен дать оценки коэффициентов разнообразия порядка 10^1–10^2.

Интересные результаты дало сравнение эффективного профиля $\widehat{D}_m$ с нижней оценкой локального профиля $\tilde{D}_m$, подсчитанной как число различных закономерностей из $\{\mu X_n^\ell \colon n \in N'\}$, допускающих m ошибок на X^L. Практически во всех задачах оказалось, что нормированный локальный профиль $p(m) = \tilde{D}_m/(\tilde{D}_0 + \cdots + \tilde{D}_L)$ является лишь слегка заниженной оценкой эффективного профиля $\widehat{D}_m$ и, как правило, сильно с ним коррелирует (Рис. 1). Данному факту пока не найдено объяснения.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00877, № 07-07-00181 и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

[1] *Cohen W. W., Singer Y.* A simple, fast and effective rule learner // Proc. of the 16 National Conference on Artificial Intelligence. — 1999. — Pp. 335–342.

[2] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Мат. вопр. киберн. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[3] *Воронцов К. В., Ивахненко А. А.* Эмпирические оценки локальной функции роста в задачах поиска логических закономерностей // *Искусственный Интеллект.* — Донецк, 2006. — С. 281–284.

[4] *Langford J.* Quantitatively tight sample complexity bounds. — 2002. — Carnegie Mellon Thesis.

Способы построения оптимальной вероятностной модели систем распознавания

Капустий Б. Е., Русын Б. П., Таянов В. А.
vtayanov@ipm.lviv.ua
Украина, Львов, Физико-механический институт им. Г. В. Карпенка НАН Украины

Актуальность задачи построения математической модели систем распознавания (СР) состоит в том, что она позволяет исследовать эту систему, не реализуя её в полном объёме. Определение параметров проводится на основании обучающей выборки. Оптимальность модели определяется её точностью и скоростью вычисления параметров [3]. Поэтому важным является применение дифференциального подхода, дающего возможность определить вероятность правильного распознавания отдельно тестируемого образа. Этот подход даёт возможность построить оптимальный вариант модели СР в условиях малых выборок [4]. Математическую модель СР можно представить в виде некоторого функционала

$$M = R(f_x, n, s, t), \tag{1}$$

где f_x — обобщённый классификатор (далее — просто классификатор); n — количество классов; s — размер класса; t — размер доверительного интервала. Модель классификатора f_x в общем виде представляется как

$$f_x = f(\psi^{[k]}, L_{xy}, h_x), \tag{2}$$

где $\psi^{[k]}$ — фрагменты функций признаков; L_{xy} — метрика в пространстве признаков; h_x — решающая функция или правило.

Если для оптимизации модели классификатора использовать последовательный анализ, а в качестве параметра оптимизации — средний размер класса в виде $s = f(\psi^{[k]}, L_{xy}, h_x)$, то задача оптимизации представляется следующим образом:

$$\underset{\psi^{[k]}, L_{xy}, h_x}{\arg\min} \; f(\psi^{[k]}, L_{xy}, h_x) = \min(s). \tag{3}$$

Модель СР включает модель классификатора с параметрами, а также компоненты, влияющие на достоверность результатов распознавания. При оптимизации модели СР важно отдельно учесть влияние на достоверность распознавания таких факторов, как мера расстояний между образами и размеры доверительного интервала и класса, связанные между собой. Основная трудность при исследовании указанных зависимостей состоит в том, что существуют влияния компонент СР как одной на другую, так и совместно — на функционал (1). Всё это усложняет построение моделей различного назначения.

Рассмотрим выражение мер расстояний между векторами признаков **x** и **y**, используемых в теории распознавания [5], через меру Манхеттена — простую линейную меру с весовыми коэффициентами a_i:

$$d(x,y) = \sum_{i=1}^{n} a_i |x_i - y_i|, \tag{4}$$

где $d(x,y)$ — произвольная мера расстояний между векторами **x** и **y**.

Меру расстояний Минковского, как наиболее обобщённую меру, используемую в теории распознавания образов, можно представить в виде

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n} a_i |x_i - y_i|\right)^{\frac{1}{p}} = C(p) \sum_{i=1}^{n} a_i |x_i - y_i|, \tag{5}$$

где $C(p) = \sum_{i=1}^{n} a_i |x_i - y_i|)^{\frac{1-p}{p}}$; $a_i = |x_i - y_i|)^{p-1}$; $p > 0$.

Из приведенного выше следует, что произвольная метрика — это фильтр в пространстве признаков, т. е. она устанавливает веса признаков при использовании решающих правил. Вес определённого признака должен быть пропорционален приращению одного из показателей при его добавлении в общее признаковое множество, используемое для дискриминации классов: вероятности правильного распознавания, среднего размера класса, дивергенции между классами или дискриминанта Фишера [1, 4, 5]. Можно использовать и другие показатели, однако способ их применения должен быть одинаковым. Если признак не даёт приращения соответствующего показателя или ухудшает его, то значение веса соответствующего признака следует принять равным нулю. Таким образом, путём дополнительного уменьшения количества признаков можно ускорить процесс распознавания, не ухудшая его качественных характеристик. Проблема оптимизации набора признаков и выбора вида метрики решена однозначно с помощью взвешенных признаков и простой линейной меры подобия между образами с весовыми коэффициентами. Задача селекции признаков в этом случае решается частично. Определяется субмножество признаков из генеральной совокупности, выбираемой при помощи того или иного алгоритма (например, ряда ортогональных преобразований). Этот алгоритм в свою очередь должен удовлетворять определённым требованиям относительно селекции признаков — таким, как минимизация энтропии образов класса или максимум дивергенции между классами. Указанным требованиям удовлетворяет метод главных компонент [5].

Последним параметром, используемым в модели, является решающая функция или правило. Условно все решающие функции можно разделить на те, что работают в признаковом пространстве и те, которые

строятся на основании функции расстояний. В признаковом пространстве, например, применяют байесовский классификатор, линейный дискриминант Фишера, метод опорных векторов, и др. В многомерном признаковом пространстве значительно усложняется процедура принятия решения при использовании этих решающих правил. Это особенно нежелательно в случаях, когда распознавание проводится непрерывно для серии образов, поступающих в блок распознавания соответствующей системы. Поэтому при практической реализации СР, работающих с достаточно большиим сериями изображений, используют решающие правила, построенные на основании функции расстояний. Принято использовать два решающих правила: по минимуму расстояния от ближайшего (1NN) и k ближайших соседей (kNN). Хотя 1NN правило наиболее простое, оно характеризуется наименьшими показателями вероятности при принятии решений. Поэтому целесообразно использовать kNN правило. При этом задача сводится к выбору значения k, оптимального для принятия решения в пределах доверительного интервала, соответствующего списку возможных претендентов. От размера класса также зависит размер доверительного интервала, в котором принимается решение. Определение условий, при которых результаты принятия решения на основании оптимального байесовского решающего правила с одной стороны, и 1NN или kNN правил — с другой, совпадают или близки, даёт возможность использовать наиболее простое решающее правило при сохранении качественных свойств процедуры принятия решения.

Литература

[1] *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* Распознавание. Математические методы. Программная система. Практические применения — Москва: Фазис, 2005. — 159 с.

[2] *Капустий Б. Е., Русын Б. П., Таянов В. А.* Новый подход к определению вероятности правильного распознавания объектов множеств // УСиМ. — 2005. — № 2. — С. 8–13.

[3] *Kapustiy B. O., Rusyn B. P., Tayanov V. A.* Tayanov Comparative analysis of different estimates of Recognition Probability // Journal of Automation and Information Sciences. — 2006. — Issue 8. — P. 8–16.

[4] *Kapustiy B. O., Rusyn B. P., Tayanov V. A.* Classifier optimization in small sample size condition // Avtomatika i vychislitel'naya tekhnika. — 2006. — vol. 40, Issue 5. — P.25–32.

[5] *Webb R. A.* Statistical Pattern recognition. — John Wiley & Sons Inc, 2nd ed., 2002.

Учет двух наборов взаимно зависимой информации об относительной важности критериев в задачах многокритериального выбора

Климова О. Н.

Klimova_O@rambler.ru

Санкт-Петербург, Санкт-Петербургский Государственный Университет

Одной из значимых проблем в области многокритериального выбора является проблема сужения множества Парето, поскольку оно зачастую оказывается достаточно широким. Как правило, дальнейшее сужение области поиска наилучшего решения происходит на основе дополнительной информации, получаемой от лица, принимающего решения (ЛПР).

В данной работе рассматривается задача, в которой дополнительная информация представляет собой количественную информацию об относительной важности критериев [2]. Она заключается в том, что выделяются группы критериев и определяется важность первой группы относительно второй. Количественно важность выражается парой наборов числовых параметров. Первый набор содержит максимальные величины выигрышей по каждому из критериев более важной группы, в том случае, если ЛПР сделает уступки по каждому из критериев менее важной группы (второй набор параметров содержит величины уступок). Более того, оказывается, что вторая группа критериев, в свою очередь, важнее первой. В данной ситуации ЛПР идет на взаимные уступки по нескольким критериям, ради прибыли по каждому из них.

Информация подобного рода, т. е. когда группа критериев A важнее группы B, а группа критериев B, в свою очередь, важнее A (причем A и B непустые и взаимно непересекающиеся группы) является взаимно зависимой [2].

Пусть X — множество возможных решений. Предпочтения ЛПР выражаются при помощи набора критериев $f_1, \ldots, f_m,\ m \geqslant 2$, образующих векторный критерий $f(x) = (f_1(x), \ldots, f_m(x))$, и бинарного отношения строгого предпочтения $\succ_X$, заданного на X. Множество выбираемых решений обозначим через $Sel(X)$.

Наряду с множествами X и $Sel(X)$ будем использовать множества возможных $Y = f(X) \subset R^m$ и выбираемых $Sel(Y) = f(Sel(X))$ векторов.

Дополнительная информация состоит из следующих двух сообщений: 1) группа критериев A важнее группы критериев B с заданными положительными параметрами $w'_i\ (\forall i \in A),\ w'_j\ (\forall j \in B)$ и группа критериев B важнее группы критериев A с заданными положительными параметрами $\gamma'_j\ (\forall j \in B),\ \gamma'_i\ (\forall i \in A)$; 2) группа A важнее группы критериев C с положительными параметрами $w''_i\ (\forall i \in A),\ w''_k\ (\forall k \in C)$ и группа C важнее группы критериев A с положительными параметрами $\gamma''_k\ (\forall k \in C)$,

γ_i'' $(\forall i \in A)$. Группы критериев A, B, C непустые и взаимно непересекающиеся. Представленный набор информации обозначим через (И). Таким образом, в рассматриваемой задаче имеются два набора взаимно зависимой информации.

Будем предполагать, что отношение предпочтения удовлетворяет четырем аксиомам, определяющим «разумный» выбор [2].

Аксиома 1 (об исключении доминируемых векторов). Для любой пары векторов $y', y'' \in Y$, удовлетворяющих соотношению $y' \succ_Y y''$, выполнено $y'' \notin Sel(Y)$.

Аксиома 2. Для отношения $\succ_Y$ существует иррефлексивное и транзитивное продолжение $\succ$ на все пространство R^m. Тем самым, отношение $\succ_Y$ является сужением $\succ$ на Y.

Аксиома 3. Каждый из критериев $f_1, \ldots, f_m$ согласован с отношением предпочтения $\succ$.

Аксиома 4. Для любых векторов $y', y'' \in R^m$ таких, что $y' \succ y''$, для любого числа $\alpha > 0$ и произвольного вектора $c \in R^m$ выполняется $\alpha y + c \succ \alpha y'' + c$.

Учет дополнительной информации происходит в два этапа. Сначала необходимо убедиться в непротиворечивости [1] предоставленной информации, а затем по определенным формулам произвести сужение исходного множества Парето.

Критерий непротиворечивости был получен в следующем виде.

Теорема 1. *Набор информации (И) непротиворечив тогда и только тогда, когда существуют номера $i_1 \in A$ и $j \in B$, для которых выполняется неравенство*

$$\frac{w'_{i_1}}{w'_j} > \frac{\gamma'_{i_1}}{\gamma'_j} \tag{1}$$

и существуют номера $i_2 \in A$ и $k \in C$, для которых выполняется неравенство

$$\frac{w''_{i_2}}{w''_k} > \frac{\gamma''_{i_2}}{\gamma''_k}. \tag{2}$$

Учет двух наборов взаимно зависимой информации описанного вида, осуществляется по формулам, полученным в следующей теореме.

Теорема 2. *Пусть отношение предпочтения $\succ$ удовлетворяет аксиомам 1–4 и задана непротиворечивая информация об относительной важности (И), причем неравенства вида (1), (2) выполняются для всех $i \in A$, $j \in B$, $k \in C$. Тогда для множества выбираемых решений $Sel(X)$ имеют место включения*

$$Sel(X) \subset P_g(X) \subset P_f(X),$$

где $P_g(X)$ — множество Парето в новой задаче многокритериального выбора с векторным критерием g размерности $q = m - (|A| + |B| + |C|) + 4 \cdot |A| \cdot |B| \cdot |C|$ и компонентами

$$
\begin{aligned}
g_{ijk} &= w'_j w''_k f_i(x) + w'_i w''_k f_j(x) + w'_j w''_i f_k(x), && \forall i \in A,\ \forall j \in B,\ \forall k \in C;\\
g_{ikj} &= \gamma'_j \gamma''_k f_i(x) + \gamma'_i \gamma''_k f_j(x) + \gamma'_j \gamma''_i f_k(x), && \forall i \in A,\ \forall j \in B,\ \forall k \in C;\\
g_{jki} &= w'_j \gamma''_k f_i(x) + w'_i \gamma''_k f_j(x) + w'_j \gamma''_i f_k(x), && \forall i \in A,\ \forall j \in B,\ \forall k \in C;\\
g_{kji} &= \gamma'_j w''_k f_i(x) + \gamma'_i w''_k f_j(x) + \gamma'_j w''_i f_k(x), && \forall i \in A,\ \forall j \in B,\ \forall k \in C;\\
g_s &= f_s, && \forall s \in I \setminus (A \cup B \cup C).
\end{aligned}
$$

Таким образом, согласно теореме 2, множество Парето $P_g(X)$, полученное относительно «нового» векторного критерия g, уже множества $P_f(X)$ и является оценкой для искомого множества $Sel(X)$.

Наконец, отметим, что теорема 2 сводится к частному случаю, если дополнительная информация (И) состоит только из одного набора взаимно зависимой информации (группа критериев A важнее группы критериев B, а группа критериев B важнее группы A) [1].

Литература

[1] *Климова О. Н., Ногин В. Д.* Учет взаимно зависимой информации об относительной важности критериев в процессе принятия решений // ЖВМиМФ. — 2006.— Т. 46, № 12.— С. 2178–2190.

[2] *Ногин В. Д.* Принятие решений в многокритериальной среде: количественный подход. — 2-е издание. — Москва: Физматлит, 2005. — 176 с.

Новый алгоритм синтеза всех неприводимых многочленов над заданным конечным полем

Леухин А. Н., Бахтин С. А.

inf@marstu.mari.ru

Йошкар-Ола, ГОУ ВПО Марийский гос. тех. университет

Проведен обзор проблемы синтеза неприводимых и примитивных многочленов над заданным конечным полем. Предложен новый быстрый детерминированный алгоритм синтеза всех неприводимых многочленов над конечным полем F_p заданной степени n, основанный на модифицированном быстром методе Крылова для вычисления характеристического многочлена матрицы Фробениуса, сопровождающей неприводимый многочлен над заданным конечным полем.

Теория многочленов степени n от одной переменной, неприводимых над конечными полями F_p, представляет существенный интерес как для исследования алгебраической структуры конечных полей $F_{q=p^n}$,

так и для многочисленных приложений в современной теории передачи информации. Такие многочлены имеют большое значение при синтезе шумоподобных кодовых последовательностей, в теории помехоустойчивого кодирования, в криптографии при решении задачи дискретного логарифмирования (к которой сводится задача логарифмирования на эллиптической кривой), в теории кольцевых счетчиков, и т. д.

Первое крупное исследование о неприводимых многочленах от одной переменной над полем F_q проведено в работе [1]. Фундаментальный обзор результатов по теории конечных полей, включающий и теорию неприводимых многочленов, приводится в работе [2]. Однако, несмотря на достигнутые успехи в теории синтеза неприводимых многочленов, имеется ряд важнейших проблем, которые до сих пор не поддаются решению. Одной из них является проблема построения неприводимых многочленов заданной степени в явном виде, а также определения периодов элементов поля — корней этих многочленов.

По существу, все подходы к синтезу неприводимых многочленов можно разделить на три большие группы. К первой группе можно отнести аналитические методы построения, позволяющие в явном виде сразу записать выражения для неприводимых многочленов. В роли таких многочленов чаще всего выступают полиномы с малым числом слагаемых — двучлены, трехчлены и четырехчлены.

В работе [2] приводятся теоремы для неприводимых двучленов и трехчленов. Конкретные примеры аналитических выражений для неприводимых многочленов, удовлетворяющие приведенным теоремам, приводятся в работе [3]. Конструкции неприводимых многочленов над полем F_2 степени $4 \cdot 3^k \cdot 5^l$, над полем F_3 степени $4 \cdot 2^k \cdot 5^l$, над полем F_p $(p > 3)$ степени $2 \cdot 2^k \cdot 3^l$ можно найти в работе [4]. Другие аналитические конструкции неприводимых многочленов над конечными полями в явном виде приведены в работах [5, 6].

К сожалению, в явном виде не удается получить выражения для неприводимых многочленов произвольной степени n над полем F_q. Кроме того, в явном виде невозможно записать все неприводимые многочлены заданной степени n.

Следующая группа методов синтеза основана на идее факторизации многочлена произвольно заданной степени n в конечном поле F_q . Неприводимость многочлена устанавливается по результатам факторизации. Первые существенны результаты получены в работе [7], опираясь на которые, в работе [8] синтезирован улучшенный и эффективный в вычислительном плане метод. В дальнейшем на основе методов факторизации появились вероятностные [9] и детерминированные [10] алгоритмы-тесты

на неприводимость многочлена произвольной степени n над полем F_q, позволяющие решать задачу за полиномиальное время.

В третью группу методов синтеза неприводимых полиномов входят алгоритмы построения неприводимых и примитивных многочленов, использующих алгебраическую структуру и внутреннее строение полей Галуа. В отличии от двух предыдущих случаев, данные методы позволяют синтезировать сразу все возможные неприводимые или примитивные многочлены степени n над F_q. Первый алгоритм — алгоритм решета — является прямым методом синтеза неприводимых многочленов [2], и для его реализации нет необходимости в использовании «начального» неприводимого многочлена. Однако этот алгоритм имеет низкую вычислительную производительность и может использоваться для малых размерностей степени многочлена n и характеристики p поля. Второй алгоритм [2] основан на свойствах минимальных многочленов степени n поля F_q. Для его реализации требуется один «начальный» неприводимый полином степени n над F_q для задания внутренней структуры поля. В работе [11] описан метод построения новых неприводимых многочленов над полем, исходя из данного неприводимого многочлена.

В ходе исследовательской работы нами был получен алгоритм синтеза всех возможных примитивных многочленов степени n над полем F_p. На первом шаге формируется матрица Фробениуса, сопровождающая «начальный» примитивный многочлен. На втором шаге определяются подмножества коэффициентов p-сопряженных элементов поля F_{p^n}, и на их основе формируются множество коофициентов, содержащее любой из элементов каждого подмножества. На третьем шаге исходная матрица возводится в степень коэффициента множества. На четверттом шаге с использованием метода Крылова формируется матрица, по которой будет вычисляться характеристический многочлен в конечном поле. Для ее формирования в качестве нулевого вектора используется вектор вида $u_0 = 1,\ u_1 = 0,\ u_2 = 0,\ \ldots,\ u_{n-1} = 0$. На пятом шаге с помощью модифицированного метода исключения Гаусса, позволяющего проводить триангуляцию матрицы даже с нулевыми элементами на главной диагонали, с учетом выполнения операций деления в конечном поле F_p, формируется матрица Крылова. Последний элемент $A_{n,n}$ этой матрицы представляет собой искомый неприводимый многочлен над полем F_{p^n}. Отметим, что процедуры возведения матрицы в степень выполняются дихотомическим способом.

Отличие предлагаемого алгоритма от рассмотренных выше заключается в том, что для его реализации не требуется введение трудоёмких в вычислительном плане операций умножения и деления многочленов в конечном поле F_{p^n}. Все операции выполняются в поле F_p, при этом

возрастает производительность и снижаются требования к объёму используемой памяти.

Программная реализация предлагаемого в работе быстрого алгоритма показала удовлетворительные результаты при сравнительном анализе быстродействия с аналогичными специализированными математическими продуктами GAP 4.4.6 группы разработчиков Gap Group и MAGMA 2.12 группы разработчиков Computational Algebra Group.

С помощью такого быстрого алгоритма синтеза, реализованного на современной элементной базе, могут быть успешно решены задачи помехоустойчивого приёма информации, кодового разделения каналов передачи информации и задач криптографии.

Работа выполнена при поддержке РФФИ, проект №07-07-00285, и гранта Президента РФ МД-63.2007.9.

Литература

[1] *Dickson L. E.* Linear Groups with an Exposition of the Galois Field Theory. — New York: 1958.

[2] *Лиддл Р., Ниддерайтер Г.* Конечные поля. — М.: Мир. Т.1,2. 1988

[3] *Gao Sh., Panario D.* Foundations of Computational Mathematics. — Springer. 1997. P.346.

[4] *Shparlinski I.* Apl. Alg. Eng. Comm. 1993. v.4. P.263.

[5] *Gao S., Mullen G.* J. Number Theory. 1994. v.49. P.118.

[6] *Menezes A., Blake I., Gao X., Mullin R., Vanstone S., Yaghoobian T.* Applications of Finite Fields. Kluwer Academic Publisher. 1993.

[7] *Butler M.* Quart. J. Math. Oxford Ser. (2). 1954. v.5. P.102.

[8] *Berlekamp E. R.* Math. Comp. 1970. v.24. P.713-735.

[9] *Rabin M. O.* SIAM J. Comp. 9. 1980. P.273.

[10] *Shoup V.* J. Symb. Comp. 20. 1996. P.363.

[11] *Варшамов Р. Р., Антонян А. М.* Докл. АН АрмССР. т.66. №4. 1978. С.197.

Эффективный ранг и эффективная размерность в вейвлет-анализе данных

Мондрус О. В.
olya@cmp.phys.msu.ru
Москва, МГУ им. М. В. Ломоносова, физический факультет

В теории измерительно-вычислительных систем (ИВС) [1] рассматривается модель $[A, \Sigma]$ измерения, выполненного по схеме

$$\xi = Af + \nu, \tag{1}$$

в которой $f \in \mathcal{R}_m$ — измеряемый сигнал, $A\colon \mathcal{R}_m \to \mathcal{R}_n$ — линейный оператор, моделирующий измерительный прибор, ν — случайный элемент $\mathcal{R}_n$ с математическим ожиданием $\mathsf{E}\nu = 0$ и ковариационным оператором $\Sigma : \mathcal{R}_n \to \mathcal{R}_n$, определенным равенством $\Sigma x = \mathsf{E}\nu(x, 0),\ x \in \mathcal{R}_n$.

Пусть $\Pi_k\colon \mathcal{R}_m \to \mathcal{R}_k$ — ортогональный проектор на k-мерное линейное подпространство $\mathcal{L}_k \subset \mathcal{R}_m$ и

$$\inf_{R:\mathcal{R}_n\to\mathcal{L}_k} \sup_{f\in\mathcal{R}_m} \mathsf{E}\|R\xi - \Pi_k f\|^2 < \infty.$$

Определение 3. *Эффективным рангом модели* $[A, \Sigma]$ *называется функция* $\rho(\varepsilon)$*, значение которой равно максимальной размерности ортогональной составляющей* f*, которую можно оценить со с.к. погрешностью, не превосходящей* ε*:*

$$\rho(\varepsilon) = \max\{k, h(\Pi_k) \leqslant \varepsilon\}, \quad \varepsilon \geqslant 0, \tag{2}$$

Определение 4. *Эффективной размерностью данных измерений, в том числе выполненных по схеме* (1)*, называется функция* $\zeta\colon [0, \infty) \to \{1, 2, \ldots\}$*, значение* $\zeta(\varepsilon)$ *которой равно минимальной размерности линейного подпространства* $\mathcal{R}^{(\varepsilon)} \subset \mathcal{R}$ *ортогональных составляющих измерений, которые приближают измерения с погрешностью, не превосходящей заданного значения* $\varepsilon \geqslant 0$ [1].

В докладе понятия эффективного ранга модели и эффективной размерности данных измерений применены для построения оптимального вейвлет-представления [2] данных измерений и их редукции.

Работа выполнена при поддержке РФФИ, грант № 05-01-00532-а.

Литература

[1] *Пытьев Ю. П.* Методы математического моделирования измерительно-вычислительных систем. — М.: ФизМатЛит, 2007.

[2] *Daubechies I.* Ten lectures on Wavelets. — SIAM, 1991 (*Добеши И.* Десять лекций по вейвлетам. — М.–Ижевск, 2001).

Об эффективности эмпирических функционалов качества решающей функции

Неделько В. М.
nedelko@math.nsc.ru
Новосибирск, Институт математики СО РАН

В работе предложен способ сравнения эффективности различных функционалов, оценивающих риск решающей функции по той же выборке, по которой построена функция, что позволяет находить в некотором смысле оптимальные функционалы.

Задача построения решающей функции

Пусть X — пространство значений переменных, используемых для прогноза, а Y — пространство значений прогнозируемых переменных, и пусть C — множество всех вероятностных мер на заданной σ-алгебре подмножеств множества $D = X \times Y$. При каждом $c \in C$ имеем вероятностное пространство: $\langle D, B, P_c \rangle$, где B — σ-алгебра, $P_c[D]$ — вероятностная мера. Параметр будем называть *стратегией природы*.

Решающей функцией называется соответствие $f\colon X \to Y$.

Качество принятого решения оценивается заданной функцией потерь $L : Y^2 \to [0, \infty)$. Под риском будем понимать средние потери:

$$R(c, f) = \int_D L(y, f(x))\, dP_c[D].$$

Пусть $\nu = \left\{(x^i, y^i) \in D \;\middle|\; i = 1, \ldots, N\right\}$ — случайная независимая выборка из распределения $P_c[D]$. Эмпирический риск определим как средние потери на выборке:

$$\widetilde{R}(\nu, f) = \frac{1}{N} \sum_{i=1}^{N} L(y^i, f(x^i)).$$

Заметим, что значение риска зависит от стратегии природы — распределения, которое неизвестно. Функционал скользящего экзамена определяется как

$$\breve{R}(\nu, Q) = \frac{1}{N} \sum_{i=1}^{N} L(y^i, f_{Q,\nu'_i}(x^i)),$$

где $\nu'_i = \nu \backslash \{(x^i, y^i)\}$ — выборка, получаемая из ν удалением i-го наблюдения, $Q\colon \{\nu\} \to \Phi$ — алгоритм построения решающих функций, $f_{Q,\nu}$ — функция, построенная по выборке ν алгоритмом Q, Φ — заданный класс решающих функций.

Доверительный интервал для риска

Доверительный интервал для R будем задавать в виде $[0, \hat{R}(\nu)]$. Здесь мы ограничиваемся односторонними оценками, поскольку на практике для риска важны именно оценки сверху. Таким образом, в данном случае построение доверительного интервала эквивалентно выбору функции $\hat{R}(\nu)$, которую будем называть оценочной функцией или просто оценкой (риска).

При этом должно выполняться условие:

$$\forall c \quad P_c\big(R \leqslant \hat{R}(\nu)\big) \leqslant \eta,$$

где η — заданная доверительная вероятность.

При построении оценок риска первая проблема, которую нужно решить, это сравнение качества различных оценок.

Можно положить, что задан функционал качества $K\big(F_{c,\hat{R}}(\cdot)\big)$, где $F_{c,\hat{R}}(\cdot)$ — функция распределения оценки $\hat{R}(\nu)$. Выбор данного функционала, так же как и выбор функции потерь, определяется практическими соображениями. Простейшим вариантом такого функционала является математическое ожидание.

При фиксированной стратегия природы c функционал K позволяет сравнивать качество оценок риска и находить оптимальную оценку.

Однако на практике распределение c неизвестно, а оценки, оптимальной при всех распределениях, может не существовать. В этом случае естественным является поиск множества Парето недоминируемых оценок.

Эмпирические функционалы качества

Известные на данный момент оценки риска (напр. [1]) строятся не как функция непосредственно выборки, а через композицию $\hat{R}(\nu) = \hat{R}_e\big(\dot{R}(\nu)\big)$, как функция значений некоторого эмпирического функционала $\dot{R}$, в качестве которого обычно выступает эмпирический риск или скользящий экзамен.

Эмпирический функционал здесь выступает в роли точечной оценки риска, на основе которой строится интервальная оценка.

В докладе представлены исследования эффективности функционалов эмпирического риска и скользящего экзамена для частных задач. При этом под эффективностью понимается, насколько хорошая интервальная оценка риска может быть построена на основе данного функционала.

Определение 1. *Оценочную функцию $\hat{R}(\nu)$ назовем согласованной с эмпирическим функционалом $\dot{R}$, если для выборок ν_1 и ν_2 одинакового объема*

$$\dot{R}(\nu_1) > \dot{R}(\nu_1) \Rightarrow \hat{R}(\nu_1) \geqslant \hat{R}(\nu_2).$$

Достаточно естественным представляется ограничиться рассмотрением только таких оценочных функций, которые согласованы с функционалами эмпирического риска и скользящего экзамена. Это означает, что оценка вероятности ошибки не должна убывать при увеличении значения эмпирического функционала.

Данное условие позволяет резко сузить пространство поиска при нахождении Парето-оптимальных оценочных функций.

Работа выполнена при поддержке РФФИ, проект № 07-01-00331-а, и СО РАН, проект № 7.

Литература

[1] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — Москва: Наука, 1974. — 415 с.

[2] *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: Изд-во ИМ СО РАН, 1999. — 211 с.

[3] *Nedelko V. M.* Estimating a Quality of Decision Function by Empirical Risk // LNAI 2734. Machine Learning and Data Mining in Pattern Recognition. Third International Conference, MLDM 2003, Leipzig. Proceedings. Berlin: Springer-Verlag, 2003. — P. 182–187.

[4] *Неделько В. М.* Оценивание смещения эмпирического риска для линейных классификаторов // Таврический вестник информатики и математики. Изд-во НАН Украины. — 2004. — № 1. — С. 47–53.

[5] *Неделько В. М.* Оценка смещения эмпирической оценки риска решающей функции // Докл. всеросс. конф. «Математические методы распознавания образов», ММРО-11 — Москва: ВЦ РАН, 2003. — С. 148–150.

Принятие решений при многих критериях на основе нечёткой информации об относительной важности критериев

Ногин В. Д.

noghin@home.eltel.net

Санкт-Петербург, Санкт-Петербургский государственный университет

Излагаются основы теории выбора решений при наличии нескольких числовых критериев в условиях нечёткого отношения предпочтения лица, принимающего решение (ЛПР), и заданном конечном наборе количественной информации об этом нечётком отношении предпочтения.

Введение

Принципиальная сложность задач выбора при многих критериях заключается в невозможности априорного определения того, что называть

наилучшим решением. Каждое ЛПР имеет право вкладывать свой смысл в это понятие. Более того, небольшое изменение обстоятельств, при которых осуществляется выбор, может привести к изменению смысла наилучшего решения. Это говорит о непродуктивности использования для решения задач выбора при многих критериях традиционного подхода, сложившегося десятки лет назад в области оптимизации с одним критерием и предполагающего обязательное формальное введение понятия оптимального решения.

В отличие от традиционного подхода в последние два десятилетия активно развивается методология, не предполагающая для своей реализации наличия строгого определения выбираемого решения (см., например, [1]). Ее суть заключается в получении тех или иных оценок сверху для заранее неизвестного множества выбираемых решений на основе определённых общих свойств поведения ЛПР в процессе принятия решений с учётом количественной информации об отношении предпочтения ЛПР.

Эта методология к настоящему времени приобрела вполне определенные контуры [1]. Ее фундамент составляет знаменитый принцип Эджворта-Парето, а основное содержание образуют результаты, показывающие, каким образом следует учитывать ту или иную количественную информацию об отношении предпочтения ЛПР для обоснованного сужения множества Парето. Нередко такую информацию удобно интерпретировать в терминах относительной важности критериев.

В соответствии с этой методологией любой выбор, подчиняющийся нескольким аксиомам, характеризующим в определённой степени «разумное» поведение ЛПР в процессе принятия решений, следует осуществлять в пределах множества Парето, которое можно построить с помощью нового векторного критерия, вычисляемого на основе исходного векторного критерия и имеющейся количественной информации об отношении предпочтения ЛПР. Тем самым, конструируется определённая оценка сверху для неизвестного множества выбираемых решений, которая является более точной, чем множество Парето. Этот подход можно охарактеризовать и таким образом: наличие указанной информации дает возможность сузить исходное множество Парето.

Нередко при выявлении информации об отношении предпочтения удается получить соответствующие количественные данные лишь в специфической нечёткой форме, когда предпочтительность того или иного решения по сравнению с другими оценивается с субъективной степенью уверенности, способной изменяться в некоторых пределах. В таких ситуациях оказывается возможным применить аппарат теории нечетких множеств и отношений [5].

Выяснилось, что разрабатываемая методология сужения множества Парето допускает распространение на более общий случай нечеткого отношения, а также нечеткого множества возможных решений (см. [2, 4]). Основы подобного распространения были заложены ещё в работе автора [3]. При этом использование нечётких множеств и отношений дает возможность разработать более гибкий аппарат, который можно применять при решении прикладных задач многокритериального выбора достаточно широкого класса.

Обзор основных результатов

Рассматривается задача многокритериального выбора, постановка которой содержит абстрактное и в общем случае нечёткое *множество возможных* (*допустимых*) *решений*, на котором заданы числовые функции (*критерии*). Относительно нечёткого *отношения предпочтения* ЛПР считаются известными некоторые его общие свойства, а также пары несравнимых критериальных оценок, в которых одна оценка предпочтительнее другой с некоторой степенью уверенности, оцениваемой числом в пределах от 0 до 1. Собственно выбор заключается в использовании всей имеющейся в наличии информации для указания такого в общем случае нечёткого подмножества множества возможных решений (так называемого *множества выбираемых решений*), которое является наиболее приемлемым для ЛПР.

При условии выполнения нескольких аксиом «разумного» поведения ЛПР формулируется так называемый *нечёткий принцип Эджворта-Парето*, согласно которому множество выбираемых решений должно быть подмножеством множества Парето. Этот принцип даёт возможность в некоторых случаях сужать область поиска решений, подлежащих выбору, до множества Парето. Иными словами, в соответсвии с этим принципом за пределами множества Парето выбираемых решений быть не должно.

Поскольку множество Парето нередко оказывается достаточно широким, для дальнейшего его обоснованного сужения предлагается использовать информацию о нечётком отношении предпочтения в форме конечного набора пар несравнимых оценок (векторов), в которых одна оценка предпочтительнее другой с некоторой степенью уверенности, оцениваемой числом а пределах от 0 до 1. Вводится понятие непротиворечивого (совместного) набора нечёткой информации об отношении предпочтения и устанавливается критерий непротиворечивости подобного набора. В соответствии с этим критерием проверка непротиворечивости сводится к решению набора задач линейного программирования специального вида.

Главными теоретическими результатами предлагаемого подхода являются теоремы, показывающие, каким образом следует учитывать име-

ющуюся нечёткую информацию для сужения множества Парето. В этих теоремах строятся определенные оценки сверху для неизвестного нечёткого множества выбираемых решений с использованием специальных множеств Парето относительно специальным образом построенных векторных критериев. Показывается, что нередко этот учёт может быть сведён к решению нескольких задач построения множества Парето.

Следует отметить, что предлагаемый подход может быть использован при решении задач многокритериального выбора с любыми множествами допустимых решений и произвольными критериями.

Работа выполнена при поддержке РФФИ, проект № 05-01-00310.

Литература

[1] *Ногин В. Д.* Принятие решений при многих критериях — 2-е издание. — Москва: ФИЗМАТЛИТ, 2005. — 176 с.

[2] *Ногин В. Д.* Принцип Эджворта-Парето и относительная важность критериев в случае нечёткого отношения предпочтения // ЖВМиМФ. — 2003. — Т. 43. № 11. — С. 1676–1686.

[3] *Ногин В. Д.* Upper estimate for fuzzy set of nondominated solutions // Fuzzy Sets and Systems. — 1994. — Vol. 67. — P. 303–315.

[4] *Ногин В. Д., Волкова Н. А.* Эволюция принципа Эджворта-Парето // Таврический вестник информатики и математики. — 2006, — № 1. — С. 21–33.

[5] *Zadeh L. A.* Fuzzy sets // Informat. Control. — 1965. — Vol. 8. — P. 338–353.

Экспертное оценивание нечеткого элемента

Пытьев Ю. П.

pytyev@phys.msu.su

Москва, МГУ им. М. В. Ломоносова, физический факультет

В докладе рассмотрена задача эмпирического построения возможности на основе заключений экспертов. Характерный момент, определяющий идею построения возможности, состоит в том, что каждый эксперт решает задачу в своей шкале, и, следовательно, метод построения возможности, учитывающий мнения всех экспертов, должен быть инвариантен относительно изотонных автоморфизмов их шкал [1].

Речь пойдет об экспертных оценках распределения нечеткого элемента ξ, принимающего значения в $X = \{x_1, \dots, x_n\}$, неформальная модель которого в той или иной степени известна экспертам. Эксперты должны оценить возможности равенств $\xi = x_1, \dots, \xi = x_n$ в своих шкалах.

На самом деле каждому эксперту достаточно оценить лишь упорядоченность значений возможностей

$$\mathrm{p}_j \triangleq \mathrm{P}(\xi = x_j), \quad j = 1, \dots$$

Пусть $\mathrm{p}_{i_1}, \mathrm{p}_{i_2}, \ldots, \mathrm{p}_{i_n}$ — упорядоченная по невозрастанию последовательность значений возможностей равенств $\xi = x_1, \ldots, \xi = x_n$, выданная i-м экспертом, $i = 1, \ldots, m$. Сопоставим i-му эксперту перестановку $\pi \triangleq (\pi_i(1), \ldots, \pi_i(n)) \triangleq (i_1, \ldots, i_n)$, упорядочивающую (по его мнению) значения возможностей

$$1 = \mathrm{p}_{i_1} \geqslant \mathrm{p}_{i_2} \geqslant \ldots \geqslant \mathrm{p}_{i_n}, \quad i = 1, \ldots, m. \tag{1}$$

Предположим, что эксперты обязаны использовать только строгое неравенство.

В таком случае расстояние $r(\pi_i, \pi_t)$ между «мнениями» i-го и s-го экспертов можно определить следующим выражением

$$r(\pi_i, \pi_t) = \left(\sum_{k=1}^{n} (\pi_i(k) - \pi_t(k))^2 \right)^{\frac{1}{2}} \tag{2}$$

Пусть в (2) $t_1, \ldots, t_n$ — номера, упорядочивающие значения возможностей, которые указал бы «эталонный» эксперт, выражая мнения всех m экспертов. Тогда перестановка $\pi_t^* \triangleq (\pi_t^*(1), \ldots, \pi_t^*(n))$, отражающая в той или иной степени мнения всех m экспертов, может быть определена как решение задачи

$$\sum_{i=1}^{m} r(\pi_i, \pi_t^*) = \min_{\pi(\cdot)} \sum_{i=1}^{m} r(\pi_i, \pi), \tag{3}$$

в которой минимум вычисляется на множестве всех перестановок $\pi(\cdot)\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$.

Так как

$$\sum_{i=1}^{m} r^2(\pi_i, \pi) = \sum_{i=1}^{m} r^2(\pi_i, \bar{\pi}) + m r^2(\bar{\pi}, \pi), \tag{4}$$

где

$$\bar{\pi}(k) = \frac{1}{m} \sum_{i=1}^{m} \pi_i(k), \quad k = 1, \ldots, n,$$

то задача (3) эквивалентна задаче отыскания перестановки π_t^*, ближайшей к функции[1] $\bar{\pi} : r^2(\bar{\pi}, \pi_t^*) = \min\limits_{\pi} r^2(\bar{\pi}, \pi)$. Если для некоторой перестановки $\hat{\pi}$ выполняется $\bar{\pi}(\hat{\pi}(1)) \leqslant \ldots \leqslant \bar{\pi}(\hat{\pi}(n))$, то очевидно, $\pi_t^* = \hat{\pi}^{-1}$.

В то время как перестановка π_t^* выражает «коллективную» экспертизу, значение первого слагаемого в правой части равенства (4) позволяет судить, насколько ей следует доверять.

[1] $\bar{\pi}(1), \ldots, \bar{\pi}(n)$, вообще говоря, не образуют перестановку $1, \ldots, n$.

Обозначим $\omega \triangleq \{\pi_1, \ldots, \pi_m\}$ — последовательность m перестановок, Π^m — класс всех $(n!)^m$ таких ω, значения $\Pr(\{\omega\}) = (n!)^{-m}$, $\omega \in \Pi^m$, определяющие вероятность на $\mathcal{P}(\Pi^m)$.

Вероятностное пространство $(\Pi^m, \mathcal{P}(\Pi^m), \Pr)$ моделирует заключения m «абсолютно некомпетентных» экспертов, которые взаимно независимо принимают решения наугад.

Статистика $s(\omega) \triangleq \sum\limits_{i=1}^{m} r^2(\pi_i, \bar{\pi})$, $\omega \in \Pi^m$, принимает значения в $S = \{s(\omega), \omega \in \Pi^m\} \triangleq \{s_1, \ldots, s_L\}$ с вероятностями $\mathrm{pr}_l \triangleq \Pr\big(\{\omega \in \in \Pi^m, s(\omega) = s_l\}\big)$, $l = 1, \ldots, L$, которые определим упорядоченными по убыванию: $\mathrm{pr}_1 \geqslant \ldots \geqslant \mathrm{pr}_L$. Если H_0 — гипотеза, согласно которой эксперты «абсолютно некомпетентны», то критерий, отвергающий H_0 ошибочно с вероятностью $\leqslant \varepsilon$, $\varepsilon > 0$, определится критическим множеством $S_\varepsilon \triangleq \{s_{l(\varepsilon)}, s_{l(\varepsilon)+1}, \ldots, s_L\}$, где $l(\varepsilon) = \min\{l, s_l \in S, \mathrm{pr}_l + \ldots + \mathrm{pr}_L \leqslant \varepsilon\}$.

Разумеется, на самом деле H_0 не обязательно свидетельствует о некомпетентности экспертов, но включение $\sum\limits_{i=1}^{m} r^2(\pi_i, \bar{\pi}) \in S_\varepsilon$ означает, что «коллективной» экспертизе π_t^* доверять не следует.

Работа выполнена при поддержке РФФИ, грант № 05-01-00532-а.

Литература

[1] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применения. — М.: ФизМатЛит, 2007.

Математические методы и адаптивные алгоритмы эмпирического построения теоретико-возможностной модели стохастического объекта

Пытьев Ю. П.

pytyev@phys.msu.ru

Москва, МГУ им. М. В. Ломоносова, физический факультет

В докладе представлены результаты исследования проблемы эмпирического построения теоретико-возможностной модели неопределенного стохастического объекта как альтернативы его теоретико-вероятностной модели, в том числе в ситуации, когда эмпирическое построение последней принципиально невозможно.

Стохастическая модель объекта охарактеризована как вероятностное пространство $(\Omega, \mathcal{P}(\Omega), \Pr)$, в котором $\Omega = \{\omega_1, \omega_2, \ldots\}$, $\mathcal{P}(\Omega)$ — класс всех подмножеств Ω, вероятности $\mathrm{pr}_i \stackrel{\mathrm{def}}{=} \Pr(\{\omega_i\})$, $i = 1, 2, \ldots$, определяющие $\Pr(\cdot)\colon \mathcal{P}(\Omega) \to [0, 1]$, попарно различны, а в остальном — произвольны.

Рассмотрены две модели взаимно независимых наблюдений за объектом: $(\Omega^n, \mathcal{P}(\Omega^n), \Pr^{(n)})$, в которой $\Pr^{(n)} \stackrel{\triangle}{=} \Pr \times \ldots \times \Pr$, и $(\Omega^n, \mathcal{P}(\Omega^n), \Pr^{(n)}_{1,\ldots,n})$, в которой $\Pr^{(n)}_{1,\ldots,n} \stackrel{\triangle}{=} \Pr_1 \times \ldots \times \Pr_n,\ n = 1, 2, \ldots$

Во второй модели (в отличие от первой) объект эволюционирует в процессе наблюдений, причем так, что эмпирическое построение его теоретико-вероятностной модели принципиально невозможно.

В каждой модели результатом n наблюдений являются частоты $\nu_1^{(n)}$, $\nu_2^{(n)}, \ldots$ элементарных событий $\omega_1, \omega_2, \ldots,\ n = 1, 2, \ldots$

Для обеих моделей наблюдений рассмотрены:

1. Адаптивные алгоритмы упорядочения и интервального оценивания значений $\mathrm{pr}_1, \mathrm{pr}_2, \ldots$
2. Алгоритм восстановления теоретико-возможностной модели $(\Omega, \mathcal{P}(\Omega), \mathrm{P})$, в которой возможность $\mathrm{P}(\cdot)\colon \mathcal{P}(\Omega) \to [0, 1]$ $\Pr_j$-стохастически измерима, $j = 1, 2, \ldots$, [1].
3. Алгоритм построения модели $(\Omega, \mathcal{P}(\Omega), \mathrm{P})$ и σ-алгебры $\mathcal{A}$ подмножеств Ω, таких, что возможность P и вероятность $\Pr_j$ в известном смысле максимально согласованы на $\mathcal{A}$ между собой, $j = 1, 2, \ldots$[1].

Для первой модели наблюдений алгоритм 1 восстановления упорядоченности

$$\mathrm{pr}_{i_1} > \ldots > \mathrm{pr}_{i_s}, \tag{1}$$

верной с априори заданной вероятностью, не меньшей $1 - \alpha,\ \alpha \in (0, 1)$, ра́вно как и алгоритм построения интервалов, покрывающих истинные значения $\mathrm{pr}_{i_1}, \ldots, \mathrm{pr}_{i_s}$ с вероятностью, не меньшей $1-\alpha$, для каждого $s = 2, 3, \ldots$ завершается на основе данных почти наверное (п.н.) конечного числа наблюдений.

Что касается восстановления Pr-стохастически измеримой возможности P, то, поскольку последняя при условии (1) определяется (с точностью до эквивалентности) конкретной упорядоченностью значений возможностей $\mathrm{p}_{i_1} \geqslant \ldots \geqslant \mathrm{p}_{i_s} \geqslant \ldots$ элементарных событий $\mathrm{p}_i \stackrel{\mathrm{def}}{=} \mathrm{P}(\{\omega_i\})$, $i = 1, 2, \ldots$, а последняя определяется условиями Pr-измеримости[1] P:

$$\mathrm{p}_{i_k} > \mathrm{p}_{i_{k+1}} \Longleftrightarrow f_{i_k} \stackrel{\mathrm{def}}{=} \mathrm{pr}_{i_1} + \ldots + \mathrm{pr}_{i_{k-1}} + 2\mathrm{pr}_{i_k} > 1;$$
$$\mathrm{p}_{i_k} = \mathrm{p}_{i_{k+1}} \Longleftrightarrow f_{i_k} < 1,\ k = 1, 2, \ldots,$$

то задача восстановления P сводится к последовательности задач проверки статистических гипотез [1]

$$\{f_{i_1} > 1 \text{ или } f_{i_1} < 1\}, \ldots, \{f_{i_s} > 1 \text{ или } f_{i_s} < 1\}, \ldots \tag{2}$$

[1]Вероятность Pr предполагается регулярной: $f_{i_k} \neq 1,\ k = 1, 2, \ldots$

Представленный в докладе алгоритм 2 решения задач (2) при любой априори заданной вероятности безошибочной проверки любого конечного числа $s = 1, 2, \ldots$ гипотез завершается на основе данных п.н. конечного числа наблюдений.

Наконец, что касается алгоритма 3, то его актуальность обусловлена тем, что восстановленная алгоритмом 2 возможность любого непустого подмножества Ω может оказаться равной единице и, следовательно, — не отражать вероятностных отличий между событиями из $\mathcal{P}(\Omega)$. В таком случае класс $\mathcal{P}(\Omega)$ должен быть сужен до σ-алгебры $\mathcal{A}$, атомы которой содержат минимальные количества элементарных событий, но достаточно «контрастны» для того, чтобы их возможности были различны.

В докладе рассмотрен алгоритм 3. построения искомой σ-алгебры $\mathcal{A}$, который для верного с априори заданной вероятностью построения любого конечного числа атомов $\mathcal{A}$ требует почти наверное конечного числа данных наблюдений.

Все перечисленные алгоритмы позволяют восстанавливать теоретико-возможностные модели стохастических объектов, эволюционирующих в процессе наблюдений, характер возможных изменений которых ограничен лишь условиями Pr_j-измеримости, $j = 1, 2, \ldots$, восстанавливаемой возможности. Эти ограничения, равно как и требования, гарантирующие п.н. конечность числа наблюдений для завершения алгоритмов с гарантированной вероятностью безошибочных решений, рассмотрены в докладе [2].

Работа выполнена при поддержке РФФИ, грант № 05-01-00532-а.

Литература

[1] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применения — М.: Физматлит, 2007.

[2] *Пытьев Ю. П.* Математические методы и алгоритмы эмпирического восстановления стохастических и нечетких моделей // IX Межд. конф. «Интеллектуальные системы и компьютерные науки», Москва — 2007.

О согласованных оценках сложности задач и алгоритмов классификации

Романов Л. Ю.

lromanov@gmail.com

Москва, Вычислительный центр РАН

В докладе рассматриваются два независимо определяемых понятия сложности: геометрическая сложность конфигурации объектов обучающей выборки и функциональная сложность выборки как сложность разделяющей поверхности. Вычислительные эксперименты подтвержда-

ют гипотезу о корреляции указанных сложностей. Строится экспериментальная зависимость, позволяющая по одной из величин произвести оценку другой.

Постановка задачи

Рассматривается стандартная задача классификации с двумя непересекающимися классами в евклидовом пространстве размерности n.

Будем изучать возможную зависимость сложностей на примере алгоритма SVM [1, 2], строящего в заданном пространстве линейную разделяющую поверхность. Для построения нелинейного решающего правила используется следующая схема: исходное пространство расширяется дополнительными осями (признаками) таким образом, чтобы SVM разделял обучающую выборку без ошибок. Значения координат по каждой из дополнительных осей являются нелинейными функциями от исходных координат: $x_{n+k} = g_k(x_1, \ldots, x_n),\ k = 1, 2, \ldots$ Построенное таким образом пространство называют *спрямляющим*. В исходном пространстве разделяющая поверхность оказывается нелинейной и разделяет обучающую выборку без ошибок.

Основной проблемой при построении спрямляющего пространства является нахождение оптимального набора дополнительных осей. Поэтому важной задачей представляется оценивание сложности спрямляющего пространства до его построения, на основе лишь геометрической сложности обучающей выборки.

Геометрическая сложность обучающей выборки

Геометрическую сложность можно определять различными способами. Наибольший интерес представляет степень взаимного проникновения классов друг в друга. Для оценки этой величины можно разными способами измерять расстояние между классами.

Определение 1. *Геометрической сложностью обучающей выборки в n-мерном пространстве будем называть отношение стороны описанного около выборки n-мерного куба к минимальному расстоянию между объектами двух классов.*

Функциональная сложность обучающей выборки

Основная идея нахождения сложности разделяющей поверхности состоит в том, что в качестве сложности поверхности принимается определенная некоторым образом сложность спрямляющего пространства, достаточного для разделения обучающей выборки без ошибок.

Будем говорить, что дополнительная ось задается функционалом $g(x_1, \ldots, x_n)$, если вводится новая координата, значения которой для рассматриваемого множества точек задаются указанным функционалом.

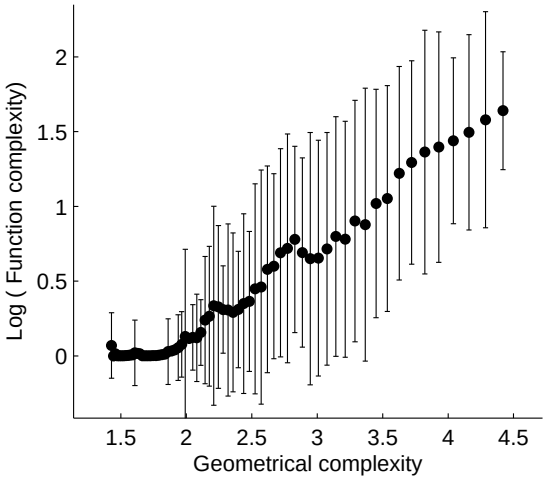


Рис. 1. Зависимость функциональной сложности от геометрической (усреднение по 50 ближайшим значениям геометрической сложности).

Функционалы будем задавать в виде суперпозиции функций некоторого базиса $\mathscr{F} = \left\{ f_i \colon \mathbb{R}^{k_i} \to \mathbb{R} \;\middle|\; i = 1, \ldots, m \right\}$.

Будем считать, что каждая функция f_i из базиса $\mathscr{F}$ обладает некоторой сложностью c_i, которая задается исследователем априори. Сложность суперпозиции функций определим как суммарную сложность входящих в нее функций.

Рассмотрим всевозможные наборы дополнительных осей, дающих линейное разделение выборки. Сложность набора осей определим как сумму сложностей функционалов, задающих оси из набора. Функциональную сложность выборки определим как минимальную сложность среди таких наборов осей.

Алгоритм нахождения функциональной сложности

Пусть мы имеем некоторую обучающую выборку. Следующий алгоритм описывает процедуру нахождения ее функциональной сложности.

Перебираются всевозможные наборы функций, образованные суперпозициями функций из базиса, в порядке увеличения сложности набора. Для каждого набора, задающего дополнительные оси спрямляющего пространства, строится линейное разделение в расширенном ими пространстве с помощью алгоритма SVM. Если расширенное пространство является спрямляющим, то функциональная сложность найдена, иначе перебор продолжается.

Эмпирическая зависимость функциональной сложности от геометрической

Для эмпирического исследования зависимости функциональной сложности от геометрической строятся обучающие выборки с заданной геометрической сложностью, для каждой из них с помощью приведенного алгоритма вычисляется функциональная сложность.

В качестве реализации метода SVM в данной работе используется алгоритм SMO [3].

В настоящей работе описанный выше метод анализа проводится при некоторых упрощениях. Во-первых, рассматриваются двумерные выборки. Во-вторых, в качестве функций, задающих координаты дополнительных осей, рассматриваются мономы $f_i(x_1, x_2) = x_1^{\alpha_1^i} x_2^{\alpha_2^i}$, сложность которых определяется как $C(\alpha_1) + C(\alpha_2)$, где

$$C(\alpha) = \begin{cases} \alpha, & \text{если } \alpha = 1, 2, 3, \ldots; \\ \beta + 1, & \text{если } \alpha = -\beta, \quad \beta = 1, 2, 3, \ldots; \\ \beta, & \text{если } \alpha = \frac{1}{\beta}, \quad \beta = 2, 3, \ldots; \\ \beta + 1, & \text{если } \alpha = -\frac{1}{\beta}, \quad \beta = 2, 3, \ldots. \end{cases}$$

График экспериментальной зависимости функциональной сложности от геометрической приведен на Рис. 1. Для отображения функциональной сложности выбрана логарифмическая шкала. Для каждого значения также отложена дисперсия.

Работа выполнена при поддержке РФФИ, проект №06-07-89315-а.

Литература

[1] *Vapnik V.* Estimation of Dependences Based on Empirical Data — Springer–Verlag, 1982.

[2] *Burges C. J. C.* A tutorial on support vector machines for pattern recognition. — Data Mining and Knowledge Discovery. — Vol. 2, No. 2. — 1998.

[3] *Platt J. C.* Sequential minimal optimization: A fast algorithm fo training support vector machines. — Technical Report MSR–TR–98–14, Microsoft Research, 1998.

Построение корректного распознающего алгоритма минимальной степени в алгебре над множеством алгоритмов вычисления оценок

Романов М. Ю.
mromanov@ccas.ru
Москва, МФТИ

В настоящей работе рассматривается метод построения распознающего алгоритма в алгебраическом расширении наименьшей степени над множеством алгоритмов вычисления оценок (АВО).

Используются обозначения, применяемые в работе [1].

Рассматриваемые алгоритмы составляются из распознающего оператора и решающего правила. Распознающий оператор вычисляет оценки близости объектов к классам, а решающее правило на основе этих оценок классифицирует объекты.

Для начальной информации I_0 и контрольной выборки $\widetilde{S}^q$ запишем задачу распознавания $Z = (I_0, \widetilde{S}^q)$. Будем считать, что задано множество распознающих операторов (РО) $B^* = \{B_1, \ldots, B_n\}$. Рассматривается алгоритм A, в котором распознающим оператором является полином над операторами $\{B_k\}$. Каждый оператор B_k строит матрицу оценок $B_k(Z) = \|\Gamma_k^{uv}\|_{q\times l}$; оператор C задан стандартным пороговым решающим правилом. В докладе рассматривается задача построения корректного алгоритма A минимальной степени, т. е. алгоритма, распознающий оператор которого является полиномом минимальной степени.

Рассматривается семейство распознающих операторов, для которых существует алгоритм вида

$$A = \left(\sum_{k=1}^{n} B_k^{x_k}\right) \circ C\,(c_1, c_2)\,,$$

корректный для задачи Z, см. [2]. Описан алгоритм нахождения набора степеней $x_k,\ k = 1, \ldots, n$, дающих корректный алгоритм минимальной степени, и рассмотрен вопрос уменьшения числа слагаемых полинома.

Так как оператор C фиксирован, в дальнейшем будем отождествлять построение корректного алгоритма A и построение соответствующего многочлена над множеством операторов B_k.

Для выборки $\{S^u\colon u = 1, \ldots, q\}$ и классов $\{K_v\colon v = 1, \ldots, l\}$ разобьём элементы матрицы ответов на 2 множества:

$$M_0 = \{(u, v) : S^u \notin K_v\};\ M_1 = \{(u, v) : S^u \in K_v\}.$$

В работе [2] показано, что для нахождения полинома минимальной степени нужно решить следующую оптимизационную задачу для нахож-

дения вектора $\tilde{y} = \{y_1, \ldots, y_n\}$:

$$\begin{cases} \forall\,(u,v) \in M_0\colon\ \varphi^{uv}\,(\tilde{y}) \leqslant c_1; \\ \forall\,(u,v) \in M_1\colon\ \varphi^{uv}\,(\tilde{y}) \geqslant c_2; \\ \max\limits_{k=1,\ldots,n} y_k \to \min; \end{cases}$$

где введены обозначения: $y_k = e^{x_k},\ \gamma_k^{uv} = \ln \Gamma_k^{uv},\ \varphi^{uv}\,(\tilde{y}) = \sum_{k=1}^{n} y_k^{\gamma_k^{uv}}$.

Результаты

Используя результаты работы [3], имеем, что для любого решения этой оптимизационной задачи выполняется $0 \leqslant y_k \leqslant ql,\ k = 1, \ldots, n$. Тогда задачу можно записать в виде

$$R = \Big\{\tilde{y} \in U,\ \max_{k=1,\ldots,n} y_k \to \min\Big\},$$

где

$$U = \left\{ \tilde{y} \;\middle|\; \begin{array}{l} \varphi^{uv}(\tilde{y}) \leqslant c_1,\ \text{для всех } (u,v) \in M_0; \\ \varphi^{uv}(\tilde{y}) \geqslant c_2,\ \text{для всех } (u,v) \in M_1; \\ 0 \leqslant y_k \leqslant ql,\ \text{для всех } k = \overline{1,n} \end{array} \right\}.$$

В работе [2] показано, что решение этой задачи содержится среди решений вспомогательных задач $R_I = \{\tilde{y} \in U_I,\ y_{i_1} \to \min\}$, с множеством ограничений $U_I = \{\tilde{y} \in U,\ y_{i_1} = y_{i_2} = \ldots = y_{i_m}\}$, для некоторого непустого множества индексов $I = \{i_1, \ldots, i_m\} \subseteq \{1, \ldots, n\}$. Используя метод линеаризации, можно свести задачу R_I к задаче квадратичного программирования [4].

При реализации метода решения задачи квадратичного программирования вместо симплекс-метода может быть использован обобщенный метод Ньютона так, как описано в статье [5]. Кроме того, метод Ньютона может быть непосредственно применен к решению задачи квадратичного программирования [6].

Описана процедура перебора задач R_I, позволяющая сократить перебор за счёт исключения вспомогательных задач, заведомо не дающих решение исходной. Указывается один метод последовательного уменьшения областей ограничений.

Следующим шагом построения корректного многочлена минимальной степени является построение многочлена минимальной степени от меньшего числа слагаемых. Рассмотрим начальный набор РО $\mathfrak{B} = \{B_k\}$, который является базисным для Z. Для него задан набор функций $\varphi^{uv}(\tilde{y})$, задающих ограничения оптимизационной задачи. Необходимо найти такой базисный поднабор $\mathfrak{B}^* \subseteq \mathfrak{B}$, для которого решение задачи $R\,(\mathfrak{B}^*)$ даёт полином наименьшей степени $F\,(\mathfrak{B}^*)$, корректный для задачи Z.

Будем говорить, что набор $\mathfrak{B}' \subseteq \mathfrak{B}$ с соответствующим набором функций $\varphi'^{uv}(\tilde{y})$ имеет тип 1, если для некоторой пары $(u, v) \in M_1$ выполняется $\varphi'^{uv}(\tilde{y}) = c_2$; соответственно имеет тип 0, если для некоторой пары $(u, v) \in M_0$ выполняется $\varphi'^{uv}(\tilde{y}) = c_1$. Задача перечисления базисных поднаборов набора $\mathfrak{B}$ эквивалентна задаче перечисления покрытий множества M_1 множествами $M(B_k)$, $k = 1, \ldots, n$. Можно показать, что решение задачи могут давать либо $\mathfrak{B}$, либо набор, дающий тупиковое покрытие, причем $\mathfrak{B}$ дает решение тогда и только тогда, когда имеет тип 1.

Здесь можно использовать метод уменьшения областей ограничений, аналогичный предыдущему.

Описанные алгоритмы могут быть эффективно распараллелены.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00718, № 06-07-89299, № 07-07-00181, и гранта Президента РФ по поддержке ведущих научных школ НШ-5833.2006.1.

Литература

[1] *Журавлев Ю. И., Исаев И. В.* Построение алгоритмов распознавания, корректных для заданной контрольной выборки // Ж. вычисл. матем. и матем. физ. — 1979. — Т. 19, № 3. — С. 726–738.

[2] *Романов М. Ю.* Об одном методе построения распознающего алгоритма в алгебре над множеством вычисления оценок // Ж. вычисл. матем. и матем. физ. — 2007. — Т. 47, № 8. — С. 1426–1430.

[3] *Рудаков К. В.* Алгебраическая теория универсальных и локальных ограничений для алгоритмов распознавания. — Дисс. докт. физ.-мат. наук, М.: ВЦ РАН, 1992.

[4] *Васильев Ф. П.* Численные методы решения экстремальных задач. — М.: Наука, 1988.

[5] *Голиков А. И., Евтушенко Ю. Г., Моллаверди Н.* Применение метода Ньютона к решению задач линейного программирования большой размерности // Ж. вычисл. матем. и матем. физ. — 2004. — Т. 44, № 9. — С. 1564–1573.

[6] *Coleman T. F., Li Y.* A Reflective Newton Method for Minimizing a Quadratic Function Subject to Bounds on some of the Variables // SIAM Journal on Optimization. — 1996. — Vol. 6, № 4. — Pp. 1040–1058.

Универсальные критерии кластеризации и вопросы устойчивости

Рязанов В. В., Арсеев А. С., Коточигов К. Л.
rvv@ccas.ru
Москва, Вычислительный центр РАН

Задачи кластеризации (или автоматической классификации, таксономии, самообучения, обучения без учителя, группировки) образуют важный раздел интеллектуального анализа данных. Существует несколько постановок кластерного анализа, но основная состоит в поиске разбиения выборок объектов (при заданных признаковых пространствах или матрицах близостей объектов) на классы эквивалентности (кластеры), причем эквивалентность объектов кластеров определяется каждым алгоритмом по-своему. Принципы, согласно которым объекты объединяются в один кластер, являются обычно «внутренним делом» конкретного алгоритма кластеризации. Пользователь, зная данные принципы, может в определенных пределах интерпретировать результаты каждого конкретного метода [1–4].

В отличие от задач распознавания, где существуют единые стандартные критерии оценки алгоритмов (оценка вероятности ошибки, эмпирический риск, и другие), в настоящее время не существует универсальных общепризнанных критериев качества решения задачи кластеризации. Соответственно, при отсутствии внешней суперцели, решения различных алгоритмов кластеризации сложно оценивать и сравнивать. Действительно, есть методы кластеризации, где ищется экстремум некоторого функционала качества разбиения (например, дисперсионный и родственные ему критерии, определитель матрицы внутригруппового разброса, и другие). Метод k-внутригрупповых средних находит группировки, где каждый объект находится ближе к среднему своей группировки, чем к среднему любой другой. Данные группировки и объявляются кластерами. Кластеры в методах иерархической группировки вычисляется согласно последовательному локальному объединению более мелких группировок в более крупные. В алгоритме ФОРЕЛЬ кластером объявляется группировка объектов, принадлежащая некоторому шару фиксированного радиуса, обладающего свойством: центр шара совпадает со средним принадлежащих ему объектов.

В настоящем докладе в основу общего универсального подхода предлагается положить идею устойчивости решений относительно малых изменений выборки. Здесь возможны различные критерии оценки качества кластеризаций. Назовем m-выборкой произвольную выборку из m объектов, а $(m-1)$-выборкой — выборку, полученную из m-выборки удалением

некоторого объекта. Определяется близость кластеризаций m-выборки и $(m-1)$-выборки на заданное число кластеров. По близостям кластеризаций m-выборки и всевозможных $(m-1)$-выборок оценивается качество кластеризации исходной m-выборки. Возможны и другие варианты определения качества кластеризаций.

Данные определения не связаны с сутью конкретного метода кластеризации, а отражают степень изменения кластеризаций относительно вариации выборок. Для некоторых алгоритмов кластерного анализа получены эффективные методы вычисления качества кластеризаций. Приводятся результаты анализа оценки качества кластеризаций на модельных и реальных задачах.

Литература

[1] *Айвазян С. А. и др.* Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989.

[2] *Дуда Р., Харт П.* Распознавание образов и анализ сцен. М.: Мир, 1976. — 511 с.

[3] *Загоруйко Н. Г.* Методы распознавания и их применение. М.: Сов. радио, 1972. — 206 с.

[4] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Изд-во Института математики, 1999.

К определению сильного перемешивания разбиений пространства

Трофимов О. Е.

trofimov@iae.nsk.su

Новосибирск, Институт автоматики и электрометрии СО РАН

Для большого количества проблем дешифровки и определения скрытых смыслов нужно уметь решать следующую задачу: «Является ли случайной предъявляемая последовательность чисел?».

А. Н. Колмогоров предложил в качестве меры сложности рекурсивно перечислимой последовательности чисел минимальную длину программы, порождающей эту последовательность [4]. Под длинной порождающей программы можно понимать логарифм номера порождающей функции в некоторой нумерации частично рекурсивных функций. А. Н. Колмогоров предложил также связать меру случайности последовательности с мерой её сложности.

Наряду с понятием случайной последовательности целесообразно рассматривать случайное разбиение натурального ряда в целом. Такое разбиение обладает свойствами сильного перемешивания. Подобные разбиения могут рассматриваться в пространствах произвольной размерности.

В частности, в непрерывном случае примером разбиения трехмерного шара может служить клубок нитей (координаты каждой нити — отдельное множество). Примером разбиения с сильным перемешиванием может быть клубок сильно перепутанных нитей. В дискретном варианте использование нумерации кортежей позволяет свести задачи в пространствах произвольной размерности к одномерному случаю.

По нашему мнению, в дискретном случае для формализации понятия разбиения сильным перемешиванием (хаотического разбиения) целесообразно использовать понятие предполной позитивной нумерации, введенное А. И. Мальцевым [5, 6].

Пусть задано некоторое семейство объектов. Будем говорить, что задана нумерация этого семейства, если каждому объекту поставлено в соответствие некоторое натуральное число. Один объект может иметь несколько номеров, но разным объектам должны соответствовать разные числа. Если речь идет, например, об эффективно вычислимых функциях, то номерами этих функций могут быть коды соответствующих программ.

Для частично-рекурсивных функций существует универсальная функция Клини $K(n, m)$ со следующим свойством. Для любой частично-рекурсивной функции $f(m)$ существует n такое, что $K(n, m) = f(m)$, число n называется клиниевским номером функции $f(m)$. В клиниевской нумерации любое число является номером некоторой функции. В дальнейшем мы будем рассматривать нумерации, в которых любое натуральное число является номером некоторого объекта. Клиниевская нумерация обладает двумя важными свойствами.

1. Для любой общерекурсивной функции $g(n)$ существует n_0, такое, что $K(g(n_0), m) = K(n_0, m)$. Это означает, что числа $g(n_0)$ и n_0 являются номерами одной и той же функции, причем число n_0 можно эффективно найти по номеру функции $g(n)$. Иными словами для любого общерекурсивного отображения $g(n)$ существует неподвижная точка, которую можно эффективно найти.
2. Для любой частично-рекурсивной функции $\varphi(n)$ существует общерекурсивная функция $\alpha(n)$ такая, что $K(\varphi(n), m) = K(\alpha(n), m)$, если $\varphi(n)$ определено, и $\alpha(n)$ есть номер нигде не определенной функции, если $\varphi(n)$ не определено.

Нумерации, удовлетворяющие свойству 1, А. И. Мальцев назвал предполными, а нумерации, удовлетворяющие свойству 2, — полными [5, 6]. В случае произвольных семейств объектов роль нигде не определенной функции играет некоторый выделенный элемент.

Если каждое натуральное число является номером некоторого элемента, то нумерация задает разбиение натурального ряда на классы эквивалентности, являющиеся номерами одного элемента. Нумерация

называется позитивной, если все ее классы эквивалентности являются рекурсивно-перечислимыми множествами. Полные нумерации не могут быть позитивными, так как множество номеров выделенного элемента не является рекурсивно-перечислимым [6].

Существуют предполные позитивные нумерации [1–3], они и дают пример рекурсивно-перечислимого хаоса [7, 8].

Здесь мы приведем пример позитивной предполной нумерации, построенный автором настоящей работы. Этот пример с согласия автора и со ссылкой на него приведен в [2].

Рассмотрим функцию $\varphi(n) = K(n, m)$, как уже говорилось выше, $K(n, m)$ — универсальная функция Клини.

Рассмотрим семейство множеств $\{A_r\}$, где r — произвольное натуральное число, а множество A_r удовлетворяет следующим условиям:

1) $r \in A_r$;
2) если $r_1, \ldots, r_n \in A_r$, то из условия $\bigcup_{i=1}^{n} r_i \bigcap \{n, \varphi(n)\} \neq \varnothing$ следует $n \in A_r \vee \varphi(n) \in A_r$;
3) других элементов в A_r нет.

Рекурсивная перечислимость множеств A_r и теорема о неподвижной точке следуют непосредственно из построения. Из построения также следует, что различные номера нигде не определенной функции принадлежат разным множествам A_r, это означает, что ни один из классов эквивалентности не совпадает со всем натуральным рядом, то есть, мы действительно получаем разбиение множества натуральных чисел.

Разбиения, построенные выше, могут быть использованы для кодирования информации. Если не известна порождающая функция разбиения, его очень трудно отличить от случайного. Предложенная конструкция может использоваться для построения статистических гипотез на независимость. Следует отметить, что для эффективной работы методов, основанных на предполных разбиениях, необходимы последовательности большого размера.

Литература

[1] *Ершов Ю.* Теория нумераций. — Москва: Наука, 1977.

[2] *Ершов Ю.* Теория нумераций 1. — Новосибирск: Наука, 1966.

[3] *Ershov Y.* Theory of enumerations, Part II // Z. Math. Log. und Grundl. Math. — 1975. — Т. 21. — Pp. 473–584.

[4] *Колмогоров А. Н.* Три подхода к определению информации // Проблемы передачи информации — 1965. — № 1. — С. 1–7.

[5] *Мальцев А. И.* Полно нумерованные множества // Алгебра и логика — 1963. — № 2. — С. 4–29.

[6] *Мальцев А. И* К теории семейств вычислимых объектов // Алгебра и логика — 1964. — № 4. — С. 5–31.

[7] *Трофимов О. Е.* О рекурсивно нумерованном хаосе // Тез. межд. конф. «Колмогоров и современная математика». — Москва: МГУ, 2003. — С. 699.

[8] *Trofimov O. E.* Chaos and Recursive Denumerrability // Int. conf. on modelling & simulation (ICMS'04-Spain), Valladolid, 2004. — Pp. 1933–194.

Эмпирическое построение неопределенного нечеткого (НН) элемента

Фаломкина О. В., Пытьев Ю. П.

pytyev@phys.msu.su

Москва, МГУ им. М. В. Ломоносова, физический факультет

В докладе рассмотрены неопределенные нечеткие (НН) модели, в которых нечеткость, неточность формулировок, относящаяся к содержанию информации, охарактеризована в терминах значений мер возможности и (или) необходимости, а их достоверность, истинность которых не может быть абсолютной в силу принципиальной неполноты знаний, охарактеризована в терминах значений мер правдоподобия и (или) доверия [6, 5].

Обозначим $(Y, \mathcal{P}(Y),\ \mathrm{P})$ — пространство с возможностью, в котором Y — множество элементарных событий, $\mathcal{P}(Y)$ — класс всех подмножеств Y, называемых событиями, $\mathrm{P}\colon \mathcal{P}(Y) \to [0,\,1]$ — мера возможности (возможность).

Возможность $\mathrm{P}(\cdot)$ определяется ее значениями $f^{\eta}(y) \triangleq \mathrm{P}(\{y\}) = \mathrm{P}(\eta = y),\ y \in Y$, на одноточечных подмножествах $\{y\} \subset Y$, а именно, $\mathrm{P}(A) \triangleq \mathrm{P}(\eta \in A) = \sup\limits_{y \in A} f^{\eta}(y),\ A \in \mathcal{P}(Y)$.

Функция $f^{\eta}\colon Y \to [0, 1]$ называется распределением возможностей значений канонического для $(Y, \mathcal{P}(Y),\ \mathrm{P})$ нечеткого элемента $\eta\colon Y \to \to (Y, \mathcal{P}(Y), \mathrm{P}(\cdot))$.

Обозначим аналогично $(\mathcal{U}, \mathcal{P}(\mathcal{U}), \mathrm{P}l(\cdot))$ — *пространство с правдоподобием*, в котором $\mathcal{U}$ — множество элементарных высказываний, $\mathcal{P}(\mathcal{U})$ — класс всех подмножеств $\mathcal{U}$ (высказываний), $\mathrm{P}l(\cdot)\colon \mathcal{P}(\mathcal{U}) \to [0,\,1]$ — мера правдоподобия. Аналогично возможности $\mathrm{P}(\cdot)$, правдоподобие $\mathrm{P}l(\cdot)$ определяется распределением правдоподобий $g^{\tilde{u}}(\cdot)\colon \mathcal{U} \to [0,\,1]$ значений канонического неопределенного элемента $\tilde{u}\colon (\mathcal{U}, \mathcal{P}(\mathcal{U}), \mathrm{P}l(\cdot)) \to \mathcal{U}$.

Определение 1 (Пытьев, [5]). *Неопределенным нечетким (НН) элементом, принимающим значения в X, называется образ $\tilde{\xi} \triangleq q(\eta, \tilde{u})$ (упорядоченной) пары $(\eta, \tilde{u})$ — нечеткого η и неопределенного $\tilde{u}$ элементов при отображении $q\colon Y \times \mathcal{U} \to X$.*

Функция $\tau_x^{\tilde{\xi}}(p) \triangleq \mathrm{P}l(P(\tilde{\xi} = x) = p) = \sup\left\{g^{\tilde{u}}(u) \;\middle|\; u \in \mathcal{U},\ f^{\xi_u}(x) = p\right\}$, $x \in X,\ p \in [0, 1]$, называется распределением правдоподобия возможностей значений НН элемента $\tilde{\xi}$, или, короче, распределением $\tilde{\xi}$. Её значение $\tau_x^{\tilde{\xi}}(p)$ определяет правдоподобие истинности «элементарного» высказывания, согласно которому p — возможность равенства $\tilde{\xi} = x \in X$.

В работах [6, 7] вопрос об эмпирическом построении НН модели не затрагивался, а вместе с тем этот вопрос при решении прикладных задач является одним из центральных. В докладах [2, 3, 4] предложены методы эмпирического построения теоретико-ваозможностных моделей, имеющих стохастический прототип, и соотвественно на основе экспертных оценок.

В докладе рассматриваются математические методы и алгоритмы эмпирического построения модели НН элемента, в которой нечеткость, неточность обусловлена случайностью, а возможность имеет стохастический прототип, неопределенность, неясность обусловлена принципиальной неполнотой знаний, а правдоподобие определяется на основе высказываний экспертов.

Работа выполнена при поддержке РФФИ, проект № 05-01-00532-а.

Литература

[1] *Пытьев Ю. П.* Стохастические и нечеткие модели. Эмпирическое построение и интерпретация // Сборник трудов 1-й международной научно-практической конференции «Современные информационные технологии и ИТ-образование». — 2005. — С. 482–492.

[2] *Пытьев Ю. П.* Экспертное оценивание нечеткого элемента // ММРО-13 (в настоящем сборнике). — 2007. — С. 52–54.

[3] *Пытьев Ю. П.* Математические методы и адаптивные алгоритмы эмпирического построения теоретико-возможностной модели стохастического объекта // ММРО-13 (в настоящем сборнике). — 2007. — С. 54–56.

[4] *Пытьев Ю. П.* Математические методы и алгоритмы эмпирического восстановления стохастических и нечетких моделей // IX Межд. конф. «Интеллектуальные системы и компьютерные науки». — 2007.

[5] *Пытьев Ю. П.* Неопределенные нечеткие модели и их применения // Интеллектуальные системы. — 2004. — № 8, Вып. 1–4. — С. 147–310.

[6] *Пытьев Ю. П., Фаломкина О. В.* О критериях оптимальности в неопределенных нечетких моделях // ММРО-12. — 2005. — С. 202–206.

[7] *Фаломкина О. В.* Нечеткие и неопределенные нечеткие модели и их применения. — Дисс. канд. физ.-мат. наук. — 2006.

Метрический подход к проблеме оценивания ошибок алгоритмов классификации

Черепнин А. А.
cherepnin@forecsys.ru
Москва, ЗАО «Форексис»

В работах [1, 2, 3] были введены понятия радиусов разрешимости и регулярности задач классификации. При их определении предполагалось, что на пространстве задач с фиксированной системой универсальных ограничений [4, 5, 6] введена метрика. Под радиусом регулярности задачи при этом понимается расстояние от нее до ближайшей нерегулярной задачи, а под радиусом разрешимости, соответственно, — расстояние до ближайшей неразрешимой.

Величины радиусов разрешимости и регулярности позволяют судить (естественно, если метрика на пространстве задач введена достаточно адекватно), например, о том, что задача оказывается разрешимой в силу избыточно точного измерения некоторых признаков.

Основная идея предлагаемого подхода состоит в том, что величины радиусов прежде всего разрешимости можно использовать для получения оценок ошибок алгоритмов классификации и для анализа отдельных признаков в описании объектов.

Для получения оценок ошибок на отдельных объектах или группах объектов предлагается проводить сравнение радиусов разрешимости «редуцированных» задач, получаемых из исходной элиминацией отдельных объектов или равномощных групп объектов. При этом в качестве оценки ошибки может быть использован некоторый монотонно убывающий функционал от радиуса разрешимости задачи, полученной в результате элиминации оцениваемого объекта (или группы). Действительно, если при элиминации определенного объекта радиус разрешимости редуцированной задачи оказывается максимальным среди всех задач, полученных элиминацией одного объекта, то это означает, что именно этот объект делает исходную задачу «максимально близкой» к неразрешимой задаче, то есть к задаче, в которой универсальные и локальные ограничения взаимно противоречивы.

При анализе групп объектов возникают проблемы переборного характера, поскольку приходится рассматривать количество задач, равное количеству сочетаний из исходного числа объектов по количеству элементов в анализируемых группах. В докладе описываются свойства метрик на пространствах задач, обеспечивающие возможность резкого снижения сложности перебора при решении этой проблемы. Отметим, что снижe-

ние сложности перебора достигается за счет решения специальных задач дискретной оптимизации.

Применение метрического подхода к проблеме оценивания ошибок алгоритмов классификации возможно, в частности, путем введения оценок ошибок на объектах как значений монотонно убывающих функционалов от радиусов разрешимости редуцированных задач, полученных элиминацией групп, в которые входит анализируемый объект. После этого получение оценки интегральной ошибки алгоритма сводится к суммированию оценок ошибок объектов, на которых алгоритм принял неправильное решение.

Отдельный интерес представляет использование предлагаемого подхода для анализа результатов использования различных алгоритмов при решении одной и той же задачи. Действительно, даже если два алгоритма допустили на обучении одинаковое количество ошибок, то может оказаться, что один из них допускал ошибки на объектах, имеющих сравнительно низкие оценки, полученные по вышеописанной методике, другой же — наоборот. В этом случае, использование второго алгоритма представляется заведомо более предпочтительным.

Отметим, что основной особенностью предложенного подхода представляется независимость получаемых с его помощью оценок от фиксации какого-либо конкретного семейства алгоритмов классификации. Использование таких семейств обычно сводится к тому, что в качестве оценки объекта выступает доля алгоритмов, давших для этого объекта при той или иной стратегии проведения экспериментов правильный результат. Очевидно, что, варьируя используемое семейство алгоритмов, для любого объекта при такой методике можно получить произвольный наперед заданный результат. Основным элементом произвола при предлагаемом подходе оказывается способ метризации пространства задач, так что в тех случаях, когда такая метризация может быть проведена на основе достаточно реалистичных предположений, можно надеяться на получение достаточно адекватных результатов.

Также следует отметить, что значительная часть способов оценивания с помощью фиксированных семейств алгоритмов классификации также включает в себя этап метризации пространств объектов, так что в этом случае выбор семейства алгоритмов оказывается дополнительным элементом произвола, избежать которого позволяет предлагаемый подход.

Работа выполнена при поддержке РФФИ, проект № 07-07-00711.

Литература

[1] *Рудаков К. В.* Универсальные и локальные ограничения в проблеме коррекции эвристических алгоритмов классификации // Кибернетика. — 1987. — № 2. — С. 30–35.

[2] *Рудаков К. В.* Полнота и универсальные ограничения в проблеме коррекции эвристических алгоритмов классификации // Кибернетика. — 1987. — № 3. — С. 106–109.

[3] *Рудаков К. В.* Симметрические и функциональные ограничения для алгоритмов классификации // Кибернетика. — 1987. — № 4. — С. 73–77.

[4] *Рудаков К. В., Черепнин А. А., Чехович Ю. В.* О метрических свойствах пространств задач классификации // Доклады РАН. — 2007. — Т. 416, № 4.

[5] *Черепнин А. А.* О радиусах разрешимости и регулярности задач распознавания // всеросс. конф. ММРО-11, Пущино, 2006. — С. 210–211.

[6] *Черепнин А. А.* Об оценках регулярности задач распознавания и классификации // ЖВМиМФ. — 1993. — № 1. — С. 155–159.

Теоретико-множественные ограничения в имитационном моделировании сложных социально-технических систем

Чехович Ю. В.
d_yura@ccas.ru
Москва, ВЦ РАН

Рассматривается класс сложных социально-технических систем, характерной чертой которых является наличие значительного числа субъектов, действующих относительно обособленно и имеющих возможность воздействовать на систему и другие субъекты путем принятия определенных субъективных решений. К системам такого класса можно отнести различные торговые системы, где субъектами, принимающими решения, являются участники торгов; транспортные автомобильные системы (субъекты — водители транспортных средств) [2], системы ведения боевых действий (субъекты — командиры различных уровней и рядовые солдаты), электоральные системы (субъекты — избиратели) и т. п.

Один из основных инструментов изучения таких систем — имитационное моделирование. В докладе описывается подход к созданию имитационных моделей сложных социально-технических систем, характеризующихся наличием элементов субъективного выбора, основанный на методологии алгебраического подхода к синтезу корректных алгоритмов [1] и теории синтеза обучаемых алгоритмов решения задач с теоретико-множественными ограничениями [4].

Опишем предположения, которые принимаются по отношению к моделируемым системам. Число субъектов в таких системах, как правило, достаточно велико — от нескольких десятков до десятков миллионов. Предполагается, что каждый субъект имеет возможность получать и анализировать данные о некоторой своей «окрестности» в рамках си-

стемы (пространственной, временной, информационной), а также может принимать (выбирать) решения из относительно бедного множества возможных решений. Для участников торгов — это, например, возможность купить, продать или воздержаться от каких-либо действий по отношению к определенному инструменту. Для водителя — начать движение, догнать, затормозить, перестроиться в другой ряд, повернуть, обогнать, не предпринимать никаких действий.

Предполагается, что субъект принимает решения, основываясь на анализе локальной ситуации. При этом различные субъекты в одной и той же ситуации (или близких ситуациях) могут принимать различные решения. Будем считать, что субъекты, принимающие одинаковые решения в сходных ситуациях, принадлежат одному классу субъектов. Разумным также представляется предположение, что количество классов субъектов намного меньше общего количества субъектов.

Таким образом, появляется возможность свести задачу имитационного моделирования поведения каждого субъекта к задаче обучения с теоретико-множественными ограничениями. В соответствии с [3] пусть пространство начальных информаций $\mathfrak{I}_i$ есть множество всех возможных описаний локальных ситуаций, а пространство финальных информаций $\mathfrak{I}_f$ — множество всех возможных решений.

Обратим внимание на то, что для каждого элемента пространства начальных информаций $I_i \in \mathfrak{I}_i$ часто имеет смысл рассматривать только соответствующее подмножество $\Pi(I_i)$ пространства финальных информаций, оставляя за пределами данного подмножества решения, которые принципиально не могут быть приняты данным субъектом в конкретной ситуации. С помощью таких теоретико-множественных ограничений можно вводить в модель некоторые «разумные» ограничения на множества принимаемых решений. Например, в экономических моделях можно «запрещать» субъекту принятие заведомо убыточных решений, в моделях транспортных потоков можно таким образом вводить предположение о стремлении водителей к безаварийной езде и т. д.

Прецедентами при такой постановке оказываются пары вида ⟨описание локальной ситуации, тип принятого решения⟩. Далее, следуя парадигме алгебраического подхода, можно проводить обучение модели, синтезируя алгоритм, который классифицирует субъекты по типам их действий в различных ситуациях в соответствии с существующими прецедентами.

Следует отметить тот факт, что введение описанной формализации для описания моделируемой системы позволяет в единых терминах соотнести различные подходы, существующие в имитационном моделировании. Например, для стохастического подхода, после введения ограни-

чений на множество допустимых решений $\Pi(I_i)$, свойственно задавать на получившемся подмножестве вероятностную меру и искомое решение выбирать случайным образом. Для детерминированных моделей подмножество $\Pi(I_i)$ допустимых решений для каждого описания локальной ситуации одноэлементно, что позволяет полностью спрогнозировать развитие системы при заданных начальных условиях. Алгебраический подход к имитационному моделированию допускает существование нескольких возможных решений для каждого описания локальной ситуации и подразумевает синтез алгоритма путем обучения на существующих прецедентах.

Следует также отметить, что для моделирования динамических систем, развивающихся во времени, введение $\mathfrak{I}_i$ как множества всех описаний локальных ситуаций позволяет интерпретировать динамику субъекта как траекторию изменения локальных ситуаций этого субъекта в рамках пространства $\mathfrak{I}_i$.

Работа выполнена при поддержке РФФИ, проект № 07-07-00711 и гранта Президента РФ поддержки молодых ученых — кандидатов наук МК-5266.2007.9.

Литература

[1] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания и классификации // Пробл. кибернетики. — 1978. — Вып. 33. — С. 5–68.

[2] *Иванов Г. Е., Рудаков К. В., Чехович Ю. В.* Об одном подходе к имитационному моделированию транспортных потоков // межд. конф. ИОИ-2006 Симферополь, 2006, — С. 96–97.

[3] *Рудаков К. В.* // Кибернетика. — 1987. № 2. — С. 30–35, 1987. № 3. — С. 106–109, 1987. № 4. — С. 73–77.

[4] *Рудаков К. В., Чехович Ю. В.* Критерии полноты для задач классификации с теоретико-множественными ограничениями // ЖВМиМФ. — 2005. — Vol 45. — № 2. — С. 344–353.

Критерии оптимизации выбора безызбыточных диагностических тестов для принятия решений в интеллектуальных диагностических системах

Янковская А. Е.

yank@tsuab.ru

Томск, Томский архитектурно-строительный университет

Безызбыточные (тупиковые [1]) безусловные диагностические тесты (ББДТ) используются для отнесения исследуемого объекта к одному из распознаваемых образов [1]. От свойств используемых тестов существенно зависит качество тестового распознавания. Однако выбор «хороших»

тестов не всегда приводит к хорошим решениям, поскольку общее количество признаков в выбранном множестве тестов может оказаться слишком большим, также как временные и стоимостные затраты или ущерб, наносимый в результате выявления значений признаков, например, при решении геоэкологических и медицинских задач. Кроме того, на качество тестов влияет способ вычисления весовых коэффициентов признаков, входящих в тесты, используемые для принятия итогового решения.

Проблема нахождения «лучших» ББДТ, позволяющих синтезировать на их основе более простые правила принятия решений [1, 2, 3], не потеряла своей актуальности. В работе [4] впервые была поставлена задача выбора множества ББДТ с заданными свойствами и предложен алгоритм её решения. Количество сформулированных автором критериев выбора оптимального подмножества ББДТ постепенно увеличивалось и достигло 5 в статье [5]. В настоящем докладе формулируется 6 критериев и постановка задачи выбора оптимального множества ББДТ. Обосновывается целесообразность введения нового критерия выбора.

Основные понятия и определения

Воспользуемся определениями и обозначениями, введенными в работах [2, 5].

Тестом называется совокупность признаков, различающих любые пары объектов, принадлежащих разным образам. Тест называется *безызбыточным*, если при удалении любого признака тест перестает быть тестом. Признак называется *обязательным*, если он содержится во всех ББДТ, и *псевдообязательным*, если он не является обязательным и входит в множество используемых при принятии решений ББДТ.

Обозначим через $N = \{N_1, \ldots, N_n\}$ — множество тестов; $Z = \{z_1, \ldots, z_m\}$ — множество признаков; L_i — множество признаков, входящих в тест $N_i, i \in \{1, \ldots, n\}$; $\mathbf{T}$ — булева матрица тестов, строки которой сопоставлены тестам N_i, где $N_i \in N,\ i \in \{1, \ldots, n\}$, столбцы x_j сопоставлены признакам $z_j \in Z,\ j \in \{1, \ldots, m\}$; n_0 — число используемых для принятия решений тестов; $\mathbf{T}_0$ — подматрица матрицы $\mathbf{T}$; N_0 — множество тестов, соответствующих строкам матрицы $\mathbf{T}_0,\ N_0 \subseteq N$.

Обозначим через w_j^r (w_j^g) весовой коэффициент признака $z_j \in Z$, $j \in \{1, \ldots, m\}$, определяемый как разделяющая способность признака [2] (информационный вес по формуле $w_j^g = \frac{k_j}{n_0}$, где k_j — количество единичных значений в столбце x_j матрицы $\mathbf{T}_0$).

Каждому тесту $N_i,\ i \in \{1, \ldots, n\}$ соответствуют:

1) вес теста $W_i^r = \sum\limits_{j \in L_i} w_j^r$;
2) вес теста $W_i^g = \sum\limits_{j \in L_i} w_j^g$;

3) стоимость теста $W'_i = \sum\limits_{j \in L_i} w'_j$, где w'_j — коэффициент стоимости признака $z_j \in Z,\ j \in \{1, \ldots, m\}$;
4) риск $W''_i = \sum\limits_{j \in L_i} w''_j$, где w''_j — ущерб (риск), наносимый в результате выявления (измерения) значения j-го признака.

Постановка задачи. Критерии оптимальности

Дано множество тестов N, представленное матрицей $\mathbf{T}$; множество признаков Z, каждый из которых содержится хотя бы в одном тесте из N; а также: весовые коэффициенты w^r_j и w^g_j, коэффициенты стоимости признаков w'_j и величина ущерба (риска) $w''_j,\ j \in \{1, \ldots, m\}$ от выявления значения j-го признака; веса W^r_j и W^g_j, стоимости W'_j, ущерб (риск) $W''_j,\ i \in \{1, \ldots, n\}$ тестов.

Необходимо выделить из матрицы $\mathbf{T}$ такую подматрицу $\mathbf{T}_0$, содержащую n_0 строк, чтобы соответствующее ей выделенное множество тестов N_0 обеспечивало выполнение следующих критериев оптимальности:

1) во множестве тестов N_0 должно содержаться максимальное число псевдообязательных признаков;
2) множество тестов N_0 должно содержать минимальное общее число признаков;
3) множество тестов N_0 должно иметь максимальный суммарный вес $W^r_0 = \sum\limits_{i \in N^0} W^r_i$;
4) множество тестов N^0 должно содержать максимальный суммарный вес $W^g_0 = \sum\limits_{i \in N^0} W^g_i$;
5) множество тестов N^0 должно иметь наименьшую суммарную стоимость;
6) множество тестов N^0 должно обеспечивать наименьший ущерб (риск).

Заметим, что критерий 4) аналогичен критерию по выбору признаков с максимальным информационным весом [1], вычисляемым только не на всём множестве ББДТ, а на множестве N^0, используемом для принятия итогового решения.

Утверждение 1. Информационный вес W^g_0 множества тестов N^0, умноженный на n_0, равен количеству K_{n_0} единичных значений в матрице $\mathbf{T}_0$.

Доказательство. Поскольку $W^g_0 = \sum\limits_{j=1}^{n_0} w^g_i = \sum\limits_{j=1}^{n_0} \frac{k_j}{n_0} = \frac{K_{n_0}}{n_0}$, следовательно, $W^g_0 n_0 = K_{n_0}$, что и требовалось доказать.

Критерий 4) можно переформулировать следующим образом: множество тестов N^0 должно включать максимальное количество признаков.

Дальнейшее расширение списка критериев выбора связано с учетом зависимости входящих в тесты признаков.

Модификация приведенных в [5] трех алгоритмов (логико-комбинаторного, на основе метода анализа иерархий, генетического) выбора оптимального множества тестов с учетом критерия 4) не представляет затруднений. Модификация алгоритмов будет воплощена в интеллектуальном инструментальном средстве ИМСЛОГ [6].

Работа выполнена при поддержке РФФИ, проект № 07-01-00452а и РГНФ, проект № 06-06-12603в.

Литература

[1] *Журавлев Ю. И., Гуревич И. Б.* Распознавание образов и анализ изображений // Искусственный интеллект: В 3-х кн. Кн.2. Модели и методы: Справ./ Под ред. Д.А.Поспелова. — М.: Радио и связь, 1990. — С. 149–191.

[2] *Янковская А. Е.* Логические тесты и средства когнитивной графики в интеллектуальной системе // Новые информационные технологии в исследовании дискретных структур: Докл. 3-ей Всерос. конф. с межд. участ., Томск: Изд-во СО РАН, 2000. — С. 163–168.

[3] *Naidenova R. A., Plaksin M. V., Shagalov V. L.* Inductive inferring all good classification test // Знание-Диалог-Решение. Сб. науч. тр. Межд. конф. Том 1. Ялта, 1995. — С. 79–84.

[4] *Янковская А. Е.* Построение логических тестов с заданными свойствами и логико-комбинаторное распознавание на них // Интеллектуализация обработки информации. Тезисы докл. — Симферополь, 2002. — С. 100–102.

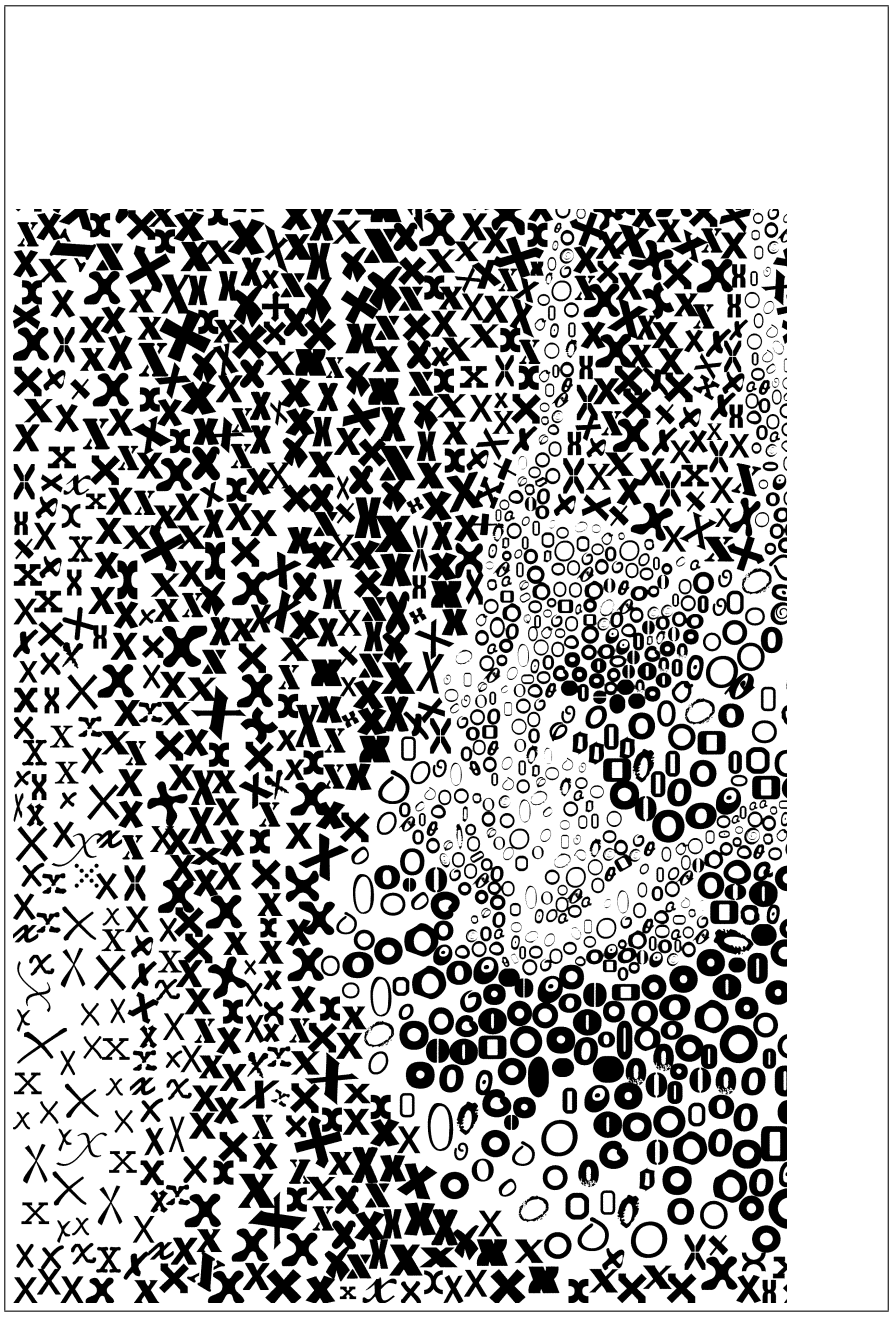
[5] *Колесникова С. И., Можейко В. И., Цой Ю. Р., Янковская А. Е.* Алгоритмы выбора оптимального множества безызбыточных диагностических тестов в интеллектуальных системах поддержки принятия решений // Труды первой межд. конф. САИТ-2005. — Т. 1. — М.: КомКнига, 2005. — С. 256–262.

[6] *Yankovskaya A. E., Gedike A. I., Ametov R. V., Bleikher A. M.* IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition // Pattern Recognition and Image Analysis. — 2003. — Vol. 13, № 4. — P. 650–657.

Методы и модели распознавания и прогнозирования

Код раздела: MM (Methods and Models)

- Дискретные (логические) модели распознавания.
- Статистические модели классификации и регрессии.
- Модели классификации на основе сходства и разделимости.
- Нейросетевые модели.
- Методы отбора и преобразования признаков.
- Методы построения алгоритмических композиций.
- Многомерный анализ.
- Теория и методы прогнозирования временных рядов.
- Обучение без учителя, кластеризация.
- Методы согласования экспертных оценок.



Предельное поведение оценки риска оптимальной групповой процедуры классификации выборки из однопараметрического экспоненциального семейства

Бабушкина Е. В., Чичагов В. В.
helvad@yandex.ru
Пермь, Пермский государственный университет

В данной работе продолжены исследования, начатые в работах [1, 2]. В работе [1] было исследовано предельное поведение оценки байесовского риска при групповой классификации с заданной границей. В данной работе также, как и в [2], эти результаты распространяются на случай классификации с использованием квазиоптимального правила. Классификация в этом случае осуществляется по областям со случайными границами. Другая отличительная часть данной работы состоит в том, что направляющая функция экспоненциального семейства распределений наблюдаемой случайной величины, как функция этой величины, имеет гамма-распределение.

Рассматривается решение следующей задачи групповой классификация n_0 объектов по измерениям их характеристик [3].

Классифицируемая выборка $\pi_{00} = \{X_{0,1}, \ldots, X_{0,n_0}\}$ может принадлежать одной из двух совокупностей $\pi_i,\ i = 1, 2$. Имеются две обучающие выборки $\pi_{i0} = \{X_{i,1}, \ldots, X_{i,n_i}\}$, элементы которых являются независимыми случайными величинами, имеющими то же распределение, что и случайная величина η_i с плотностью $f(y, \theta_i),\ i = 1, 2$.

Решение задачи будем искать при следующих предположениях.

(C_1). Распределение вероятностей случайной величины η_i принадлежит однопараметрическому экспоненциальному семейству с плотностью

$$f(y, \theta_i) = h(y) \exp\left\{\theta_i\ T(y) + V(\theta_i)\right\}, \quad y \in G \subset \mathbb{R}, \quad \theta_i \in \Theta \subset \mathbb{R}.$$

Здесь $h(y),\ T(y)$ — известные борелевские функции, $V(\theta)$ — непрерывно дифференцируемая функция параметра θ.

(C_2). Сумма $S_{n_i} = \sum_{j=1}^{n_i} T(X_{i,j})$ является достаточной статистикой параметра θ_i по выборке $\pi_{i0},\ g_m(t, \theta_i)$ — плотность распределения случайной суммы $\sum_{j=1}^{m} T(X_{i,j})$.

(C_3). Случайная величина $Y_i = T(\eta_i),\ i = 1, 2$, имеет гамма-распределение с плотностью

$$f_{Y_i}(t; \sigma_i, \nu) = \frac{t^{\nu-1}}{\sigma_i^{\nu}\Gamma(\nu)} \cdot e^{-t/\sigma_i}, \quad t > 0, \quad \sigma_i = -\frac{1}{\theta_i} > 0, \quad \nu > 0.$$

(C_4). Для случайных событий $B_i = \left\{\left|\sum_{j=1}^{n_0}\left(T(X_{0,j}) - \nu\sigma_i\right)\right| \geqslant n_i^{\gamma_i}\right\}$, определённых при некотором $\gamma_i \in \left(0, \frac{1}{4}\right)$, справедливы соотношения $\lim\limits_{n_i\to\infty} n_i \,\mathsf{P}\,(B_i) = 0,\ \ i = 1, 2.$

Не нарушая общности рассуждений, далее предполагаем, что $\sigma_1 > \sigma_2$, оба параметра неизвестны. Следуя [2, 3], сформулируем утверждение.

Теорема 1. *Если выполнены условия (C_1), (C_3) то оптимальное решающее правило групповой классификации π_{00} имеет вид:*

$$\pi_{00} \in \pi_1, \quad \textit{если} \ \ q(t) = \ln\frac{\omega_1 g_{n_0}(t,\theta_1)}{\omega_2 g_{n_0}(t,\theta_2)} \geqslant 0, \tag{1}$$

где ω_1, ω_2 — некоторые заданные числа. Неравенство (1) осуществляет разбиение числовой прямой на два интервала

$$J_1 = \{t\colon q(t) < 0\} = \{t\colon t < c\}\,, \qquad c = -\frac{\ln\frac{\omega_1}{\omega_2} - n_0\nu\ln\frac{\sigma_1}{\sigma_2}}{\frac{1}{\sigma_2} - \frac{1}{\sigma_1}}.$$
$$J_2 = \{t\colon t \geqslant c\}\,,$$

Основной качественной характеристикой правила (1) является байесовский риск

$$R = \omega_1 \,\mathsf{P}_{\theta_1}\{\pi_{00} \in \pi_2\} + \omega_2 \,\mathsf{P}_{\theta_2}\{\pi_{00} \in \pi_1\} = \sum_{j=1}^{2}\omega_j \int_{J_j} g_{n_0}(t,\theta_j)\,dt.$$

Определим квазиоптимальное решающее правило групповой классификации π_{00}:

$$\pi_{00} \in \pi_1, \quad \text{если} \ \ q(t \mid S_{n_1}, S_{n_2}) = \ln\frac{\omega_1\widehat{g}_{n_0}(t \mid S_{n_1})}{\omega_2\widehat{g}_{n_0}(t \mid S_{n_2})} \geqslant 0, \tag{2}$$

где $\widehat{g}_{n_0}(t|S_{n_j})$ — несмещенная оценка плотности $g_{n_0}(t,\theta_j)$, являющаяся, согласно [4], плотностью бета-распределения.

Неравенство (2) осуществляет разбиение числовой прямой на две области

$$J_1(S_{n_1}, S_{n_2}) = \{t\colon q(t \mid S_{n_1}, S_{n_2}) < 0\}\,,$$
$$J_2(S_{n_1}, S_{n_2}) = \{t\colon q(t \mid S_{n_1}, S_{n_2}) \geqslant 0\}\,.$$

Определим оценку риска квазиоптимального правила (2), также как и в [2], следующим образом

$$\widetilde{R} = \widetilde{R}(S_{n_1}, S_{n_2}) = \sum_{j=1}^{2}\omega_j \int_{J_j(S_{n_1},S_{n_2})} \widehat{g}_{n_0}(t \mid S_{n_j})dt. \tag{3}$$

Подобно теореме 5 из [2], можно доказать следующее утверждение, определяющее предельное поведение оценки риска (3).

Теорема 2. *Пусть выполнены условия* (C_1)–(C_4), $c > 0$, $n_1 \leqslant n_2$, *неравенство правила (2) обращается в равенство в единственной точке* $\widetilde{c}$,

$$[\widetilde{\sigma}_R]^2 = \sum_{j=1}^{2} \frac{[\omega_j \widetilde{\sigma}_j]^2}{n_j}, \quad [\widetilde{\sigma}_j]^2 = \frac{\left[\widetilde{c} \cdot \widehat{g}_{n_0}(\widetilde{c} \mid S_{n_j})\right]^2}{\nu}, \quad \widetilde{c} > 0. \tag{4}$$

Тогда при $n_1 \to \infty$ *последовательность нормированных оценок квазиоптимального риска* $(\widetilde{R} - R)/\widetilde{\sigma}_R$ *сходится по распределению к стандартной нормальной случайной величине.*

Замечание 1. При конечных объемах обучающих выборок для некоторых сочетаний параметров возможно существование двух решений $\widetilde{c}$, или же отсутствие хотя бы одного решения. В первом случае для расчета $[\widetilde{\sigma}_R]^2$ можно либо воспользоваться теоремой 5 из [2], либо видоизменить формулы (4). Возникновение второй ситуации означает вырожденность квазиоптимального правила групповой классификации.

В дальнейшем планируется изучение правомерности применения асимптотических результатов, полученных в данной работе и в [1, 2] для конечных объемов обучающих выборок, а также сравнение изложенных в этих работах подходов к оценке асимптотической дисперсии байесовского риска.

Работа выполнена при поддержке РФФИ, проект № 05-01-00229.

Литература

[1] *Бабушкина Е. В., Чичагов В. В.* Применение несмещенных оценок к оцениванию риска процедуры групповой классификации // Статистические методы оценивания и проверки гипотез. — Перм. ун-т, 2006. — С. 4–11.

[2] *Чичагов В. В.* Построение статистических выводов, основывающихся на несмещенных оценках, по интервалам случайной длины // Статистические методы оценивания и проверки гипотез. — Перм. ун-т, 2007. — С. 59–71.

[3] *Абусев Р. А., Лумельский Я. П.* Статистическая групповая классификация. — Перм. ун-т, 1987. — 92 с.

[4] *Воинов В. Г., Никулин М. С.* Несмещенные оценки и их применения. — М.: Наука, 1989. — 440 с.

Повышение обобщающей способности бустинга в задачах с перекрывающимися классами

Баринова О. В., Вежневец А. П., Вежневец В. П.
olga.barinova@gmail.com, {avezhnevets, vvp}@graphics.cs.msu.ru
Москва, МГУ им. М. В. Ломоносова

Проблема переобучения является одной из центральных в машинном обучении. Известно, что алгоритмы бустинга (boosting) на одних задачах демонстрируют хорошую способность к обобщению, а на других склонны к переобучению. В работе [1] описан ряд методов для уменьшения переобучения в бустинге. В данной работе предлагаются два новых усовершенствования бустинга и проводится сравнительный анализ их обобщающей способности.

Рассмотрим задачу классификации на два класса. Обозначим обучающую выборку через $T = \{(x_i, y_i)\}_{i=1}^N$, где $x_i \in X$ — вектор признаков, $y_i \in \{-1, 1\}$ — метка класса. Пусть прецеденты (x_i, y_i) взяты из неизвестного распределения $P(x, y)$. В основе алгоритмов бустинга лежит процедура минимизации эмпирического риска:

$$R_N(F) = \frac{1}{N} \sum_{i=1}^{N} C\left(y_i, F(x_i)\right) \to \min,$$

где $F(x)$ — выход бустинга, $C\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ — функция потерь.

Удаление из обучающей выборки путающих прецедентов

Будем называть *путающими* те прецеденты из обучающей выборки, на которых ошибается идеальный байесовский классификатор:

$$\left\{(x_i, y_i) \in T \;\middle|\; P(-y_i \,|\, x_i) > 0.5 > P(y_i \,|\, x_i)\right\}.$$

Путающие прецеденты встречаются в задачах с перекрывающимися распределениями классов. Неправильная классификация этих прецедентов предпочтительнее, чем правильная, однако процедура минимизации эмпирического риска ведет к настройке на *путающих* прецедентах, что может приводить к переобучению.

Первый предлагаемый подход к повышению обобщающей способности бустинга состоит в удалении *путающих* прецедентов из обучающей выборки.

Использование оценки математического ожидания функции потерь с меньшей дисперсией

Второй предлагаемый подход состоит в замене минимизации эмпирического риска при настройке бустинга минимизацией другой оценки

Алгоритм 1. Алгоритм оценивания апостериорной вероятности.

Вход:
Обучающая выборка $T = \{(x_i, y_i)\}_{i=1}^N$;
число итераций алгоритма K;

Выход:
Оценки апостериорной вероятности $\bar{p}(1 \mid x_i),\ i = 1, \ldots, N$;

1: **для всех** $k = 1, \ldots, K$
2: Разделить обучающую выборку случайным образом на три равные части $T_k^1 \cup T_k^2 \cup T_k^3 = T$: $T_k^i \cap T_k^j = \varnothing,\ i \neq j$;
3: Обучить бустинг на первой части выборки T_k^1, получить классификатор F_k;
4: Оценить параметры калибровки A, B для F_k (оптимальные параметры сигмоиды) на второй части выборки T_k^2 при помощи шкалирования Платта [3];
5: Вычислить апостериорные вероятности на третьей части выборки по формуле: $p^k(1 \mid x_i) = \dfrac{1}{1 + \exp\left(AF_k(x_i) + B\right)}$;
6: Вычислить среднее значение апостериорной вероятности для каждого прецедента: $\bar{p}(1 \mid x_i) = \frac{1}{K} \sum_{k=1}^K p^k(1 \mid x_i)$;

математического ожидания функции потерь:

$$R'_N(F) = \frac{1}{N} \sum_{i=1}^N \Big(P(1 \mid x_i) C\big(1, F(x_i)\big) + P(-1 \mid x_i) C\big(-1, F(x_i)\big) \Big).$$

Нетрудно показать, что для дисперсий справедливо соотношение $\mathsf{D}R'_N(F) < \mathsf{D}R_N(F)$, и скорость сходимости $R'_N(F)$ к математическому ожиданию функции потерь значительно выше, чем у $R_N(F)$.

Соответствующее изменение легко встраивается в бустинг. Составляется расширенная обучающая выборка $T' = \big\{(x_i, y_i)\big\}_{i=1}^N \cup \big\{(x_i, -y_i)\big\}_{i=1}^N$ и изменяется инициализация весов бустинга [2]: вместо начальных весов $D_1(i) = \frac{1}{N},\ i = 1, \ldots, N$, берутся веса $D'_1(i) = p(y_i \mid x_i),\ i = 1, \ldots, 2N$.

Оценивание апостериорной вероятности

В обоих методах используются апостериорные вероятности прецедентов из обучающей выборки. Апостериорные вероятности неизвестны, однако их можно оценить. Для этого предлагается Алгоритм 1.

В первом подходе полученная оценка апостериорной вероятности $\bar{p}(1 \mid x_i)$ используется для нахождения *путающих* прецедентов. Прецеденты, для которых $\bar{p}(y_i \mid x_i) < 0,5 < \bar{p}(-y_i \mid x_i)$, удаляются из обучающей

Задача	Оценка ошибки	Исход.	Сокращ.	Расшир.
Breast	3.73	4.6	3.65	4.58
Australian	13.06	15.2	13.8	12.75
German	24.66	25.72	25.35	24.8
Heart	18.34	21.41	18.36	16.67
Pima	24.03	25.58	23.99	23.31
Spam	6.49	6.19	6.02	6.06
Vote	4.51	4.75	4.61	4.37

Таблица 1. Ошибка кроссвалидации (%) бустинга, построенного по полной, сокращенной и расширенной обучающей выборке

выборки. Доля *путающих* прецедентов, обнаруженных в обучающей выборке, дает оценку ошибки обобщения.

Во втором подходе полученная оценка апостериорной вероятности $\bar{p}(1 \mid x_i)$ используется для инициализации весов бустинга: $D'_1(i) = \bar{p}(y_i \mid x_i),\ i = 1, \ldots, 2N$.

Эксперименты

В экспериментах использовался алгоритм бустинга [2] с решающим деревом глубины 1 в качестве базового классификатора и числом итераций 100. Использовалась 50×2 кроссвалидация: данные 50 раз делились случайным образом на 2 равные части — на одной части настраивались алгоритмы, другая часть использовалась для вычисления ошибки.

В таблице 1 приведены результаты экспериментов на данных из репозитория UCI. В первом столбце приведена доля *путающих* прецедентов. Во втором столбце — ошибка кроссвалидации для бустинга, настроенного на исходной обучающей выборке. В третьем — ошибка кроссвалидации для бустинга, настроенного по сокращенной обучающей выборке (первый метод). В четвертом — ошибка кроссвалидации для бустинга, настроенного на расширенной обучающей выборке (второй метод).

Эксперименты показывают, что предлагаемые методы повышают обобщающую способность бустинга и помогают избежать переобучения.

Литература

[1] *Friedman J.* Greedy function approximation: a gradient boosting machine // Annals of Statistics. — 2001. — Vol. 29, № 5. — Pp. 1189–1232

[2] *Schapire R., Singer Y.* Improved boosting algorithms using confidence-rated predictions // Machine Learning. — 1999. — Vol. 37, № 3. — Pp. 297–336

[3] *Niculescu-Mizil A., Caruana R.* Obtaining calibrated probabilities from boosting // 21st Conf. on Uncertainty in Artificial Intelligence, Edinburgh, Scotland, 2005.

Оценивание точности восстановления вещественнозначной функции на основе обучения распознаванию классов её значений

Блыщик В. Ф., Донской В. И.
donskoy@ccssu.crimea.ua
Симферополь, Таврический национальный университет им. В. И. Вернадского

Для восстановления вещественнозначных функций, заданных в конечном числе точек, обычно используются методы регрессионного анализа, и точность аппроксимации оценивается невязкой — степенью «близости» значений модельной функции к значениям восстанавливаемой функции по совокупности заданных точек. В данной работе рассматривается подход к восстановлению функций на основе теории распознавания, обозначаемый нами как CBFA — Classification Based Function Approximation.

Пусть X^n — признаковое пространство, $F\colon X^n \to \mathbb{R}$ — вещественнозначная функция, известная в k точках $x_1, \ldots, x_k$, т. е. заданная множеством пар $T = \{(x_m, y_m) \mid y_m = F(x_m),\ m = 1, \ldots, k\}$.

Реализация подхода CBFA предполагает выполнение следующих этапов.

1. Указать значения μ и M такие, что $\mu < y_m < M$ для всех $m = 1, \ldots, k$. Промежуток $[\mu, M]$ определяет множество допустимых значений функции F на X^n.
2. Разбить отрезок $[\mu, M]$ на l промежутков $\pi_j = [\lambda_{j-1}, \lambda_j),\ j = 1, \ldots, l$, так, чтобы в каждом промежутке разбиения находились точки из множества $\{x_1, \ldots, x_k\}$. Считать $\lambda_0 = \mu;\ \lambda_l = M$.
3. Поставить в соответствие каждому промежутку π_j среднее значение $\hat{F}_j$ по всем y_m таким, что $y_m \in \pi_j$. В результате будет получена кусочно-постоянная аппроксимация функции F.
4. Свяжем с построенным разбиением функцию $\omega\colon X^n \to \{1, \ldots, l\}$, принимающую значение j тогда и только тогда, когда $y \in \pi_j$. Функция ω порождается аппроксимируемой функцией F и заданным разбиением отрезка $[\mu, M]$. По множеству T строится обучающая выборка $T_0 = \{(x_m, \omega(x_m)) \mid m = 1, \ldots, k\}$. Эта выборка используется для обучения распознаванию значений функции ω.
5. При помощи некоторого алгоритма обучения распознаванию по выборке T_0 длины m строится решающее правило $D\colon X^n \to \{1, \ldots, l\}$, определяющее номер класса $j = D(x)$ для любого $x \in X^n$. Этот номер класса j определяет значение $\hat{F}_j$ восстанавливаемой функции в точке x. Будем полагать, что для используемого алгоритма распознава-

ния существует оценка вероятности ошибки вычисления класса (значения $\omega(x)$) такая, что $\mathsf{P}\{D(x) \neq \omega(x)\} < \delta$.

Величина $1-\delta$ мажорирует вероятностную меру события, состоящего в правильном нахождении номера класса $\omega(x)$, что эквивалентно истинности соотношения $\omega(x) = D(x)$ для произвольного $x \in X^n$. Это событие равносильно верному определению правилом D промежутка $[\lambda_{j-1}, \lambda_j)$, в котором находится значение $F(x)$ восстанавливаемой функции. Поэтому $\mathsf{P}\{F(x) \in [\lambda_{j-1}, \lambda_j)\} > 1 - \delta$, где $\mathsf{P}\{\cdot\}$ — вероятностная мера события. Полученный результат может быть оформлен в виде следующего утверждения.

Теорема 1. *Пусть вероятность ошибки правила D не превышает δ. Тогда для любой точки $x \in X^n$, отнесенной правилом D к классу j, значение восстанавливаемой функции F с вероятностью большей, чем $1 - \delta$, будет принадлежать промежутку $[\lambda_{j-1}, \lambda_j)$, $j = 1, \ldots, l$.*

Следствие 1. *Определяемое правилом D в точке $x \in X^n$ значение $\hat{F}_j$ аппроксимируемой функции с вероятностью большей, чем $1 - \delta$, будет отличаться от истинного значения $F(x)$ на величину, не превышающую $\varepsilon = \max\{\hat{F}_j - \lambda_{j-1}, \lambda_j - \hat{F}_j\}$.*

Таким образом, качество восстановления значений функции определяется «детальностью» разбиения и вероятностью ошибки применяемого алгоритма распознавания. Основным достоинством метода CBFA является возможность получения аналитических описаний классов значений восстанавливаемых функций при условии выбора подходящих для этой цели алгоритмов обучения распознаванию. Метод CBFA был использован для построения логических описаний классов значений псевдобулевой платежной функции в матричных играх с булевыми стратегиями и частично-заданной начальной информацией [1, 2]. CBFA позволяет достаточно просто оценивать результаты аппроксимации. Недостатком CBFA является необходимость «измельчения» разбиения для получения требуемой точности и, как следствие, — большое число классов на этапе обучения (обычно до нескольких десятков).

Литература

[1] *Блыщик В. Ф.* Решение игр с булевыми стратегиями и неполной информацией на основе синтеза ДНФ // Искусственный интеллект. — 2000. — № 2. — С. 9–12.

[2] *Блыщик В. Ф.* Алгоритм построения классов значений платежной функции по прецедентной начальной информации // Искусственный интеллект. — 2006. — № 2. — С. 10–13.

Калибровка метода многоклассовой классификации один-против-всех для бустинга

Вежневец А. П., Соболев А. А., Вежневец В. П.
avezhnevets@graphics.cs.msu.ru, neusobol@yandex.ru,
dmoroz@graphics.cs.msu.ru
Москва, МГУ им. Ломоносова, лаборатория машинной графики и мультимедиа

В данной статье рассматривается задача многоклассовой классификации; показывается, что классический метод один против всех может быть существенно улучшен с помощью метода шкалирования Платта выходов бинарных классификаторов [1, 3].

Введение

Для сведения задачи классификации со многими классами к бинарной существует множество способов. Самый простой из них — один-против-всех [2]. Для бинарных классификаторов, основанных на методе опорных векторов, было показано, что [2] такой простой метод, при регуляризации бинарных классификаторов, не уступает многим более сложным и вычислительно трудным методам, основанных на самокорректирующихся кодах [4]. В данной статье показывается, что метод один-против-всех также очень эффективен и в случае использования в качестве бинарных классификаторов комитетов деревьев решений, построенных бустингом и откалиброванных алгоритмом Платта.

Описание подхода

Основная идея заключается в использовании независимого шкалирования выходов бинарных классификаторов методом Платта [1] для их лучшей согласованности, что повышает качество работы алгоритма. Пусть $f_c(x)\colon X \to R$ — бинарный классификатор, настроенный на распознавание класса $c \in Y = [1, \ldots, C]$. Предполагается что бинарный классификатор возвращает *уверенность* в том, что прецедент принадлежит классу c (например, возвращаемое значение может быть отступом от разделяющей поверхности). Тогда многоклассовый классификатор по классическому методу один-против-всех строится как

$$F(x) = \arg\max_{c \in Y} f_c(x).$$

В классическом методе не предполагается никакой калибровки выходов бинарных классификаторов. Оценим апостериорные вероятности следующим образом:

$$P(c|x) \approx \widetilde{P}(c|x) = \frac{1}{1 + \exp(Af_c(x) + B)},$$

где параметры A и B оцениваются алгоритмом Платта. Будем строить финальный классификатор следующим образом:

$$F(x) = \arg\max_{c \in Y} \widetilde{P}(c|x).$$

В результате получатся более согласованные бинарные классификаторы, что ведет к уменьшению ошибки.

Эксперименты

Для сравнения были взяты три метода: один-против-всех, самокорректирующиеся коды (ECC) и один-против-всех со шкалированием Платта. В качестве бинарных классификаторов использовались деревья классификации глубины 3, усиленные бустингом. Ниже представлены графики зависимостей ошибки на контрольных данных (использовался скользящий контроль) от итераций алгоритмов (Рис. 1): для ECC — каждая последующая точка соответствует ошибке на длине кодовых слов, большей на единицу, а для один против всех — количество итераций бинарных классификаторов. Для самокорректирующихся кодов в качестве бинарных классификаторов использовались комитеты из 40 деревьев, построенных бустингом (количество комитетов определяется длиной кодового слова).

Заключение

Эксперименты показывают, что применение шкалирования Платта к алгоритму один-против-всех дает существенное уменьшение количества ошибок этого метода — на трех из пяти представленных выборках метод один против всех, откалиброванный методом Платта, работает качественнее ECC.

Литература

[1] *J. Platt* Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. // Advances in Large Margin Classifiers, 1999. — pp. 61–74.

[2] *R. Rifkin, A. Klautau.* In Defense of One-Vs-All Classification. // The Journal of Machine Learning Research, 2004. — pp. 101–141.

[3] *A. Niculescu-Mizil and R. Caruana* Predicting good probabilities with supervised learning. // Proceedings of the 22nd international conference on Machine learning, 2005. — pp. 625–632

[4] *Chun-Nan Hsu and Yu-Shi Lin* Boosting Multiclass Learning with Repeating Codes // Journal of Artificial Intelligence Research, 2006. — pp. 263–286.

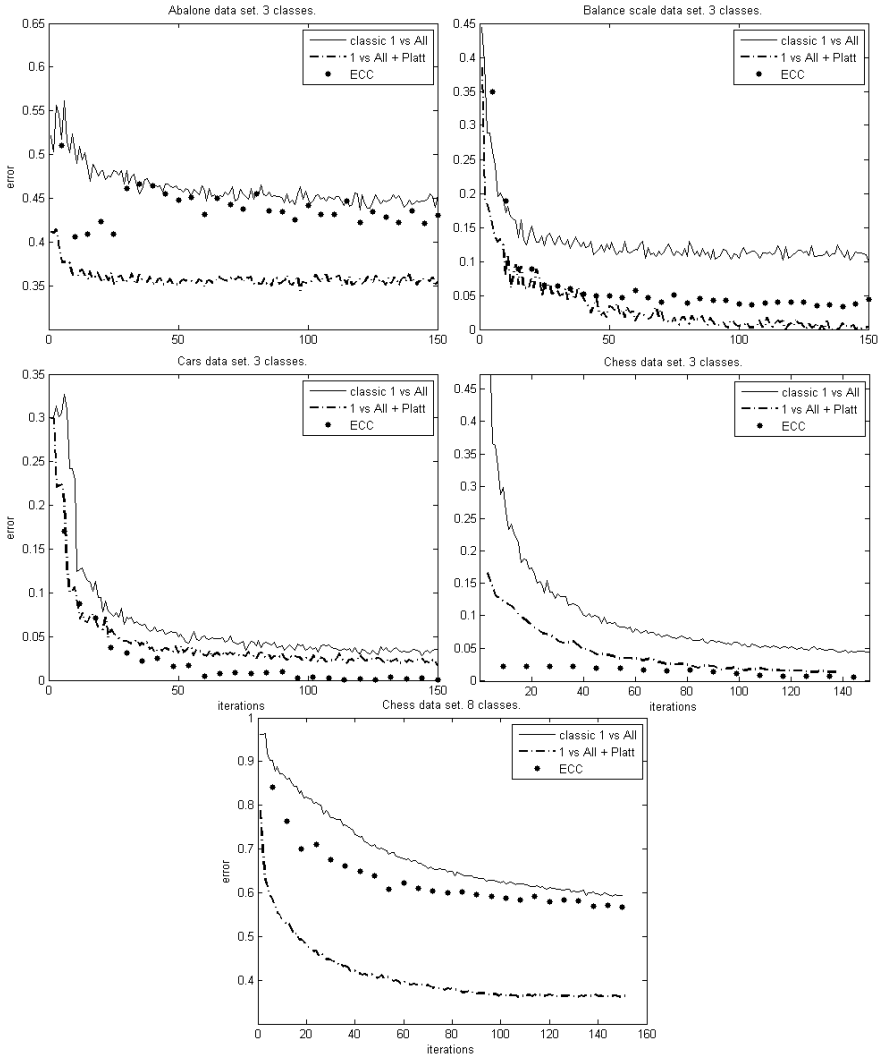


Рис. 1. Результаты экспериментов.

Проблема переобучения при отборе признаков в линейной регрессии с фиксированными коэффициентами

Венжега А. В., Ументаев С. А., Орлов А. А., Воронцов К. В.
voron@ccas.ru
Москва, Вычислительный центр РАН

Проблема отбора информативных признаков часто возникает при решении задач классификации и прогнозирования. В условиях, когда обучающая выборка мала, а признаков много, увеличивается риск переобучения — найденная функция регрессии может допускать на новых данных существенно больше ошибок, чем на обучении. Это может быть связано как с неудачным отбором признаков, так и с переоптимизацией параметров модели [1, 2]. В данной работе рассматривается задача линейной регрессии, в которой коэффициенты регрессии фиксированы, т. е. не настраиваются по обучающей выборке. Это позволяет исследовать переобучение, обусловленное исключительно отбором признаков.

Задача линейной регрессии с фиксированными коэффициентами

Рассмотрим задачу восстановления зависимости $y^*\colon X \to \mathbb{R}$ по выборке объектов $X^\ell = \{x_i\}_{i=1}^\ell \subset X$ с известными значениями $y_i = y^*(x_i)$. Объекты описываются p признаками $f_j\colon X \to \mathbb{R},\ j = 1, \ldots, p$. Зависимость восстанавливается в классе линейных функций $a_J(x) = \sum_{j\in J} w_j f_j(x)$, где $J \subseteq \{1, \ldots, n\}$ — подмножество признаков, веса w_j фиксированы, например, $w_j \equiv 1$. Требуется по обучающей выборке X^ℓ сформировать набор признаков $J = J(X^\ell)$ как можно меньшей мощности, при котором функция $a_J(x)$ аппроксимирует $y^*(x)$ на всём X как можно точнее. Качество аппроксимации на выборке $U \subset X$ будем характеризовать либо средним отклонением $E(J,U) = \frac{1}{|U|}\sum_{u\in U} |a_J(u) - y^*(u)|$, либо частотой ошибок $\nu(J,U) = \frac{1}{|U|}\sum_{u\in U}\big[|a_J(u) - y^*(u)| > \theta\big]$, где θ — порог ошибки.

Несмотря на упрощённость постановки, данная регрессионная задача имеет ряд приложений в социологии и экономике. Приведём примеры.

— Требуется выбрать представительный набор магазинов для оценивания суммарного объёма потребительского спроса. Здесь объекты — это промежутки времени, признаки соответствуют магазинам, значения признаков — это объёмы продаж некоторого товара или группы товаров в данном магазине за данный промежуток времени.
— Требуется выбрать представительный набор территориальных округов для прогнозирования результатов политических выборов по всей стране. Здесь объектами являются партии, либо пары ⟨партия,

год выборов⟩. Признаки соответствуют регионам; значения признаков — это число голосов, отданных за партию в регионе.

— Требуется выявить крупных участников биржевых торгов, совершающих покупки и продажи крупных пакетов акций синхронно с движением цены. Роль объектов играют промежутки времени, признаки соответствуют участникам, значениями признаков являются разности объёмов покупки и продажи.

Отметим, что в первых двух задачах целевая зависимость по определению есть сумма всех признаков, $y^*(x) = \sum_{j=1}^{p} f_j(x)$.

Задачи такого типа возникают, в частности, когда регулярный сбор данных по всем признакам (регионам, магазинам, участникам торгов, и т. п.) слишком дорог, либо вообще невозможен; но имеются результаты однократного сбора данных по всем признакам и возможность организовать регулярный сбор данных по части признаков. Необходимо найти набор признаков, который будет давать наиболее точные прогнозы.

Величину $\delta(X_n^\ell, X_n^k) = \nu(J_n, X_n^k) - \nu(J_n, X_n^\ell)$ будем называть *переобученностью* набора $J_n = J(X_n^\ell)$ при n-м разбиении выборки X^L на обучающую подвыборку X_n^ℓ и контрольную X_n^k, где $n \in N \subseteq \{1, \ldots, C_L^\ell\}$ — множество разбиений. В данной работе исследуется зависимость переобученности от параметров метода отбора признаков.

Методы отбора признаков

Рассматриваются три эвристических метода отбора признаков.

М1. Выбор t лучших признаков. Для данной обучающей выборки X^ℓ признаки упорядочиваются по возрастанию среднего отклонения $E(\{j\}, X^\ell)$, $j = 1, \ldots, p$, и в набор $J(X^\ell)$ включаются t первых признаков. Фактически, перебор подмножеств в этом методе отсутствует.

М2. Перебор подмножеств t из T лучших признаков является обобщением предыдущего. Признаки также упорядочиваются по возрастанию среднего отклонения, и из первых T признаков выбирается набор J, $|J| = t$, для которого значение $E(J, X^\ell)$ минимально. Если число вариантов C_T^t превышает R, перебираются R случайных наборов.

М3. Жадное добавление признаков. К набору J последовательно добавляется по одному признаку. Каждый раз добавляется такой признак j, для которого среднее отклонение $E(J \cup \{j\}, X^\ell)$ минимально.

Эксперименты и выводы

Эксперименты проводились на данных по результатам выборов в Государственную думу РФ. Число признаков (субъектов федерации) $p = 89$, число объектов (политических партий) $L = 35$, значения признаков $f_j(x_i)$ — это число голосов, отданных за i-ю партию в j-м регионе.

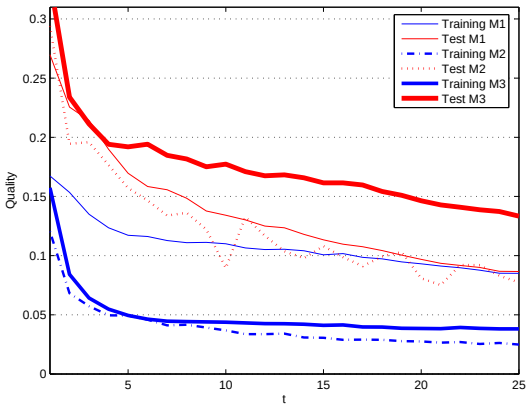


Рис. 1. Зависимость средней частоты ошибок на обучении и на контроле от числа выбранных признаков t для методов M1, M2, M3. Параметры эксперимента: $\ell = 17$, $k = 18$, $T = 35$, $R = 1300$; усреднение проводилось по $|N| = 30$ случайным разбиениям.

Оказалось, что в данной задаче переобучение, связанное с отбором признаков, возникает практически всегда.

Чем больше возможных вариантов порождает процедура отбора признаков, тем сильнее переобучение. Метод M1 наименее подвержен переобучению. Однако по критерию средней частоты ошибок на контрольной выборке он уступает методу M2, который является лидером соревнования. Наилучшие результаты M2 показывает при $T - t = 2$ или 3, то есть когда перебор делается с целью выкинуть 2–3 наименее удачных признака из T лучших признаков. Метод M3 наиболее переобучен и показывает наихудшую точность прогнозов на контроле.

Работа выполнена при поддержке РФФИ, проект № 05-01-00877 и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

[1] *Miller A.* Subset selection in regression. — Chapman & Hall/CRC, 2002.

[2] *Zhang P.* Inference after variable selection in linear regression models // Biometrika. — 1992. — No. 79. — Pp. 741–746.

Инвариантный метод настройки параметров в разреженном байесовском обучении

Ветров Д. П., Кропотов Д. А.
vetrovd@yandex.ru, dkropotov@yandex.ru
Москва, ВМиК МГУ, ВЦ РАН

Одним из популярных современных подходов для решения задач классификации является байесовское обучение, в частности, метод релевантных векторов (RVM), в котором используется независимая регуляризация весов объектов, а параметры регуляризации определяются автоматически в процессе обучения [2]. В RVM применяется регуляризация с помощью гауссовского априорного распределения. Однако, известно, что лапласовское априорное распределение может обеспечивать существенно более разреженные решающие правила [1]. Тем не менее, непосредственное применение лапласовского априорного распределения в RVM приводит к интегралам, недоступным для вычисления как аналитически, так и численно. Кроме того, метод RVM оказывается неинвариантным относительно линейных преобразований базисных функций, входящих в решающее правило.

В данной работе предлагается подход, в рамках которого возможно применение любых типов априорных распределений в разреженном байесовском обучении. При этом классификатор становится инвариантным относительно линейных преобразований базисных функций.

Допустим, что имеется набор объектов обучения $\{(\mathbf{x}_i, t_i)\}_{i=1}^n = (\mathcal{X}, \mathcal{T})$, представленных d-мерным вектором признаков $\mathbf{x} \in \mathbb{R}^d$ и меткой класса, принимающей два значения $t \in \{-1, +1\}$. В качестве семейства классификаторов выберем обобщенные линейные модели

$$y(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{M} w_i \varphi_i(\mathbf{x}) = \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle, \tag{1}$$

где $\mathbf{w}$ — набор весов, определяющих классификатор, $\varphi(\mathbf{x}) = \{\varphi_i(\mathbf{x})\}_{i=1}^N$ — набор базисных функций (обобщенных признаков). Тогда логарифм правдоподобия корректной классификации обучающей выборки может быть записан как

$$L(\mathcal{T}|\mathcal{X}, \mathbf{w}) = -\sum_{i=1}^{n} \log\bigl(1 + \exp(-t_i y(\mathbf{x}_i, \mathbf{w}))\bigr). \tag{2}$$

Множество возможных классификаторов определяется априорной функцией распределения на веса $P(\mathbf{w}|\alpha)$. Оптимальные значения весов находятся как $\mathbf{w}_{MP} = \arg\max_{\mathbf{w}} \exp(L(\mathcal{T}|\mathcal{X}, \mathbf{w}))P(\mathbf{w}|\alpha)$. В RVM в качестве априорного распределения выбирается нормальное распределение

$w_i \sim \mathcal{N}(0, \alpha_i^{-1})$ со своим параметром регуляризации α_i для каждого веса. Для поиска значений α используется максимизация обоснованности:

$$P(\mathcal{T}|\alpha) = \int P(\mathcal{T}|\mathcal{X}, \mathbf{w}) P(\mathbf{w}|\alpha) d\mathbf{w} \to \max_{\alpha}. \tag{3}$$

Интеграл (3) не берется аналитически. В RVM используется аппроксимация подынтегральной функции (регуляризованного правдоподобия) с помощью гауссианы, от которой интеграл может быть найден аналитически. При применении других типов априорных распределений, в частности, лапласовского, аппроксимация регуляризованного правдоподобия с помощью гауссианы является неадекватной.

Основная идея предлагаемого подхода заключается в аппроксимации функции правдоподобия гауссианой, интерпретации собственных векторов матрицы ковариации гауссианы в качестве новых координатных осей в пространстве весов и регуляризации вдоль этих осей с подбором коэффициентов регуляризации путем максимизации обоснованности. После такой аппроксимации значение обоснованности (3) может быть записано как

$$P(\mathcal{T}|\alpha) \approx P(\mathcal{T}|\mathcal{X}, \mathbf{w}_{ML}) \int \exp\left(\frac{1}{2}(\mathbf{w} - \mathbf{w}_{ML})^T H (\mathbf{w} - \mathbf{w}_{ML})\right) P(\mathbf{w}|\alpha) d\mathbf{w},$$

где $\mathbf{w}_{ML}$ и $(-H)^{-1}$ — математическое ожидание и ковариационная матрица аппроксимирующей гауссианы.

Перейдем к новым переменным в пространстве весов, определяемых собственными векторами матрицы H: $\mathbf{u} = Q\mathbf{w}$, где $H = Q^T \Lambda Q$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$, $\{\lambda_i\}_{i=1}^M$ — собственные значения H. Логарифм правдоподобия (2) — вогнутая функция, поэтому гессиан $H \leqslant 0$ и все его собственные значения $\{\lambda_i\}_{i=1}^M$ неположительны. Обозначим $h_i = -\lambda_i \geqslant 0$. Независимая регуляризация относительно новых переменных $\mathbf{u}$ означает, что априорная функция распределения может быть записана как

$$P(\mathbf{u}|\alpha) = \prod_{i=1}^{M} P(u_i|\alpha_i).$$

Основная цель подобной регуляризации — это представление обоснованности (3) в виде произведения одномерных интегралов

$$P(\mathcal{T}|\alpha) \approx P(\mathcal{T}|\mathcal{X}, \mathbf{u}_{ML}) \prod_{i=1}^{M} \underbrace{\int \exp\left(-\frac{h_i}{2}(u_i - u_{ML,i})^2\right) P(u_i|\alpha_i) du_i}_{f_i(h_i, u_{ML,i}, \alpha_i)}. \tag{4}$$

Каждый из одномерных интегралов может быть взят аналитически либо численно в зависимости от используемого типа регуляризатора. Поиск оптимальных значений α можно осуществлять независимо для каждого α_i, решая одномерную задачу оптимизации. Такая процедура обучения получила название метода релевантных собственных векторов (REVM).

В случае использовании гауссовского априорного распределения $u_i \sim \mathcal{N}(0, \alpha_i^{-1})$ значения одномерных интегралов $f_i(h_i, u_{ML,i}, \alpha_i)$ в выражении (4) могут быть вычислены аналитически:

$$f_i(h_i, u_{ML,i}, \alpha_i) = \sqrt{\frac{\alpha_i}{h_i + \alpha_i}} \exp\left(-\frac{h_i \alpha_i u_{ML,i}^2}{2(h_i + \alpha_i)}\right). \tag{5}$$

В зависимости от значений h_i и $u_{ML,i}$ интеграл (5), как функция от α_i, имеет один максимум либо монотонно возрастает на интервале $[0, +\infty)$. Приравнивая производную (5) по α_i к нулю, получаем оптимальное значение α_i:

$$\alpha_i^{\text{опт}} = \begin{cases} \frac{h_i}{h_i u_{ML,i}^2 - 1}, & \text{если } h_i u_{ML,i}^2 > 1; \\ +\infty, & \text{иначе.} \end{cases}$$

Теорема 1. *Рассмотрим наборы базисных функций $\varphi_1(\mathbf{x})$ и $\varphi_2(\mathbf{x}) = A\varphi_1(\mathbf{x})$, где A — невырожденная матрица размера $M \times M$. Обозначим решающее правило вида* (1)*, полученное методом классикации по набору базисных функций φ по выборке $(\mathcal{X}, \mathcal{T})$ через $y(\mathbf{x}; \varphi, \mathcal{X}, \mathcal{T})$. Тогда*

$$y_{RVM}(\mathbf{x}; \varphi_1, \mathcal{X}, \mathcal{T}) \not\equiv y_{RVM}(\mathbf{x}; \varphi_2, \mathcal{X}, \mathcal{T});$$
$$y_{REVM}(\mathbf{x}; \varphi_1, \mathcal{X}, \mathcal{T}) \equiv y_{REVM}(\mathbf{x}; \varphi_2, \mathcal{X}, \mathcal{T}).$$

Результаты экспериментов показывают, что по сравнению с RVM, REVM работает на порядок быстрее и приводит к более разреженным решающим правилам (в терминах переменных $\mathbf{u}$).

Работа выполнена при поддержке РФФИ, проекты № 06-01-08045, № 05-07-90333, № 06-01-00492, № 07-01-00211.

Литература

[1] *Williams P. M.* Bayesian regularization and pruning using a Laplace prior // Neural Computation. — 1995. — V. 7, № 1. — Pp. 117–143.

[2] *Tipping M. E.* Sparse Bayesian Learning and the Relevance Vector Machine // Journal of Machine Learning Research. — 2001. — Vol. 1, № 5. — Pp. 211–244.

О выборе наилучшего квадратичного регуляризатора в обобщенных линейных моделях классификации

Ветров Д. П., Кропотов Д. А.
VetrovD@yandex.ru, DKropotov@yandex.ru
Москва, ВМиК МГУ, ВЦ РАН

Обобщенные линейные модели (generalized linear models) в последние годы являются популярным средством решения задач классификации и восстановления регрессии. Примерами таких моделей могут служить методы опорных и релевантных векторов, логистическая регрессия, и др. Настройка весов производится путем оптимизации суммы некоторого функционала качества, связанного с ошибкой на обучающей выборке, и регуляризатора, предотвращающего перенастройку на данные.

Рассмотрим стандартную задачу классификации на два класса по заданной обучающей выборке $(X, T) = \{\mathbf{x}_i, t_i\}_{i=1}^m$, где $\mathbf{x} \in \mathbb{R}^d$, а $t \in \{-1, 1\}$.

Статистические модели обучения

При статистическом подходе в качестве функционала, связанного с ошибкой на обучении, используется логарифм правдоподобия правильной классификации обучающей выборки

$$L(T|X, \mathbf{w}) = -\sum_{i=1}^{m} \log\left(1 + \exp\left(-t_i \sum_{j=1}^{N} w_j \varphi_j(\mathbf{x}_i)\right)\right), \qquad (1)$$

где $\{\varphi_j(\mathbf{x})\}_{j=1}^N$ — базисные функции, зафиксированные до начала обучения. Настройка весов производится путем оптимизации суммы (1) и регуляризатора, штрафующего большие значения весов $\mathbf{w}$ во избежание перенастройки на данные. Наиболее популярным регуляризатором является квадратичный, в простейшем случае имеющий вид

$$R_l(\mathbf{w}, \lambda) = -\lambda \mathbf{w}^{\mathrm{T}} I \mathbf{w} = -\lambda \sum_{j=1}^{N} w_j^2.$$

По такой схеме работает классический метод логистической регрессии, в котором коэффициент λ подбирается с помощью трудоемкой процедуры кросс-валидации. Популярность квадратичного регуляризатора обуславливается, помимо прочего, тем, что он соответствует ридж-регуляризации гессиана (часто вырожденного или плохо обусловленного) логарифмического правдоподобия, облегчающей процедуру оптимизации весов.

В методе релевантных векторов [3] используется более сложный регуляризатор вида

$$R_r(\mathbf{w}, \Lambda) = -\mathbf{w}^{\mathrm{T}}\Lambda\mathbf{w} = -\sum_{j=1}^{N} \lambda_j w_j^2,$$

где Λ — неотрицательная диагональная матрица. Таким образом, каждому весу присваивается собственный коэффициент регуляризации. Выбор значений λ_j проводится с помощью процедуры байесовского обучения путем максимизации *обоснованности* модели

$$\Lambda = \arg\max E(\Lambda) = \arg\max \int \exp\left(L(T|X, \mathbf{w}) + R_r(\mathbf{w}, \Lambda)\right) d\mathbf{w}.$$

Недиагональная регуляризация

Авторы предлагают рассмотреть более общий случай произвольной неотрицательно определенной матрицы регуляризации

$$R_i(\mathbf{w}) = -\mathbf{w}^{\mathrm{T}}A\mathbf{w}, \quad A^{\mathrm{T}} = A, \quad A \geqslant 0.$$

Настройка матрицы A производится также путем оптимизации обоснованности модели $E(A)$.

Пусть $\mathbf{w}_{ML} = \arg\max L(T|X, \mathbf{w})$, $H = -\nabla\nabla L(T|X, \mathbf{w})|_{\mathbf{w}=\mathbf{w}_{ML}} \geqslant 0$. Обозначим $M = H(H + A)^{-1}A$. Тогда можно показать [1], что

$$\frac{\partial \log E(A)}{\partial A} = -0.5\left[M^{-1} - \mathbf{w}_{ML}\mathbf{w}_{ML}^{\mathrm{T}}\right].$$

Приравнивая производную нулю отсюда можно получить

$$A^{-1} = \mathbf{w}_{ML}\mathbf{w}_{ML}^{\mathrm{T}} - H^{-1}.$$

Матрица A является симметричной, но имеет не более одного положительного собственного значения. В силу симметрии существует ортогональная матрица U, такая что $D = U^{\mathrm{T}}A^{-1}U$ является диагональной матрицей. Заменяя все отрицательные собственные значения матрицы D^{-1} на $+\infty$ (что соответствует наибольшему значению обоснованности среди неотрицательных собственных значений в случае, когда экстремум достигается в отрицательной области) получаем, что матрица регуляризации A будет штрафовать с бесконечно большим коэффициентом все значения весов $\mathbf{w}$, кроме тех, которые лежат вдоль вектора $\mathbf{u}$, соответствующего единственному положительному собственному значению d^{-1}

Задача	**SVM**	**RVM**	**LogReg**	**IREVM**	**VRVM**
Bupa	28.46	33.33	57.97	29.68	30.78
Heart	19.33	18.15	22.44	18.00	17.56
Hepatitis	17.55	14.32	19.48	16.26	13.03
Votes	4.78	5.56	5.98	4.92	5.61
WPBC	21.72	23.64	23.84	21.11	24.04
Laryngeal1	17.75	16.81	19.44	17.28	17.37
Weaning	10.99	15.70	13.31	12.72	13.64
Ранг	**17.00**	**21.00**	**32.00**	**14.00**	**21.00**
Цвет	Место 1	Место 2	Место 3	Место 4	Место 5

Таблица 1. Ошибки классификаторов (в %).

в матрице D^{-1}. Соответственно, в ходе обучения решается задача одномерной оптимизации $\mathbf{w}_{MP} = \theta\mathbf{u} = \arg\max_{\theta} \left(L(T|X, \theta\mathbf{u} + d^{-1}\theta^2)\right)$.

Заметим, что в отличие от метода релевантных векторов, в предложенном подходе не требуется итеративного подбора коэффициентов регуляризации, а формула для оптимального регуляризатора сразу выписывается в явном виде.

Эксперименты

В таблице 1 представлены результаты экспериментов, проведенные на реальных задачах из репозитория UCI [4]. Сравнение проведено с классическим (RVM) и вариационным (VRVM) [2] методом релевантных векторов, методом опорных векторов (SVM) и логистической регрессии (LogReg). Отдельно можно отметить, что скорость обучения практически не отличается от скорости обучения логистической регрессии и значительно быстрее обучения метода релевантных векторов.

Работа выполнена при поддержке РФФИ, проекты № 06-01-08045, № 05-07-90333, № 06-01-00492 и № 07-01-00211.

Литература

[1] *Kropotov D. A., Vetrov D. P.* Optimal Bayesian Linear Classifier with Arbitrary Gaussian Regularizer // 7th Open German-Russian Workshop on Pattern Recognition and Image Understanding (OGRW2007), Ettlingen, 2007.

[2] *Bishop C. M., Tipping M. E.* Variational Relevance Vector Machines // Uncertainty in artificial intelligence (UAI-2000), 2000— P. 46–53.

[3] *Tipping M. E.* Sparse Bayesian Learning and the Relevance Vector Machine // Journal of Machine Learning Research. — 2001. — Vol. 1, № 5. — P. 211–244.

[4] *Asuncion A., Newman D. J.* UCI (Machine Learning Repository) — 2007. — `www.ics.uci.edu/~mlearn/MLRepository.html`.

Новый метод обучения байесовской логистической регрессии с использованием лапласовского регуляризатора

Ветров Д. П., Кропотов Д. А., Курчин О. В.
VetrovD@yandex.ru, DKropotov@yandex.ru, 4education@mail.ru
Москва, ВМиК МГУ, ВЦ РАН, ВМиК МГУ

Рассмотрим стандартную задачу классификации на два класса по заданной обучающей выборке $\mathcal{D} = (X, T) = \{\mathbf{x}_i, t_i\}_{i=1}^m$, где $\mathbf{x} \in \mathbb{R}^d$, $t \in \{-1, 1\}$. Основным недостатком многих алгоритмов классификации является эффект переобучения. Одним из наиболее популярных подходов к его устранению является байесовская регуляризация, суть которой заключается в том, что задаются априорные распределения вероятности некоторых (возможно всех) весов классификатора. При этом настройка весов производится путем оптимизации суммы некоторого функционала качества, связанного с ошибкой на обучающей выборке, и регуляризатора, предотвращающего перенастройку на данные. Данную концепцию используют такие широко известные модели, как метод опорных векторов, метод релевантных векторов, логистическая регрессия.

Логистическая регрессия

Классическим подходом, особенно популярным при решении задач классификации в медицине, позволяющим вычислить апостериорную вероятность принадлежности объекта к одному из двух классов, является логистическая регрессия, основанная на линейной комбинации признаков объектов:

$$p(t|\mathbf{x}) = \frac{1}{1 + \exp(-t\hat{y}(\mathbf{x}))},$$

где $t \in \mathcal{Y} = \{-1; 1\}$, $\hat{y}(\mathbf{x}) = w_0 + w_1 x^1 + \cdots + w_d x^d$. Тогда отрицательный логарифм правдоподобия принимает вид:

$$L(\mathcal{Y}^m|\mathcal{X}^m, \mathbf{w}) = \sum_{i=1}^{n} \ln\bigl(1 + \exp(-t_i\hat{y}(\mathbf{x}_i))\bigr). \tag{1}$$

Поиск значений весов $\mathbf{w}$ осуществляется путём максимизации (1). Полученный классификатор обладает следующей особенностью: практически ни один из его весов w_i не равен в точности нулю. Добавление к критерию (1) лапласовского регуляризатора позволяет получать более разреженные решения [1]:

$$F = L(\mathcal{Y}^m|\mathcal{X}^m, \mathbf{w}) + \lambda R(\mathbf{w}), \tag{2}$$

где $R(\mathbf{w}) = \sum_{i=1}^d |w_i|$.

Устранение параметра распределения λ

Заметим, что параметр распределения λ в критерии (2) заранее неизвестен. Для его определения может быть проведено усреднение по данному параметру при использовании для него некоторого априорного распределения. Поскольку данный параметр является параметром масштаба, то закономерным является использование несобственного распределения Джеффри, $p(\lambda) \propto 1/\lambda$ — аналог равномерного распределения в логарифмической шкале. Условная плотность вероятности $p(\mathbf{w}|\lambda)$ для лапласовского распределения вычисляется как

$$p(\mathbf{w}|\lambda) = \left(\frac{\lambda}{2}\right)^N \exp\bigl(-\lambda R(\mathbf{w})\bigr) = \prod_{i=1}^{N} \frac{\lambda}{2} \exp\bigl(-\lambda|w_i|\bigr),$$

где N — количество ненулевых весов классификатора.

Для определения априорной плотности распределения весов классификатора и устранения параметра λ производится усреднение $p(\mathbf{w}) = \int p(\mathbf{w}|\lambda)p(\lambda)d\lambda$. В результате приходим к следующему критерию оптимизации:

$$Q = L(\mathcal{Y}^m|\mathcal{X}^m, \mathbf{w}) + N \ln R(\mathbf{w}). \tag{3}$$

Полученная модель (BLogReg), в основе которой лежит логистическая регрессия, была разработана в 2006 году английскими учеными Коули и Тэлбот [3]. Ими же была предложена процедура настройки данной модели. Основным достоинством данной модели является разреженность получаемого решения. Однако, разрывность целевой функции (3) привела к тому, что данная процедура могла использовать только метод оптимизации первого порядка, что существенно сказывалось на длительности процедуры обучения. В работе [2] был предложен похожий подход усреднения путем интегрирования по параметру масштаба с использованием распределения Джеффри для гауссовского распределения.

Разработанный метод настройки

Целевая функция (3) не является всюду гладкой, и, более того, является разрывной, поэтому мы не можем использовать методы оптимизации второго порядка. Для решения данной проблемы предлагается заменить критерий (3) на его непрерывный аналог:

$$\hat{Q} = L(\mathcal{T}|\mathcal{X}, \mathbf{w}) + \hat{N} \log R(\mathbf{w}); \tag{4}$$

$$\hat{N} = n - \sum_{i=1}^{n} \exp\left(-\frac{w_i^2}{2\sigma^2}\right). \tag{5}$$

Здесь $\sigma > 0$ — некоторый положительный коэффициент нечеткости. Рассмотрим гипероктант $\mathcal{H}_{ML}$, в котором находится точка максимума

правдоподобия (1). Тогда можно показать, что достаточно проводить оптимизацию (4) в области $\mathcal{M} = \{\bar{w} \in \mathcal{H}_{ML}, |w_i| \geqslant \varepsilon > 0, \forall i = 1, \ldots, n\}$. При этом критерий (4) является гладким в области $\mathcal{M}$ и для оптимизации может быть использован метод Ньютона второго порядка с ограничениями, что приводит к значительному увеличению быстродействия процедуры обучения.

Теорема 1. *Пусть ε — погрешность оптимизации, а σ — коэффициент нечеткости в выражении* (5)*. Тогда применение нечеткой версии критерия* (4) *возможно тогда и только тогда, когда*

$$\frac{1}{\sigma^2} = \bar{o}(\varepsilon^{-2} \ln^{-1} \varepsilon).$$

Эксперименты

Был проведен ряд экспериментов на задачах из UCI репозитория [4] по сравнению разработанного метода с методами опорных и релевантных векторов, а также байесовской логистической регрессии. Полученные результаты свидетельствуют о том, что разработанный метод позволяет получить сравнимое с остальными методами качество распознавания и разреженность, но при этом обучается быстрее BLogReg. На выборках размерности порядка 10 скорость обучения выросла в 2–5 раз, на выборках размерности порядка 100 — в 8–20 раз.

Литература

[1] *Williams, P. M.* Bayesian regularization and pruning using a Laplace prior. // Neural Computation. — 1995. — Vol. 7. — P. 117–143.

[2] *Figueiredo M.* Adaptive sparseness for supervised learning. // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2007. — Vol. 25. — P. 1150–1159.

[3] *Cawley G. C., Talbot N. L.* Gene selection in cancer classiffication using sparse logistic regression with bayesian regularization. // Bioinformatics. — 2007. — Vol. 22. — P. 2348–2355.

[4] *Asuncion A., Newman D. J.* UCI Machine Learning Repository — 2007. — `www.ics.uci.edu/~mlearn/MLRepository.html`.

Расширение метода Expectation Propagation на случай логистического правдоподобия

Ветров Д. П., Кропотов Д. А., Пташко Н. О.
vetrovd@yandex.ru, dkropotov@yandex.ru, ptashko@inbox.ru
Москва, ВМиК МГУ, ВЦ РАН

Рассматривается стандартная задача классификации на два класса. Дана обучающая выборка $D = \{(x_i, t_i)\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^d$, $t_i \in \{-1, 1\}$. Алгоритм классификации строится в виде $y(\mathbf{x}) = \text{sign}(\mathbf{w}^T\varphi(\mathbf{x}))$, где $\mathbf{w} \in \mathbb{R}^M$, $\varphi(\mathbf{x}) = [\varphi_1(\mathbf{x}), \ldots, \varphi_M(\mathbf{x})]^{\text{T}}$ — некоторые базисные функции. Значения $\mathbf{w}$ находятся путем максимизации регуляризованного правдоподобия:

$$p(\mathbf{t}|X, \mathbf{w})p(\mathbf{w}|\alpha) = p(\mathbf{w}|\alpha)\prod_{i=1}^N p(t_i|\mathbf{x}_i, \mathbf{w}) = p(\mathbf{w}|\alpha)\prod_{i=1}^N \Psi(t_i\mathbf{w}^T\varphi(\mathbf{x}_i); 0, 1),$$

где $\mathbf{t} = \{t_i\}_{i=1}^N$, $X = \{\mathbf{x}_i\}_{i=1}^N$, $\Psi(y; m, s^2) = \frac{1}{\sqrt{2\pi}s}\int\limits_{-\infty}^{y-m} \exp\left(-\frac{x^2}{2s^2}\right) dx$ — гауссовская функция распределения (пробит-функция), $p(\mathbf{w}|\alpha) \sim \mathcal{N}(0, A^{-1})$, где $A = \text{diag}(\alpha_1, \ldots, \alpha_M)$. Значения гиперпараметров α находятся с помощью максимизации обоснованности [1]:

$$p(\mathbf{t}|X, \alpha) = \int p(\mathbf{t}|X, \mathbf{w})p(\mathbf{w}|\alpha)d\mathbf{w} \to \max_\alpha.$$

Данный интеграл, как правило, не удается вычислить аналитически. Поэтому возникает проблема его адекватной аппроксимации. Одним из популярных способов решения этой задачи является алгоритм expectation propagation (EP) [2].

Алгоритм Expectation Propagation

Алгоритм EP использует тот факт, что правдоподобие является произведением простых множителей. Аппроксимируя каждый из них, получаем аппроксимацию всего апостериорного распределения.

$$p(\mathbf{w}|X, \mathbf{t}, \alpha) \propto p(\mathbf{w}|\alpha)\prod_{i=1}^N p(t_i|\mathbf{x}_i, \mathbf{w}) \equiv p(\mathbf{w}|\alpha)\prod_{i=1}^N g_i(\mathbf{w}) \approx$$

$$\approx p(\mathbf{w}|\alpha)\prod_{i=1}^N \widetilde{g}_i(\mathbf{w}) = q(\mathbf{w}).$$

Алгоритм аппроксимирует каждый множитель таким образом, чтобы результирующее апостериорное распределение было близко к аппроксимированному в смысле меры Кульбака-Лейблера (называемой также

Алгоритм 1. Expectation Propagation.

Вход: $g_i(x), i = 1, \ldots, N$;
Выход: $\widetilde{g}_i(x), i = 1, \ldots, N$;

1: инициализация: $\widetilde{g}_i = 1$; $v_i = \infty$; $m_i = 0$; $s_i = 1$; $q(\mathbf{w}) = p(\mathbf{w}|\alpha)$;
2: **цикл** *// пока все* (m_i, v_i, s_i) *не сойдутся*
3: **для** $i = 1, \ldots, N$
4: (a) Удаляем $\widetilde{g}_i$ из $q(\mathbf{w})$.
Получим $q^{\setminus i}(\mathbf{w}) \propto q(\mathbf{w})/\widetilde{g}_i \sim \mathcal{N}(\mathbf{m}_{\mathbf{w}}^{\setminus i}, \mathbf{V}_{\mathbf{w}}^{\setminus i})$;
$\mathbf{V}_{\mathbf{w}}^{\setminus i} = \mathbf{V}_{\mathbf{w}} + \frac{V_{\mathbf{w}}\varphi_i(V_{\mathbf{w}}\varphi_i)^{\mathrm{T}}}{v_i - \varphi_i^{\mathrm{T}} V_{\mathbf{w}}\varphi_i}$; $\mathbf{m}_{\mathbf{w}}^{\setminus i} = \mathbf{m}_{\mathbf{w}} + (\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i)v_i^{-1}(\varphi_i^{\mathrm{T}}\mathbf{m}_{\mathbf{w}} - m_i)$;
5: (b) Полагаем $\widehat{p}(\mathbf{w}) \propto g_i(\mathbf{w})q^{\setminus i}(\mathbf{w})$.
Находим $q(\mathbf{w})$, минимизирующее $KL(\widehat{p}(\mathbf{w})\|q(\mathbf{w}))$;
$\mathbf{m}_{\mathbf{w}} = \mathbf{m}_{\mathbf{w}}^{\setminus i} + \mathbf{V}_{\mathbf{w}}^{\setminus i}\rho_i\varphi_i$; $\mathbf{V}_{\mathbf{w}} = \mathbf{V}_{\mathbf{w}}^{\setminus i} + (\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i)\frac{\rho_i(\varphi_i^{\mathrm{T}}\mathbf{m}_{\mathbf{w}} + \rho_i)}{\varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i + 1}(\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i)^{\mathrm{T}}$;
$Z_i = \int_{\mathbf{w}} g_i(\mathbf{w})q^{\setminus i}(\mathbf{w})d\mathbf{w} = \Psi(z_i)$, где
$z_i = \frac{(\mathbf{m}_{\mathbf{w}}^{\setminus i})^{\mathrm{T}}\varphi_i}{\sqrt{\varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i + 1}}$; $\rho_i = \frac{1}{\sqrt{\varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i + 1}}\frac{\mathcal{N}(z_i;0,1)}{\Psi(z_i)}$;
6: (c) Используя $\widetilde{g}_i = Z_i\frac{q(\mathbf{w})}{q^{\setminus i}(\mathbf{w})}$, получаем:
$v_i = \varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i\left(\frac{1}{\rho_i(\varphi_i{}^{\mathrm{T}}\mathbf{m}_{\mathbf{w}} + \rho_i)} - 1\right) + \frac{1}{\rho_i(\varphi_i{}^{\mathrm{T}}\mathbf{m}_{\mathbf{w}} + \rho_i)}$;
$m_i = \varphi_i^{\mathrm{T}}\mathbf{m}_{\mathbf{w}}^{\setminus i} + (v_i + \varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i)\rho_i$;
$s_i = \Psi(z_i)\sqrt{1 + v_i^{-1}\varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i}\exp\left(\frac{1}{2}\frac{\varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i + 1}{\varphi_i^{\mathrm{T}}\mathbf{m}_{\mathbf{w}}^{\setminus i} + \rho_i}\rho_i\right)$;
7: Подсчет нормализующей константы и обоснованности:
$B = (\mathbf{m}_{\mathbf{w}})^{\mathrm{T}}(\mathbf{V}_{\mathbf{w}})^{-1}(\mathbf{m}_{\mathbf{w}}) - \sum_i \frac{m_i^2}{v_i}$;
$p(\mathbf{t}|X, \alpha) \approx \int\prod_{i=1}^N \widetilde{g}_i(\mathbf{w})d\mathbf{w} = \frac{|\mathbf{V}_{\mathbf{w}}|^{1/2}}{(\prod_j \alpha_j)^{1/2}}\exp(B/2)\prod_i s_i$;

KL-дивергенцией):

$$KL(p\|q) = \int q(x)\log\frac{p(x)}{q(x)}dx.$$

Каждый аппроксимирующий член $\widetilde{g}_i(\mathbf{w})$ выбирается равным $\widetilde{g}_i(\mathbf{w}) = s_i\exp\left(-\frac{1}{2v_i^2}(t_i\mathbf{w}^T\varphi(\mathbf{x}_i) - m_i)^2\right)$, где (m_i, v_i, s_i) — параметры.

Можно показать, что если функция $q(\mathbf{w})$ — гауссиана, то условие близости эквивалентно условию равенства первых и вторых моментов функций $p(\mathbf{w}|X, \mathbf{t}, \alpha)$ и $q(\mathbf{w})$. Алгоритм 1 представляет собой метод EP (для краткости $t_i\varphi_i$ обозначено как φ_i, а $\Psi(z; 0, 1)$ через $\Psi(z)$).

Модификация EP

Заметим, что приведенный алгоритм неприменим для популярной модели логистической регрессии, в которой правдоподобие задается выражением

$$p_\sigma(\mathbf{t}|X,\mathbf{w}) = \prod_{i=1}^{N} \sigma(t_i \mathbf{w}^T \varphi(\mathbf{x}_i)) = \prod_{i=1}^{N} \frac{1}{1+\exp(-t_i \mathbf{w}^T \varphi(\mathbf{x}_i)))}.$$

Для этого при текущем векторе α находим максимум $\mathbf{w}_{MP}$ функции $p(\mathbf{w}|X,\mathbf{t},\alpha)$. Далее для каждого $i = 1,\ldots,N$ приближаем функцию $\sigma(y)$ в точке $y^i_{MP} = t_i \mathbf{w}^T_{MP} \varphi(\mathbf{x}_i)$ пробит-функцией $\Psi(y; m_i, {s_i}^2)$, получая приближение $p_\Psi(\mathbf{w}|\mathbf{t},\alpha) = \prod_{i=1}^{N} \Psi(y; m_i, {s_i}^2)$. Параметры пробит-функции m_i и ${s_i}^2$ находятся из требований совпадения значения нулевой и первой производных логистической и пробит-функции в точке y^i_{MP}

$${s_i}^2 = \frac{1}{\sqrt{2\pi} S(1-S)} \exp\left(-\frac{1}{2}\Psi^{-1}(S;0,1)\right), \quad m_i = y^i_{MP} - \Psi^{-1}(S;0,1)s_i,$$

где $S = \sigma(y^i_{MP})$, а $\Psi^{-1}(y;0,1)$ — функция, обратная к пробит-функции[1]. Далее применяем алгоритм EP для поиска приближения функции $p_\Psi(\mathbf{w}|\mathbf{t},\alpha) = \prod_{i=1}^{N} \Psi(y; m_i, {s_i}^2)$.

Следует отметить, что предлагаемое в работе приближение использует тот факт, что логистическая и пробит-функция очень близки по значениям (и значит можно приблизить одну другой), в то время как их логарифмы (которые используются при обучении логистической и пробит-регрессии соответственно) сильно отличаются, поэтому сами методы классификации приводят к существенно различным решающим правилам. В частности, пробит-регрессия значительно менее робастна.

Работа выполнена при поддержке РФФИ, проекты №№ 07-01-00211, 06-01-08045, 05-07-90333.

Литература

[1] *Tipping M.* Sparse Bayesian Learning and the Relevance Vector Machine // Journal of Machine Learning Research. — 2001. — Vol. 1, № 5. — P. 211–244.

[2] *Qi Y. A., Minka T. P., Picard R. W., Ghahramani Z.* Predictive Automatic Relevance Determination by Expectation Propagation // 21-st International Conference on Machine Learning, Banff, Canada, 2004.

[1] Реализованная, например, в среде MATLAB в виде процедуры norminv.

Проблема переобучения функций близости при построении алгоритмов вычисления оценок

Воронцов К. В., Ульянов Ф. М.
voron@ccas.ru
Москва, Вычислительный центр РАН

Модель алгоритмов вычисления оценок (АВО) была предложена Ю. И. Журавлёвым в начале 70-х [1]. В данной работе предлагается новый метод обучения АВО, основанная на *принципе явной максимизации отступов* — direct optimization of margin [4]. Известно, что максимизация отступов повышает обобщающую способность линейных композиций классификаторов. АВО как раз и является такой композицией, причём роль базовых классификаторов в ней играют функции близости. Вводится понятие переобученности функций близости и предлагается эмпирическая методика подбора управляющих параметров, позволяющая снижать переобученность функций близости в процессе их построения.

Постановка задачи и используемый вариант АВО

Пусть X — пространство объектов, Y — множество имён классов, $y^*\colon X \to Y$ — целевая функция, значения которой известны только на объектах конечной обучающей выборки $X^\ell = (x_i, y_i)_{i=1}^\ell \subset X \times Y$, $y_i = y^*(x_i)$. Задача заключается в том, чтобы построить алгоритм классификации $a\colon X \to Y$, аппроксимирующий y^* на всём множестве X.

Пусть на множестве X заданы функции расстояния $r_j\colon X \times X \to \mathbb{R}_+$, $j = 1, \ldots, n$, не обязательно метрики. Когда объекты из X описываются n признаками $f_j\colon X \to \mathbb{R}$, можно положить $r_j(x, x') = \bigl|f_j(x) - f_j(x')\bigr|$.

Рассмотрим алгоритм вычисления оценок, имеющий вид

$$a(x) = \arg\max_{y \in Y} \Gamma_y(x); \qquad \Gamma_y(x) = \sum\nolimits_{i \in I_y} B_i(x);$$

где $\Gamma_y(x)$ — оценка объекта x за класс y; $I_y \subseteq \{1, \ldots, \ell\}$ — множество индексов «наиболее типичных» обучающих объектов класса y, называемых *эталонами*; $B_i(x)$ — функция близости, оценивающая сходство объекта x с эталоном x_i по *опорному множеству* $\omega_i \subseteq \{1, \ldots, n\}$. В данной работе функции близости задаются в виде

$$B_i(x) = \bigl[\, \rho_i(x, x_i) \leqslant R_i \,\bigr]; \qquad \rho_i(x, x_i) = \sum\nolimits_{j \in \omega_i} \tfrac{1}{\varepsilon_j} r_j(x, x_i);$$

где ε_j — нормировочные коэффициенты. Каждая функция близости $B_i(x)$ представляет собой *шар* с центром в эталонном объекте x_i и *радиусом* R_i относительно расстояния $\rho_i(x, x_i)$. Будем говорить, что шар $B_i(x)$ *выделяет* объект x, если $B_i(x) = 1$. Итак, алгоритм $a(x)$ задаётся набором параметров $\langle I_y, \omega_i, x_i, R_i, \varepsilon_j \rangle$, где $y \in Y$, $i \in I_y$, $j = 1, \ldots, n$.

Особенностью данного варианта АВО является то, что с каждым эталонным объектом x_i связывается своё (и ровно одно) опорное множество ω_i и свой радиус R_i, однозначно задающие шар $B_i(x)$.

Метод обучения АВО

Ставится задача построить надёжный, хорошо интерпретируемый алгоритм. Для этого шаров должно быть не слишком много; опорные множества ω_i должны иметь невысокую мощность; каждый шар $B_i(x)$ должен быть закономерностью класса y_i, т. е. выделять как можно больше объектов класса y_i и как можно меньше объектов остальных классов; наконец, шар должен обладать обобщающей способностью, т. е. оставаться закономерностью класса y_i на объектах, не вошедших в состав обучения.

Предлагается следующий метод обучения данного варианта АВО.

Сначала вычисляются нормировочные коэффициенты ε_j как среднее значение функции расстояния ρ_j по всем парам обучающих объектов.

Затем начинается поиск «хороших» шаров, реализуемый тремя вложенными циклами перебора. На внешнем цикле обучающие объекты по очереди рассматриваются как кандидаты в эталоны (центры шаров). Для каждого кандидата x_i методом случайного поиска с адаптацией [2] перебираются опорные множества ω_i. Для каждого опорного множества перебираются такие значения радиуса R_i, при которых шар $B_i(x)$ выделяет различные (по составу объектов) подвыборки. Из построенных шаров выбирается тот, который максимизирует критерий $W(B_i)$, определяемый ниже.

Отступом объекта $x \in X$ называется величина

$$M(x) = \Gamma_{y^*(x)}(x) - \max_{y \in Y \setminus \{y^*(x)\}} \Gamma_y(x).$$

Чем больше $M(x)$, тем надёжнее классифицируется объект x. Известно, что оптимальным с точки зрения понижения вероятности ошибки является такое распределение отступов, при котором все они принимают одинаковое и как можно большее значение [4].

Добавление шара $B_i(x)$ изменяет значение отступа $M(x)$ на величину $m(x) = B_i(x)\big(2\,[y_i{=}y^*(x)] - 1\big) \in \{\pm 1, 0\}$. Это изменение поощряется или наказывается путём назначения объекту x веса $w(M(x), m(x))$, где функция $w(M, m)$ задаётся из следующих эвристических соображений:

— увеличение малых (близких к нулю) значений отступа $M(x)$ должно поощряться, уменьшение — наказываться;
— для больших положительных значений отступа, наоборот, уменьшение должно поощряться, увеличение — наказываться, так как это способствует выравниванию распределения отступов;

— для наименьших отрицательных значений отступа увеличение должно поощряться, если ставится задача безошибочно классифицировать обучающую выборку; в противном случае объект может считаться шумовым выбросом, тогда поощряться должно уменьшение отступа.

Критерием качества шара является суммарный вес покрытых им объектов обучающей выборки: $W(B_i) = \sum_{i=1}^{\ell} w(M(x_i), m(x_i))$. Максимизация данного критерия приводит к тому, что каждый следующий шар стремится допустить как можно меньше ошибок, и при этом покрыть объекты, неуверенно классифицируемые композицией всех предыдущих шаров. Одновременно выделяются объекты-выбросы.

Эмпирическое оценивание переобученности шаров

Качество шара $B_i(x)$ на конечной выборке $U \subset X$ характеризуется частотой его ошибок $\nu(B_i, U) = \frac{1}{|U|} \sum_{x \in U} \big[B_i(x) \neq [y^*(x) = y_i]\big]$.

Допустим, что шар $B_i(x)$ был построен по обучающей выборке X^ℓ и имеется непересекающаяся с ней контрольная выборка X^k. *Переобученностью* шара $B_i(x)$ называется разность частоты его ошибок на контроле и на обучении: $\delta(B_i, X^\ell, X^k) = \nu(B_i, X^k) - \nu(B_i, X^\ell)$.

Предлагается методика эмпирического оценивания переобученности, основанная на скользящем контроле. Фиксируется множество разбиений *полной* выборки X^L на обучающую и контрольную, $X^L = X^\ell_n \cup X^k_n$, $L = \ell + k$, $n = 1, \ldots, N$. По каждой обучающей выборке строится АВО. Для каждого шара, полученного в процессе поиска, вычисляются следующие характеристики, зависящие только от обучающей выборки: количество покрытых объектов; количество ошибок; мощность опорного множества; вес шара $W(B_i)$, и др. Исследуется зависимость средней переобученности шаров от этих характеристик с целью понять причины переобучения и скорректировать управляющие параметры алгоритма поиска шаров. Практически во всех экспериментах наблюдались следующие закономерности. Шары, выделяющие меньше объектов, более переобучены. С увеличением числа перебираемых шаров переобученность сначала уменьшается, затем возрастает, однако оптимальная «глубина перебора» в каждой задаче своя. Мощность опорного множества практически не влияет на переобученность.

Результаты экспериментов

В экспериментах на реальных задачах из репозитория UCI качество предложенного алгоритма оказалось сопоставимым с лучшими из известных логических алгоритмов классификации [3]. В задаче распознавания участков генных последовательностей (promoters) точность классификации у АВО примерно втрое лучше, чем у конкурентов, см. Таблицу 1.

Задача	C4.5 Trees	C4.5 Rules	C5.0 Rules	RIP-PER	SLIP-PER	ABO
german	27.5	27.0	28.3	28.6	27.2	**25.5**
australian	18.8	18.8	20.1	**15.2**	15.7	15.8
ionosphere	10.3	10.3	12.3	10.3	7.7	**6.7**
liver	37.5	37.5	31.9	**31.3**	32.3	32.3
promoters	22.7	18.1	22.7	18.1	18.9	**5.5**
breast-cancer	6.6	5.2	5.0	3.7	4.2	**3.6**
hepatitis	20.8	20.0	21.8	19.7	17.4	**16.6**

Таблица 1. Процент ошибочных классификаций при 10-кратном скользящем контроле на 7 реальных задачах из репозитория UCI.

Работа выполнена при поддержке РФФИ, проект №05-01-00877, и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

[1] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. кибернетики. — 1978. — Т. 33. — С. 5–68.

[2] *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных. — Новосибирск: Наука, 1981.

[3] *Cohen W. W., Singer Y.* A simple, fast and effective rule learner // Proc. of the 16 National Conference on Artificial Intelligence. — 1999. — Pp. 335–342.

[4] *Mason L., Bartlett P., Baxter J.* Direct optimization of margins improves generalization in combined classifiers: Tech. rep.: Australian National Univ., 1998.

Методы коррекции локально возмущенных полуметрик

Громов И. А.

igor_gromov@mail.ru

Москва, МГУ им. М. В. Ломоносова

Для решения задач интеллектуального анализа данных широко применяются метрические методы. Эффективность их использования существенно зависит от выбора функции сходства (например, полуметрики) на обрабатываемых объектах. Нередко выбор полуметрики для решения конкретной задачи субъективен, а значит, можно модифицировать полуметрику с тем, чтобы она точнее отражала характерные сходства и различия исследуемых объектов.

Постановка задачи. Рассматривается задача коррекции полуметрики в ситуации, когда эксперт (в предметной области) требует изменить расстояние между ровно одной парой объектов. Пусть дано конечное множество объектов мощности N с заданной на нем полуметрикой R.

Эксперт выбирает одну пару объектов (i, j) и по своему усмотрению изменяет расстояние между ними: $r_{ij} \mapsto r'_{ij}$. В общем случае это влечет нарушение неравенств треугольника. Требуется предложить методы коррекции $A\colon R' \mapsto \widetilde{R}$, которые позволят построить полуметрику $\widetilde{R}$ такую, что $r'_{ij} = \tilde{r}_{ij}$, и при этом будут

1) универсальны, т. е. применимы для любых R и R';
2) «наиболее полно» соответствовать экспертной интерпретации внесенного возмущения;
3) строить полуметрику $\widetilde{R}$, «максимально схожую» с исходной полуметрикой R.

Формализация задачи. Для того, чтобы коррекция отвечала требованию 2, предложена процедура обучения алгоритма коррекции. Она позволяет взаимодействовать с экспертом на более естественном для него языке и полнее учитывать его требования к проводимой коррекции.

Для формализации понятия сходства метрик (а при некоторых допущениях также и полуметрик) введен ряд интерпретируемых функционалов их сравнения:

$$Q_w(R, \widetilde{R}) = w_u Q_u(R, \widetilde{R}) + w_p Q_p(R, \widetilde{R}), \qquad w_u, w_p \geqslant 0; \tag{1}$$

$$Q_u(R, \widetilde{R}) = \frac{\sum\limits_{(kl)\in E_N} w_{kl} \left(\tilde{r}_{kl} - r_{kl}\right)^2}{\sum\limits_{(kl)\in E_N} \tilde{r}_{kl}^2}, \qquad w_{kl} \geqslant 0; \tag{2}$$

$$\begin{aligned} Q_p(R, \widetilde{R}) = \sum_{\triangle(klm)} & w_{kl,km} \left(\frac{\tilde{r}_{kl}}{\tilde{r}_{km}} - \frac{r_{kl}}{r_{km}}\right)^2 + \\ & + w_{kl,lm} \left(\frac{\tilde{r}_{kl}}{\tilde{r}_{lm}} - \frac{r_{kl}}{r_{lm}}\right)^2 + w_{km,lm} \left(\frac{\tilde{r}_{km}}{\tilde{r}_{lm}} - \frac{r_{km}}{r_{lm}}\right)^2, \\ & w_{kl,km},\ w_{kl,lm},\ w_{km,lm} \geqslant 0. \end{aligned} \tag{3}$$

В ряде случаев (для некоторых значений весов в (2), (3)) получены методы коррекции, доставляющие минимум указанным функционалам. Таким образом, удалось напрямую согласовать функционалы сравнения метрик с формулами коррекции.

Трехэтапная схема построения алгоритмов коррекции полуметрики. Существуют различные подходы к решению поставленной задачи коррекции полуметрики. Автором предложена трехэтапная схема коррекции возмущенных полуметрик. В рамках данной схемы в ходе коррекции рассматриваются тройки попарно различных объектов и треугольники, в которых вершинами являются такие объекты, а длинами

сторон — расстояния между ними. На первом этапе коррекции рассматриваются только тройки объектов вида (ijk), на втором — (ikl) и (jkl), на третьем — (klm), $k,l,m \notin \{i,j\}$. На каждом из этапов коррекции модифицируются расстояния в тех и только тех треугольниках, в которых неравенства треугольника нарушены. В общем случае данная схема коррекции требует исследования всех троек объектов и имеет сложность $O(N^3)$. Для достаточно больших значений N вычислительная сложность данной схемы становится препятствием к ее практическому применению.

В результате исследования конечных полуметрик были выявлены новые свойства. На их основании предложен ряд алгоритмов, не требующих рассмотрения в процессе коррекции всех троек объектов. Было доказано, что при их использовании достаточно проведения только первых двух этапов коррекции для построения полуметрики. Таким образом, данные алгоритмы имеют квадратичную $O(N^2)$, а в специальном случае — линейную сложность, т. е. полуметрика строится в ходе первого этапа. Вычислительная эффективность предложенных алгоритмов позволяет проводить преобразование полуметрики интерактивно.

Универсальный алгоритм коррекции полуметрики $\mathcal{A}$. В данном разделе сформулирован *универсальный* алгоритм коррекции полуметрики $\mathcal{A}$, т. е. алгоритм, гарантировано строящий полуметрику для любых исходных полуметрик и любых локальных возмущений, внесенных экспертом.

Пусть эксперт модифицировал в полуметрике R одно расстояние: $r_{ij} \mapsto r'_{ij}$ и требует сохранить указанное им значение r'_{ij}. Кроме того, эксперт зафиксировал функционал сравнения полуметрик (тем самым давая интерпретацию внесенного возмущения). Тогда для того, чтобы скорректировать возникшие вследствие этого нарушения неравенств треугольника в R', предлагается следующий алгоритм $\mathcal{A}$.

1-й этап: коррекция $\triangle(ijk)$, в которых неравенства треугольника нарушены. Какой именно метод коррекции при этом должен быть применен, определяется выбором функционала сравнения полуметрик.

2-й этап: коррекция $\triangle(ikl)$ и $\triangle(jkl)$, в которых неравенства треугольника нарушены. Коррекция проводится по следующему правилу:

$$\tilde{r}_{kl} = \alpha \tilde{r}_{kl}^{\min} + (1-\alpha)\tilde{r}_{kl}^{\max},\ \forall (k,l) : k,l \notin \{i,j\},$$

где $\tilde{r}_{kl}^{\min} = \max\{|\tilde{r}_{ik} - \tilde{r}_{il}|, |\tilde{r}_{jk} - \tilde{r}_{jl}|\}$, $\tilde{r}_{kl}^{\max} = \min\{(\tilde{r}_{ik} + \tilde{r}_{il}), (\tilde{r}_{jk} + \tilde{r}_{jl})\}$, $\alpha \in [0,1]$ и α фиксировано *для всех* $\triangle(ikl)$, $\triangle(jkl)$.

3-й этап: не требуется.

На первом этапе выполнения алгоритма $\mathcal{A}$ требуется рассмотреть $N-2$ треугольников, на втором — $(N-2)(N-3)$ треугольников. Если

сложность вычисления первого этапа — $O(N)$, то сложность алгоритма $\mathcal{A}$ — $O(N^2)$.

Параметр α может быть либо задан экспертом, либо настроен для получения величины $\tilde{r}_{kl}$, наиболее близкой к r_{kl}.

В универсальном алгоритме $\mathcal{A}$ не накладывается никаких специальных ограничений на методы коррекции, используемые на первом этапе, что, однако, компенсируется весьма жесткими условиями, налагаемыми на процедуру коррекции на втором. В ходе исследования алгоритмов коррекции в рамках трехэтапной схемы удалось перераспределить мощность требований между этапами выполнения алгоритма в духе алгебраического подхода к синтезу алгоритмов распознавания. За счет сужения множества методов коррекции на первом этапе были существенно смягчены условия второго этапа.

Литература

[1] *Майсурадзе А. И.* Гомогенные и ранговые базисы в пространствах метрических конфигураций // ЖВМиМФ. — 2006. — Т. 46, № 2. — С. 344–361.

[2] *Deza M., Dutour M.* Data mining for cones of metrics, quasi-metrics, semimetrics, and super-metrics. — 2006.

Нелинейные монотонные композиции классификаторов

Гуз И. С.
ivanguz@mail.ru
Москва, МФТИ

Объединение нескольких алгоритмов в композицию во многих случаях позволяет повысить качество классификации. Линейные и выпуклые композиции хорошо исследованы как теоретически, так и эмпирически [1, 2]. Нелинейные монотонные композиции были предложены в [3] и относительно мало исследованы. Цель данной работы — сравнить эффективность монотонных и линейных композиций, а также исследовать влияние структуры композиции и параметров метода настройки на обобщающую способность монотонной композиции.

Стандартная постановка задачи классификации на два класса

Пусть X — множество допустимых описаний объектов, $Y = \{-1, 1\}$ — множество меток классов, $y^*\colon X \to Y$ — неизвестная *целевая зависимость*, $X^\ell = (x_i)_{i=1}^\ell \subset X \times Y$ — обучающая выборка, $y_i = y^*(x_i)$. Требуется построить алгоритм классификации $a\colon X \to Y$, аппроксимирующий целевую зависимость $y^*(x)$ на всём множестве X.

Пусть фиксирован *метод обучения* μ, который по выборке X^ℓ с заданными весами объектов $W^\ell = (w_i)_{i=1}^\ell$ строит *алгоритмический оператор*

$b\colon X \to \mathbb{R}$ путём минимизации функционала средней взвешенной ошибки $Q(b) = \sum_{i=1}^{\ell} w_i \big[a(x_i) \neq y_i\big]$, где $a(x) = \operatorname{sign} b(x)$ — алгоритм классификации, соответствующий оператору b.

Монотонные композиции классификаторов

Композицией алгоритмических операторов $b_1, \ldots, b_T$ называется алгоритм вида $a(x) = F\big(b_1(x), \ldots, b_T(x)\big)$, где отображение $F\colon \mathbb{R}^T \to Y$ называется *корректирующей операцией* [1].

Требование монотонности F как отображения из $\mathbb{R}^T$ в Y означает, что $F(b_1, \ldots, b_T)$ не должно уменьшаться при увеличении выходных значений операторов $b_1, \ldots, b_T$, Это вполне естественное требование, если предполагать, что операторы настраиваются на решение одной и той же задачи. Монотонные корректирующие операции образуют более широкое семейство функций по сравнению с выпуклыми (линейными с неотрицательными коэффициентами). Это позволяет точнее настраиваться на данные, но, возможно, повышает риск переобучения.

Пара объектов $x_j, x_k \in X^\ell$ называется *дефектной парой* набора алгоритмических операторов $b_1, \ldots, b_T$, если $y_j < y_k$ и $b_t(x_j) \geqslant b_t(x_k)$ для всех $t = 1, \ldots, T$. Число дефектных пар можно рассматривать как функционал качества композиции [4]. Операторы $b_1, \ldots, b_T$ строятся и добавляются в композицию последовательно. Перед настройкой очередного оператора b_t методом μ вес w_i каждого объекта x_i, $i = 1, \ldots, \ell$, устанавливается пропорционально числу дефектных пар набора операторов $b_1, \ldots, b_{t-1}$, в которых участвует данный объект. Кроме того, вводится параметр λ, управляющий стратегией настройки очередного оператора: при $\lambda = 0$ оператор b_t настраивается только на аппроксимацию обучающей выборки без учёта ранее построенных операторов $b_1, \ldots, b_{t-1}$; при $\lambda = 1$ оператор b_t настраивается только на компенсацию ошибок, допущенных композицией $F(b_1, \ldots, b_{t-1})$; при промежуточных значениях $\lambda \in (0, 1)$ реализуется компромиссная стратегия настройки.

Наряду с дефектными парами при расчёте весов w_i можно учитывать специальным образом определяемые дефектный тройки [4], что значительно увеличивает вычислительную сложность алгоритма. Поэтому в данной работе исследуется вопрос о целесообразности учета троек.

Эксперименты и результаты

Исследование монотонной коррекции проводились на четырёх реальных задачах из репозитория UCI. В качестве метода обучения μ использовались два метода: байесовский классификатор с локальным восстановление плотности по Парзену-Розенблатту и переменной шириной окна; и метод опорных векторов (SVM). Выпуклая композиция, с которой проводилось сравнение, строилась алгоритмом AdaBoost [5].

Метод \ Задача	ionospere	house-vote	bupa	diabetes
Monotone (SVM)	9.7% (3)	**3.2**% (5)	**31.3**% (2)	23.6% (2)
Monotone (Parzen)	**8.0**% (2)	5.6% (5)	32.7% (3)	30.2% (2)
AdaBoost (SVM)	11.5% (65)	4.1% (40)	**30.7**% (15)	**22.7**% (15)
AdaBoost (Parzen)	12.0% (15)	6.0% (11)	33.0% (34)	29.0% (23)
SVM	13.1%	4.5%	42.2%	23.0%
Parzen	15.0%	6.2%	33.8%	30.7%

Таблица 1. Доля ошибок на контрольных данных, усредненная по 50 разбиениям. В скобках указано среднее число алгоритмов в композиции.

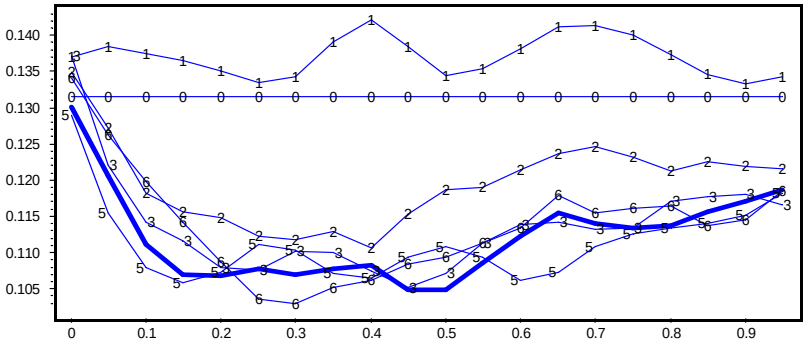


Рис. 1. Зависимость доли ошибок на контроле от параметра λ на примере монотонной коррекции SVM в задаче `ionosphere`. Цифрами 1–6 отмечено значение T. Выделена кривая $T = 4$, для которой положение минимума наиболее устойчиво относительно λ. Линия, помеченная цифрой 0, показывает уровень ошибок на контроле для отдельного SVM.

Для оценивания обобщающей способности использовался скользящий контроль: выборка разбивалась 50 раз случайным образом на обучающую (80%) и контрольную (20%). Оптимальное число T операторов в композиции определялось по минимуму средней доли ошибок на контрольной выборке. Результаты представлены в Таблице 1.

Оптимальное число операторов в монотонных композициях в большинстве случаев равно 2 или 3, что гораздо меньше, чем в линейных композициях. При дальнейшем увеличении T число дефектных пар быстро исчерпывается до нуля и возникает переобучение.

Параметр λ имеет смысл брать в пределах 0.1–0.4, Рис 1. Настройка базовых алгоритмов только лишь на устранение совокупного дефекта предыдущих алгоритмов ($\lambda = 1$) ведёт к переобучению.

При $T = 2$ учет дефектных троек немного улучшает обобщающую способность; при $T \geqslant 3$ только ухудшает. Поскольку подсчёт дефектных троек — трудоемкая операция, то от него вообще можно отказаться.

Таким образом, монотонные композиции, так же как и линейные, позволяют улучшать качество классификации, но при этом состоят из гораздо меньшего числа алгоритмов.

Работа выполнена при поддержке РФФИ, проект №05-01-00877, и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

[1] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. кибернетики. — 1978. — Т. 33. — С. 5–68.

[2] *Kuncheva L.* Combining pattern classifiers. — John Wiley & Sons, Inc., 2004.

[3] *Рудаков К. В., Воронцов К. В.* О методах оптимизации и монотонной коррекции в алгебраическом подходе к проблеме распознавания // Докл. РАН. — 1999. — Т. 367, № 3. — С. 314–317.

[4] *Воронцов К. В.* Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания // ЖВМ и МФ. — 2000. — Т. 40, № 1. — С. 166–176.

[5] *Freund Y., Schapire R. E.* Experiments with a new boosting algorithm // International Conference on Machine Learning. — 1996. — Pp. 148–156.

Кластеризация элементов множества на основе взаимных расстояний и близостей

Двоенко С. Д.

dsd@uic.tula.ru

Тула, Тульский государственный университет

Принцип несмещеной кластеризации лежит в основе известных алгоритмов кластер-анализа. Рассмотрены их модификации, когда доступна только матрица расстояний или близостей между объектами. Показана связь с алгоритмами экстремальной группировки признаков.

Кластеризация объектов

В кластер-анализе предполагается, что объекты $\omega_i \in \Omega$, $i = 1, \ldots, N$, расположены в n-мерном пространстве (обычно евклидовом), где каждый представлен вектором $\mathbf{x}_i = (x_{i1}, \ldots x_{in})$, а все — матрицей данных $X(N, n)$. Каждый признак представлен своими наблюдениями $X_j = (x_{1j}, \ldots x_{Nj})^{\mathrm{T}}$, $j = 1, \ldots n$. Объекты образуют K локальных сгущений (кластеры, классы, таксоны), которые следует выделить. Алгоритмы

кластер-анализа (например, K-средних [1], семейство FOREL [2]) предполагают наличие признаков и строят «несмещенную» [3] кластеризацию: в ней все «представители» кластеров совпадают с их «центрами» $\widetilde{\mathbf{x}}_k = \overline{\mathbf{x}}_k$. Иначе центры назначаются представителями, и кластеры переопределяются. Центр кластера $\overline{\mathbf{x}}_k$ может не совпадать с его элементами $\mathbf{x}_i \in \Omega_k$.

В матрице расстояний $D(N,N)$ объект $\omega(\overline{\mathbf{x}}_k)$ не представлен. Обычно «центром» кластера выбирают объект $\overline{\omega}_k$, наименее удаленный от объектов кластера Ω_k. При $\widetilde{\omega}_k = \overline{\omega}_k$ кластеризация может оказаться смещенной в пространстве, т. к. «центр» $\mathbf{x}(\overline{\omega}_k)$ не совпадет со средним $\overline{\mathbf{x}}_k$.

Относительно любого объекта $\omega_k \in \Omega$, взятого как начало координат, и любой пары объектов ω_i, ω_j; $i,j = 1,\ldots,N$, по известной теореме косинусов определяется их скалярное произведение $c_{ij} = (d^2_{ki} + d^2_{kj} - d^2_{ij})/2$, где $c_{ii} = d^2_{ki}$ и $d_{pq} = d(\omega_p, \omega_q)$ — расстояние. В матрицах $C_k(N,N)$, $k = 1,\ldots,N$, их главные диагонали представляют квадраты расстояний от соответствующего начала координат ω_k до объектов ω_i, $i = 1,\ldots,N$. Янгом и Хаусхолдером в задаче метрического шкалирования [4] предложено восстанавливать евклидово пространство как разложение $C_k = XX^{\mathrm{T}}$, где матрица $C_k(N-1,N-1)$ с рангом $n < N$ положительно полуопределена, $X(N-1,n)$ — матрица проекций объектов $\omega_1,\ldots\omega_{k-1},\omega_{k+1},\ldots\omega_N$ на n ортогональных осей с началом координат ω_k. В методе главных проекций Торгенсона [5] восстанавливается евклидово пространство с началом координат в центре тяжести множества объектов $\omega_i \in \Omega$, $i = 1,\ldots,N$.

Центры $\overline{\omega}_k$ кластеров Ω_k, $k = 1,\ldots,K$ немедленно определяются по методу Торгенсона. Для начала координат в центре тяжести кластера Ω_k получим матрицу скалярных произведений $\overline{C}_k(N,N)$. Диагональные элементы $\overline{c}^k_{ii} = d^2(\omega_i, \overline{\omega}_k)$, $i = 1,\ldots,N$ представляют центр $\overline{\omega}_k$ кластера Ω_k квадратами расстояний до остальных объектов $\omega_i \in \Omega$, $i = 1,\ldots,N$:

$$d^2(\omega_i, \overline{\omega}_k) = \frac{1}{N_k}\sum_{p=1}^{N_k} d^2_{ip} - \frac{1}{2N_k^2}\sum_{p=1}^{N_k}\sum_{q=1}^{N_k} d^2_{pq}; \quad \omega_p, \omega_q \in \Omega_k,$$

где N_k — число объектов в Ω_k. Отсюда сразу определяются алгоритмы кластеризации для расстояний, например K-средних и FOREL.

Положительно полуопределенную матрицу $S(N,N)$ попарных близостей $s_{ij} = s(\omega_i,\omega_j) \geqslant 0$ объектов $\omega_i, \omega_j \in \Omega$ можно считать матрицей скалярных произведений векторов $\mathbf{x}_i = \mathbf{x}(\omega_i)$, $i = 1,\ldots,N$, в пространстве размерности не выше N. Скалярные произведения объектов ω_i, ω_j относительно $\omega_k \in \Omega$ представлены как $s_{ij} = (d^2_{ki} + d^2_{kj} - d^2_{ij})/2$. Поскольку $s_{ii} = d^2_{ki}$, то $d^2_{ij} = s_{ii} + s_{jj} - 2s_{ij}$, и можно получить кластеризацию по расстояниям.

Пусть объекты разбиты на K кластеров Ω_k. Представим центр кластера $\overline{\omega}_k$ своими близостями к остальным объектам $\omega_i \in \Omega$:

$$s(\omega_i, \overline{\omega}_k) = \frac{1}{N_k} \sum_{p=1}^{N_k} s_{ip}; \quad \omega_p \in \Omega_k,$$

где N_k — число объектов в Ω_k. Отсюда сразу определяются алгоритмы кластеризации для близостей, например K-средних и FOREL.

Несмещенная кластеризация минимизирует дисперсию кластера и максимизирует среднюю близость объектов в кластере:

$$\sigma_k^2 = \frac{1}{2N_k^2} \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} d_{ij}^2 = \frac{1}{N_k} \sum_{i=1}^{N_k} s_{ii} - \frac{1}{2N_k^2} \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} s_{ij}; \quad \omega_i, \omega_j \in \Omega_k.$$

После нормировки $s'_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}$ получим $\sigma_k^2 = 1 - \delta_k$.

Кластеризация признаков

Группировка n признаков по их корреляциям $R(n,n)$ выполняется алгоритмом K-средних для близостей $S(n,n)$, где $s_{ij} = r_{ij}^2$ или $s_{ij} = |r_{ij}|$, максимизируя величины (n_k — число признаков ω_i в группе Ω_k):

$$I_1 = \sum_{k=1}^{K} n_k \delta'_k = \sum_{k=1}^{K} \sum_{i=1}^{n_k} r^2(\omega_i, \overline{\omega}_k) \text{ и } I_2 = \sum_{k=1}^{K} n_k \delta''_k = \sum_{k=1}^{K} \sum_{i=1}^{n_k} |r(\omega_i, \overline{\omega}_k)|.$$

В алгоритмах экстремальной группировки «квадрат» и «модуль» [6] функционалы $J_1 = \sum_{k=1}^{K} \sum_{i=1}^{n_k} r^2(\omega_i, \pi_k)$ и $J_2 = \sum_{k=1}^{K} \sum_{i=1}^{n_k} |r(\omega_i, \mu_k)|$, $\omega_i \in \Omega_k$, где π_k — главный «фактор», μ_k — центроидный «фактор» группы Ω_k, характеризуют качество разбиения признаков на K групп, в каждой из которой признаки наиболее сильно коррелируют со своим фактором. Факторы строятся как одновременное решение основных факторных задач: построение K общих факторов и их косоугольное вращение [7].

Представим центр $\overline{\omega}_k$, главный π_k и центроидный μ_k факторы группы Ω_k своими близостями к признакам $\omega_i \in \Omega_k$:

$$\begin{cases} s(\omega_i, \overline{\omega}_k) = (1/n_k) \sum\limits_{j=1}^{n_k} s_{ij}; \\ s(\omega_i, \pi_k) = \sum\limits_{j=1}^{n_k} \alpha_j^k s_{ij} = \lambda_k \alpha_j^k, \quad \boldsymbol{\alpha}_k = (\alpha_1^k, \ldots \alpha_{n_k}^k); \\ s(\omega_i, \mu_k) = \sum\limits_{j=1}^{n_k} s_{ij}; \end{cases}$$

где λ_k — максимальное собственное значение, $\boldsymbol{\alpha}_k$ — соответствующий ему собственный вектор подматрицы близостей $S(n_k, n_k)$ признаков $\omega_i \in \Omega_k$.

Легко увидеть, что близости $s(\omega_i, \overline{\omega}_k)$ и $s(\omega_i, \mu_k)$ совпадают с точностью до множителя. Поэтому группировки по «модулю» — несмещенные. Группировки по «квадрату» — смещенные. Также очевидно, что $J_1 \geqslant I_1$. В итоге, оба алгоритма K-средних для расстояний и для близостей, примененные для кластеризации признаков, аналогичны алгоритму «модуль».

В основе алгоритма FOREL лежит процедура несмещенной кластеризации, которую можно представить как алгоритм «1-среднего». Следовательно, техника кластер-анализа, развитая для семейства FOREL, адекватно применима для группировки признаков.

Работа выполнена при поддержка РФФИ, проект № 05-01-00679 и INTAS, проект № 04-77-7347.

Литература

[1] *Tou J. T., Gonzalez R. C.* Pattern recognition principles. — London: Addison-Wesley, 1981.

[2] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Изд. Ин-та матем., 1999.

[3] *Шлезингер М.И.* О самопроизвольном различении образов // Читающие автоматы, Киев: Наукова думка, 1965. — С. 38–45.

[4] *Young G., Householder A. S.* Discussion of a set of points in terms of their mutual distances // Psychometrika. — 1938. — V. 3. — P. 19–22.

[5] *Torgenson W. S.* Theory and methods of scaling. — N.Y.: J. Wiley, 1958.

[6] *Браверман Э. М.* Методы экстремальной группировки параметров и задача выделения существенных факторов // Автоматика и телемеханика. — 1970. — № 1. — С. 123–132.

[7] *Harman H. H.* Modern factor analysis. — Chicago: Univ. Chicago Press, 1976.

Алгоритм распознавания, основанный на построении метрических закономерностей

Дедовец М. С., Сенько О. В.

senkoov@ccas.ru

Москва, ВЦ РАН, ВМиК МГУ

Ранее были предложены связанные с решением задач распознавания средства анализа данных, основанные на поиске закономерностей. При этом под закономерностью понимается подобласть признакового пространства, содержащая объекты одного класса (полная закономерность) или преимущественно одного класса (частичная закономерность). При этом геометрическая форма подобласти определяется заранее заданной моделью. Так, наиболее часто используемыми моделями явля-

ются модели логических закономерностей [1, 2, 3], в которых искомые области признакового пространства имеют форму гиперпараллелепипедов. Рассматривались также закономерности, формируемые с помощью линейных границ с произвольной ориентацией относительно координатных осей [4]. Несомненно, что априорное задание геометрической формы является существенным ограничением, затрудняющим выявление закономерностей, реально существующих в данных, но не удовлетворяющих сделанным предположениям. В связи с этим был предложен новый тип закономерностей, который далее будет называться метрическим. Под метрической закономерностью понимается подобласть признакового пространства, задаваемая как окрестность некоторого набора точек обучающей выборки, принадлежащих к одному классу K с минимальным включением объектов, не принадлежащих K.

При задании закономерностей используется понятие G-смежности между объектами одного из классов. Предположим, что у нас задана некоторая метрика ρ. Два объекта S_1 и S_2 из пересечения класса K и обучающей выборки назовём G-смежными, если в обучающей выборке не существует такого объекта S, не принадлежащего классу K, что одновременно выполняются два неравенства: $\rho(S, S_1) \leqslant \rho(S_1, S_2)$ и $\rho(S, S_2) \leqslant \rho(S_1, S_2)$. Смысл отношения смежности между объектами одного класса состоит в том, что оно обеспечивает отсутствие между ними объектов других классов. Классу K может быть сопоставлен граф, вершинами которого являются объекты класса. Двум вершинам ставится в соответствие ребро, если соответствующие объекты являются G-смежными. Под метрической закономерностью в настоящем исследовании нами понималась окрестность компоненты связности графа G-смежности. Объект S принадлежит метрической закономерности, если существует хотя бы одна пара объектов S_1 и S_2 таких, что: $\rho(S, S_1) \leqslant \rho(S_1, S_2)$ и $\rho(S, S_2) \leqslant \rho(S_1, S_2)$.

Был разработан алгоритм распознавания, основанный на голосовании по представительным системам метрических закономерностей. При этом закономерности строились в пространствах, задаваемых парами или тройками признаков. В систему включались только те закономерности, которые содержали не менее $0.25\, m_K$ объектов, где m_K — число объектов класса K в обучающей выборке. Для вычисления оценки объекта S за класс K использовалась формула $\gamma(S, K) = \sum_{S_i \in K} V(i)/V_0(i)$, где суммирование ведётся по всевозможным объектам обучающей выборки из класса K, $V_0(i)$ — число метрических закономерностей, в которые вошёл объект S_i, $V(i)$ — число метрических закономерностей, в которые одновременно вошли объекты S и S_i.

Оценка эффективности алгоритма проводилась на совокупности реальных прикладных задач. Проведённые исследования показали достаточно высокую точность распознавания на контрольной информации, сравнимую с точностью, даваемой альтернативными подходами — методом статистически взвешенных синдромов и методом опорных векторов. Можно сделать вывод о перспективности модели и необходимости дальнейших исследований в данном направлении.

Работа выполнена при поддержке РФФИ, проекты № 06-01-00492, № 06-01-08045.

Литература

[1] *Журавлёв Ю. И., Рязанов В. В., Сенько О. В.* Распознавание. Математические методы. Программная система. Применения. — Москва: Фазис, 2006.

[2] *Богомолов В. П., Виноградов А. П., Журавлёв Ю. И., Катериночкина Н. Н., Ларин С. Б.,Рязанов В. В.,Сенько О. В.* Программная система распознавания ЛОРЕГ: алгоритмы распознавания, основанные на голосовании по системам логических закономерностей. — М.: ВЦ РАН, 1998. — 63 с.

[3] *Кузнецов В. А., Сенько О. В., Кузнецова А. В. и др.* Распознавание нечетких систем по методу статистически взвешенных синдромов и его применение для иммуногематологической нормы и хронической патологии // Химическая физика, 1996. — Т. 15, № 1. — С. 81–100.

[4] *Dokukin A. A., Senko O. V.* About new pattern recognition method for the universal program system Recognition. Proc. of the Int. Conf, I.Tech-2004, Varna (Bulgaria), 14–24 June 2004. — Pp. 54–58.

[5] *Сенько О. В., Кузнецова А. В.* Алгоритмы распознавания, основанные на голосовании по системам закономерностей различных типов // Докл. всеросс. конф. ММРО-12. — Москва, 2005, — С. 200–203.

Непараметрический иерархический классификатор для случая многих классов

Добротворский Д. И., Пестунов И. А., Синявский Ю. Н.

pestunov@ict.nsc.ru

Новосибирск, Институт вычислительных технологий СО РАН

В настоящее время проблема выбора информативных признаков в рамках параметрического подхода хорошо изучена, предложен ряд эффективных подходов к ее решению. Однако практически отсутствуют методы выбора признаков для непараметрических классификаторов [2]. В докладе представлен иерархический непараметрический классификатор на основе оценок Розенблатта-Парзена и связанный с ним метод выделения информативных признаков.

Традиционный подход к построению непараметрических правил классификации, основанных на оценках Розенблатта-Парзена, заключается в подстановке в байесовское решающее правило вместо неизвестных вероятностных характеристик классов соответствующих им оценок, полученных по обучающим выборкам [3]. Общий вид этих правил для $(0,1)$-матрицы потерь можно представить выражением

$$\hat{\delta}_0 = \hat{\delta}_0(x;V) = \underset{i\in\{1,\ldots,M\}}{\arg\max}\ q_i\hat{f}_i(x).$$

Здесь $x \in \mathbb{R}^k$; $V = \bigcup_{i=1}^{M} V^{(i)}$ — обучающая выборка объема $N = \sum_{i=1}^{M} N_i$, $V^{(i)} = \left\{x_j^{(i)} \in \mathbb{R}^k\right.$ — наблюдение из i-го класса$\left.\right\}$; q_i, $i = 1,\ldots,M$ — априорная вероятность i-го класса; $\hat{f}_i(x)$ — оценка условной плотности распределения i-го класса $f_i(x)$ в точке $x \in \mathbb{R}^k$, определяемая выражением

$$\hat{f}_i(x) = \frac{1}{N_i c^k}\sum_{j=1}^{N_i}\Phi\left(\frac{x - x_j^{(i)}}{c}\right),$$

где Φ — ядро, c — параметр сглаживания. Для случая двух классов Ω_1 и Ω_2 это правило можно переписать следующим образом:

$$\begin{cases} x \in \Omega_1, & \text{если } \hat{h}(x) = -\ln\frac{\hat{f}_1(x)}{\hat{f}_2(x)} < t, \\ x \in \Omega_2, & \text{в противном случае,} \end{cases}$$

где $\hat{h}(x)$ — непарамерическая оценка функции $h(x) = -\ln(f_1(x)/f_2(x))$, а $t = \ln(q_1/q_2)$ — решающий порог.

В соответствии с методом [2], выделение информативных признаков сводится к нахождению матрицы признаков решающей границы Σ_{DB} и вычислению ее собственных векторов $(v_1,\ldots,v_k)$, задающих ортонормированный базис пространства признаков. Бóльшим собственным значениям соответствуют более информативные признаки.

Пусть $n(x)$ — единичный вектор нормали к решающей границе S в точке x. Тогда матрица Σ_{DB} определяется следующим образом:

$$\Sigma_{DB} = \int_S n(x)n^{\mathrm{T}}(x)f(x)dx \Big/ \int_S f(x)dx,$$

где $f(x)$ — плотность распределения вектора признаков в точке x.

Нахождение поверхности S и нормалей к ней осуществляется следующим образом. Пусть точки $x^{(1)}$ и $x^{(2)}$ правильно классифицированы и относятся к разным классам. Тогда отрезок, соединяющий эти точки,

должен пересекать решающую границу. Поэтому, двигаясь вдоль этого отрезка, можно найти точку $x \in S$. В предлагаемом алгоритме поиск осуществляется методом деления отрезка пополам.

Уравнение байесовской решающей границы можно записать в виде $h(x) = t$. Поскольку функция $h(x)$ в непараметрическом случае неизвестна, вектор нормали к S в точке x приближенно выражается следующим образом:

$$\bigtriangledown h(x) = \frac{\partial h}{\partial x_1}x_1 + \frac{\partial h}{\partial x_2}x_2 + \cdots + \frac{\partial h}{\partial x_n}x_n \approx \frac{\Delta \hat{h}}{\Delta x_1}x_1 + \frac{\Delta \hat{h}}{\Delta x_2}x_2 + \cdots + \frac{\Delta \hat{h}}{\Delta x_n}x_n.$$

Алгоритм реализации описанного метода, предложенный в работе [2], является вычислительно сложным, что существенно ограничивает его применимость к выборкам большого объема. В работе [1] представлен быстрый алгоритм оценивания матрицы Σ_{DB}, быстродействие которого достигается за счет уменьшения объема выборки, участвующей в построении классификатора.

Выделение признаков в многоклассовом случае является достаточно сложной задачей. Для поиска информативных признаков в случае M классов ($M > 2$) в работе [2] предлагается традиционный способ сведения ее к решению нескольких двухклассовых задач. В этом случае матрица Σ_{DB} определяется по формуле

$$\Sigma_{DB} = \sum_{(\Omega_i, \Omega_j)} q_i q_j \Sigma_{DB}(\Omega_i, \Omega_j). \qquad (1)$$

Такой подход часто приводит к необоснованным вычислительным затратам и снижению качества выбираемых признаков (Рис. 1).

Представленный в докладе непараметрический иерархический классификатор сначала выделяет изолированные группы близких классов (на Рис. 1 обведены пунктирной линией), затем, при необходимости, разделяет их на более мелкие группы. На каждом уровне иерархии для классификации используется свой минимально достаточный набор информативных признаков, определяемый на основе матрицы (1). Это позволяет снизить трудоемкость алгоритма классификации без потери качества.

Статистическое моделирование на многочисленных модельных и реальных данных показывает, что предлагаемый метод построения непараметрического иерархического классификатора позволяет более чем на порядок сократить объем требуемых вычислений.

Работа выполнена в рамках интеграционного проекта СО РАН и ДВО РАН №86.

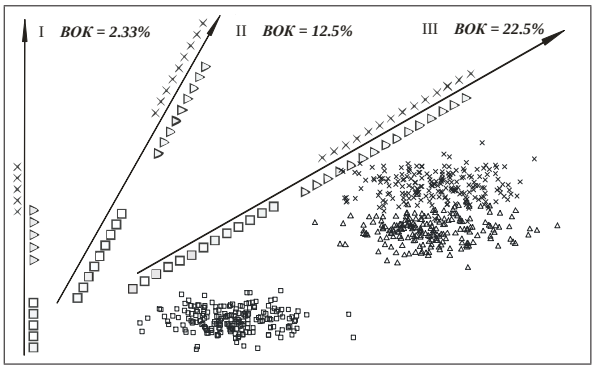


Рис. 1. Двумерная модель, состоящая из трех нормально распределенных классов. Стрелками показан оптимальный информативный признак (III) и признаки, выделенные по методу главных компонент (I) и по формуле (1) (II).

Литература

[1] *Добротворский Д. И., Пестунов И. А.* Быстрый алгоритм извлечения признаков для непараметрического классификатора на основе решающей границы // Межд. конф. ВИТ-2006, Павлодар: ТОО НПФ «ЭКО», 2006. — Т. I. — С. 409–417.

[2] *Lee C., Landgrebe D. A.* Decision Boundary Feature Extraction for Non-Parametric Classification // IEEE Trans. on System, Man and Cybernetics. — 1993. — Vol. 23, N,2. — С. 433–444.

[3] *Харин Ю. С.* Робастность в статистическом распознавании образов. — Минск: Университетское, 1992. — 232 с.

Индуктивный поиск оптимального алгоритма вычисления оценок

Докукин А. А.
dalex@ccas.ru
Москва, ВЦ РАН

В статье описывается индуктивная модификация разработанного ранее метода построения оптимального алгоритма вычисления оценок (АВО) при динамическом пополнении обучающей выборки.

Алгоритм построения АВО

В этой главе будет описана задача построения АВО и кратко изложены предыдущие результаты.

Алгоритм вычисления оценок для задачи распознавания [2] ищется в виде полинома над элементарными алгоритмами максимальной высоты [3], т. е. разности между максимальной оценкой неправильных пар (объект, класс) и минимальной оценкой правильных пар. Задача максимизации высоты элементарных слагаемых подробно рассматривалась ранее [1, 4], были построены алгоритмы, точные и приближенные, получены оценки сложности и проведены практические исследования на модельных и реальных задачах.

Напомним схему алгоритма. Работа метода начинается с перехода к вспомогательной задаче. Для каждой пары (обучающий объект S_j, контрольный объект S^i) в новой задаче порождается объект $|S^i - S_j|$, который относится в один из классов: правильный или неправильный, в зависимости от правильности пары $\big(S^i, K(S_j)\big)$, где $K(S_j)$ — класс объекта S_j. Модуль берется покоординатно. Между АВО, который задается набором порогов функции близости, и гиперпараллелепипедами вспомогательной задачи существует соответствие, при этом показано, что АВО максимальной высоты соответствует правильный прямоугольник, т. е. минимальный гиперпараллелепипед, содержащий некоторую комбинацию правильных объектов. Таким образом, задача поиска АВО максимальной высоты сводится к перебору множества правильных прямоугольников или некоторого их подмножества в случае приближенных алгоритмов.

В результате тестирования на модельных и реальных задачах в качестве алгоритма для поиска оптимальных слагаемых был выбран алгоритм покоординатного подъема, в котором перебор правильных прямоугольников осуществляется по одной координате за раз, и на каждом шаге выбирается алгоритм максимальной высоты.

Построение полинома начинается с разбиения обучающей выборки на собственно обучающую и контрольную. Важно отличать контрольную выборку от тестовой, которая используется для проверки качества и не используется при обучении. Разбиение производится случайно в некоторой пропорции, которая является параметром настройки алгоритма. Далее множество контрольных объектов перебирается, и для каждого объекта строится элементарный АВО максимальной высоты с использованием всей обучающей выборки.

Такая схема показала качество распознавания на уровне аналогов, а в некоторых случаях даже превосходит их. Однако, при пополнении обучающей информации (в общем случае и контрольной, и обучающей выборок) алгоритм требует полной перенастройки. Тем не менее, достаточно естественным образом схема может быть обобщена на случай дина-

мического пополнения информации, описание такой модификации приводится в следующем разделе.

Индуктивная модификация

При переходе к динамической задаче необходимо пройти все основные этапы обычного алгоритма. Пусть имеется обучающая выборка и алгоритм A, настроенный на её распознавание. Алгоритм A представляет собой полином из слагаемых максимальной высоты. Пусть теперь обучающая выборка пополняется новым объектом. Прежде всего надо определить, в какую из выборок его отнести: контрольную или собственно обучающую.

Согласно общей схеме, разделение происходит случайным образом с заданной вероятностью. Если объект оказывается отнесенным в контрольную выборку, то необходимо всего лишь построить дополнительное слагаемое максимальной высоты, взяв новый объект в качестве центрального, и определить его степень согласно статической схеме.

Если же объект оказывается отнесенным в обучение, надо перенастроить все найденные ранее слагаемые. Для этого предлагается использовать тот же покоординатный метод, но на значительно меньшей решетке. Фактически, на n-мерном кубе, образованном новым объектом и вершиной найденного ранее для этого слагаемого правильного параллелепипеда. В целях улучшения качества схему можно дополнить полным переобучением некоторых слагаемых после заданного количества индуктивных шагов.

Работа выполнена при поддержке грантов РФФИ № 06-01-08045 офи, № 05-01-00332, № 05-07-90333, № 06-01-00492, а также гранта Президента РФ, НШ-5833.2006.1.

Литература

[1] *Докукин А. А.* Об одном подходе к оптимизации АВО // Доклады 11-й Всероссийской конференции Математические методы распознавания образов ММРО-11, Москва, 2003. — С. 68–71.

[2] *Журавлев Ю. И.* Корректные алгебры над множеством некорректных (эвристических) алгоритмов II // Кибернетика. — 1977. — № 6. — С. 21–27.

[3] *Журавлев Ю. И., Исаев И. В.* Построение алгоритмов распознавания, корректных для заданной контрольной выборки // ЖВМиМФ. — 1979. — Т. 19, № 3.

[4] *Dokukin A. A.* Optimal method for constructing AEC of maximal height in context of pattern recognition // Pattern recognition and image analysis. — 2005. — Vol. 15, No. 1.

Об алгоритме классификации на основе полного решающего дерева

Дюкова Е. В., Песков Н. В.
djukova@ccas.ru, peskov@ccas.ru
Москва, ВЦ РАН

Решающие деревья — это один из наиболее популярных инструментов для решения задач классификации по прецедентам в случае, когда исследуемые объекты описываются в признаковом пространстве.

Синтез решающего дерева представляет собой итерационный процесс. Как правило, для построения очередной вершины дерева выбирается признак, наилучшим образом удовлетворяющий некоторому критерию ветвления. По значениям этого признака и осуществляется ветвление, далее указанная процедура повторяется для каждой из ветвей. Описанный подход обуславливает основные достоинства метода, а именно, решающее правило строится быстро, получается достаточно простым и хорошо интерпретируемым.

Однако, в случае, когда два или более признака удовлетворяют рассматриваемому критерию ветвления в равной или почти равной мере, выбор одного из этих признаков осуществляется практически случайным образом. При этом в зависимости от выбранного признака построенные деревья могут существенно отличаться как по составу используемых признаков, так и по своим распознающим качествам.

В докладе рассматривается следующий подход к решению указанной проблемы. При возникновении ситуации, когда два или более признака удовлетворяют критерию в равной мере, предлагается проводить ветвление по каждому из этих признаков независимо. Полученная в результате конструкция названа полным решающим деревом.

Данный подход продемонстрирован на примере усовершенствования алгоритма построения допустимого разбиения (далее АДР). Критерий ветвления в АДР представляет собой набор условий с разным приоритетом. При выборе очередной вершины сначала проверяется условие с наибольшим приоритетом. Ищется признак с наименьшим номером, для которого это условие выполняется. Если ни один признак не удовлетворяет рассматриваемому условию, то проверяется следующее по порядку условие с более низким приоритетом.

Проведенное экспериментальное исследование на реальных задачах показало, что точность распознавания при использовании полного решающего дерева существенно выше точности распознавания АДР. Кроме того, было проведено сравнение новой модели, названной ПРД, с алгоритмом C4.5, широко используемого в промышленных системах анализа

данных. На рассмотренных задачах новая модель ведет себя не хуже алгоритма C4.5, а на некоторых существенно превосходит C4.5.

Работа выполнена при поддержке РФФИ, проекты № 07-01-00516 и № 06-07-89299 и гранта Президента РФ по поддержке ведущих научных школ НШ № 5833.2006.1.

Литература

[1] *Донской В. И., Башта А. И.* Дискретные модели принятия решений при неполной информации. — Симферополь: Таврия, 1992. — С. 33–74.

О подходах к синтезу случайных и решающих лесов

Дюличева Ю. Ю.

dyulicheva_yu@mail.ru

Симферополь, Таврический национальный университет им. В. И. Вернадского

Увеличение сложности структуры решающего дерева (РД) и уменьшение его обобщающей способности наблюдаются при все более точной, безошибочной «настройке» РД на исходную обучающую информацию. Такое поведение, характерное для большинства алгоритмов обучения распознаванию, называется переподгонкой (overfitting) или переобучением. Разработка критериев редукции (pruning) [5] и наращивания (grafting) [9] ветвей решающих деревьев, а также алгоритмов синтеза совокупностей решающих деревьев относительно «простой» структуры — лесов — направлена на поиск компромисса между излишним усложнением структуры РД и получением как можно более высокой оценки качества решающих правил, синтезируемых по обучающей выборке.

Перечислим вкратце основные подходы к синтезу случайных и решающих лесов. Существенным требованием при построении лесов является поиск и синтез различных решающих деревьев, входящих в состав леса. Корректные на обучающей выборке деревья случайного решающего леса (random decision forest), предложенного в работе [6], строятся на подмножествах признаков, случайно выбранных из исходного множества признаков. В работе [4] предложена модель случайных лесов (random forests). При построении каждого дерева случайного леса осуществляется выбор наиболее информативного признака на основе случайным образом сгенерированного для каждой вершины РД подмножества признаков. В работе [8] предложен алгоритм «дровосека» (lumberjack algorithm) для синтеза решающего леса с переходами по ссылкам (linked decision forest), основанный на анализе структуры РД с установкой ссылки перехода на другое дерево леса при появлении в синтезируемом дереве одинаковых поддеревьев. В работе [7] предложен эволюционный подход к построению решающего леса. В работах [1, 2, 3] предложена индуктивная модель эм-

пирического решающего леса. При построении каждого дерева решающего леса формируется область отказа. Для построения следующего дерева леса используются признаки, не участвовавшие в обучении предыдущих деревьев, и объекты, попавшие в пересечение областей отказа всех деревьев, входящих в состав леса, после чего производится дообучение дерева на оставшихся объектах обучающей выборки.

В докладе приводится сравнительный анализ описанных выше подходов к синтезу случайных и решающих лесов, а также обсуждается ряд улучшений алгоритма синтеза эмпирического решающего леса, предложенного в работах [1, 2, 3].

Литература

[1] *Донской В. И., Дюличева Ю. Ю.* Индуктивная модель r-корректного эмпирического леса // Труды международной конференции по индуктивному моделированию, Львов. — 2002. — № 2. — С. 54–58.

[2] *Дюличева Ю. Ю.* Оценка VCD r-редуцированного эмпирического леса // Таврический вестник информатики и математики. — 2003. — № 2. — С.35–43.

[3] *Дюличева Ю. Ю.* Применение эмпирического решающего леса для фильтрации обучающих данных // Таврический вестник информатики и математики. — 2006. — № 1. — С. 55–61.

[4] *Breiman L.* Random Forests // Machine Learning. — 2001. — No 45. — Pp. 5–32.

[5] *Esposito F., Malerba D., Semeraro G. A.* Comparative Analysis of Methods for Pruning Decision Trees // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 1997. — Vol. 19, No 5. — Pp. 476–491.

[6] *Ho T. K.* The Random Subspace Method for Constructing Decision Forests // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 1998. — Vol. 20, No 8. — Pp. 832–844.

[7] *Rouwhorst S. E., Engelbrecht A. P.* Searching the Forest: Using Decision Trees as Building Blocks for Evolutionary Search in Classification Databases // Congress on Evolutionary Computation CECOO. — 2000. — Vol. 1. — Pp. 633–638.

[8] *Uther William T. B., Veloso Manuela M.* The Lumberjack Algorithm for Learning Linked Decision Forests // Symp. on Abstraction, Reformulation and Approximation, LNAI. — Springer Verlag, 2000. — Vol. 1864. — Pp. 219–230.

[9] *Webb G. I.* Decision Tree Grafting // 15th Int. Joint Conf. on Artificial Intelligence, Nagoya, Japan. — Morgan Kaufmann, 1997. — Pp. 846–851.

Матричная коррекция несовместных систем линейных алгебраических уравнений как обобщение метода наименьших квадратов

Ерохин В. И.
erohin_v_i@mail.ru
Борисоглебск, Борисоглебский гос. пед. ин-т

Рассматриваются несовместные системы линейных алгебраических уравнений (СЛАУ) и связанные с ними специфические задачи математического программирования (задачи матричной коррекции), заключающиеся в минимальном (по некоторой матричной норме) изменении коэффициентов исследуемых СЛАУ, обеспечивающем их совместность. Анализируется устойчивость решений скорректированных систем, которая сравнивается с устойчивостью псевдорешений, полученных с помощью метода наименьших квадратов (МНК) и его модификаций, а также с помощью Тихоновской регуляризации. Указываются некоторые возможные приложения задач матричной коррекции несовместных СЛАУ в качестве инструментов параметрической идентификации линейных и линеаризуемых моделей.

Задачи матричной коррекции несовместных СЛАУ

Пусть $Ax = b$, где $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ — некоторая СЛАУ, $X(A,b) \triangleq \{x \,|Ax = b\}$ — допустимое множество ее решений, соотношения между рангом матрицы A и ее размером — произвольные. Будем предполагать, что $X(A,b) = \varnothing$. Символом $\|\cdot\|$ будем, в зависимости от контекста, обозначать евклидову векторную или матричную норму.

Следующие две (относительно простые) задачи матричной коррекции исторически были исследованы одними из первых:

$$Z_1[1,2,3]\colon \quad \big\|[H\ \ -h]\big\| \to \inf_{X(A+H,b+h)\neq\varnothing};$$

$$Z_2[1,3]\colon \quad \|H\| \to \inf_{X(A+H,b)\neq\varnothing};$$

где $H \in \mathbb{R}^{m\times n}$ — некоторая матрица, $h \in \mathbb{R}^m$ — некоторый вектор.

Заметим, что классический МНК может быть записан в виде

$$Z_0\colon \quad \|h\| \to \inf_{X(A,b+h)\neq\varnothing},$$

а за задачей Z_1 в зарубежной литературе закрепилось название TLS (Total Least Norm — обобщенный метод наименьших квадратов).

В настоящее время усилиями отечественных и зарубежных ученых задачи Z_1 и Z_2 подверглись многочисленным усложнениям и модификациям. Указанные усложнения и модификации выразились, например,

в переходе от евклидовой матричной нормы к другим, например — полиэдральным нормам [4], использованию взвешенных различными способами норм [3, 4], в переходе к коррекции несобственных задач линейного программирования (ЛП) [1, 4, 5], в появлении дополнительных ограничений в виде запрета на коррекцию некоторых строк и (или) столбцов матрицы (расширенной матрицы) исследуемой системы [3, 4] и даже в виде запрета на коррекцию произвольного множества элементов расширенной матрицы коэффициентов [4]. Получены определенные результаты для задач матричной коррекции несовместных СЛАУ, матрицы (расширенные матрицы) которых имеют специальную структуру, например, блочную [6] или структуру матриц Тёплица (Ганкеля) [7, 8].

Сравнение МНК, регуляризации Тихонова и матричной коррекции

В этом параграфе будем рассматривать переопределенные несовместные неоднородные СЛАУ. Очевидная причина такого выбора — распространенность такого рода систем в задачах обработки результатов наблюдений. Менее очевидная причина связана с условиями существования решений задач матричной коррекции [3, 4]. Под псевдорешением, полученным с помощью МНК, будем понимать нормальное псевдорешение СЛАУ $\hat{x} = A^{+}b$, где A^{+} — псевдообратная матрица (обобщенная обратная матрица Мура-Пенроуза). Под псевдорешением, полученным с помощью регуляризации Тихонова [9], будем понимать вектор $x_{\mu\delta}$, получаемый из решения задачи T относительно $A_{\mu\delta}$, $b_{\mu\delta}$, $x_{\mu\delta}$:

$$T\colon \begin{cases} \|x_{\mu\delta}\| \to \min; \\ \|A - A_{\mu\delta}\| \leqslant \mu; \quad \|b - b_{\mu\delta}\| \leqslant \delta; \quad x_{\mu\delta} \in X(A_{\mu\delta}, b_{\mu\delta}); \end{cases}$$

где $\mu \geqslant 0,\ 0 \leqslant \delta < \|b\|$ — параметры, воплощающие априорную информацию о величине отклонений элементов матрицы A и вектора b от гипотетических (неизвестных) точных значений. Наконец, под псевдорешениями, получаемыми с помощью задач Z_1 и Z_2 будем понимать векторы $x_{[H^*-h^*]} \in X(A+H^*, b+h^*)$ и $x_{H^*} \in X(A+H^*, b)$, где $[H^*\ -h^*]$ — решение задачи Z_1, H^* — решение задачи Z_2.

Наиболее интересным является случай, при котором гипотетическая точная матрица СЛАУ не имеет полного столбцевого ранга, однако вследствие ошибок наблюдений, погрешностей дискретизации или погрешностей, обусловленных вычислениями в конечной разрядной сетке, предъявляемая исследователю матрица A оказывается полноранговой. Хорошо известно, что в указанных условиях классический МНК приводит к неустойчивым решениям, а метод Тихонова при использовании достоверных значений μ и δ от указанного недостатка свободен. Что касается решений, получаемых с помощью матричной коррекции,

то их устойчивость при решении практических задач хорошо известна. В то же время, теоретически этот вопрос до настоящего времени был изучен очень слабо, однако автору удалось показать, что при выполнении некоторых дополнительных ограничений, и, в частности, соответствующей «малости» параметров μ и δ задачи Z_1, Z_2 и им подобные становятся эквивалентными задаче T, что позволяет теоретически обосновать устойчивость получаемых с их помощью псевдорешений.

Так, справедлива следующая теорема.

Теорема 1. *Если решение задачи T существует и выполняется условие $\mu < (\|b - A\hat{x}\| - \delta)/\|\hat{x}\|$, то найдется подходящая задача матричной коррекции, дающая то же псевдорешение $x_{\mu\delta}$ исследуемой СЛАУ, что и задача T. При этом, если $\bar{A}x = \bar{b}$ — гипотетическая точная совместная СЛАУ, $\bar{x}$ — ее нормальное решение, $\left\|\bar{A} - A\right\| \leqslant \mu$, $\left\|\bar{b} - b\right\| \leqslant \delta$, то $\lim\limits_{\mu,\delta\to 0} x_{\mu\delta} = \bar{x}$, т. е. вектор $x_{\mu\delta}$ является устойчивым приближением к вектору $\bar{x}$.*

Применение матричной коррекции в задачах идентификации

Представленный ниже перечень задач является иллюстративным и не претендует на полноту.

1. Оценивание параметров линейной системы в случае, когда ошибкам подвержены не только элементы вектора b но и матрицы A. (Несложно показать, что если ошибки в расширенной матрице исследуемой системы подчиняются нормальному распределению с нулевым средним и одинаковой для всех коэффициентов дисперсией, то псевдорешения исследуемой СЛАУ, получаемые в результате решения задачи Z_1, совпадут с решением, получаемым методом максимального правдоподобия).
2. Идентификации параметров зашумленного стационарного временного ряда. (При использовании модификаций метода де'Прони сводится к матричной коррекции несовместных СЛАУ с матрицами Ганкеля или Тёплица)
3. Решение обратной задачи для модели, заданной интегральным уравнением. (Несовместная СЛАУ появляется в результате дискретизации модели. При этом ошибки в коэффициентах матрицы системы могут быть обусловлены погрешностями дискретизации и погрешностями, вносимыми арифметикой с конечной разрядностью).

Литература

[1] *Еремин И. И., Мазуров В. Д., Астафьев Н. Н.* Несобственные задачи линейного и выпуклого программирования. — М: Наука, 1983. — 336 с.

[2] *Van Huffel S.* Analysis of the total least squares problem and its use in parameter estimation // PhD thesis, Dept. of Electr. Eng., Kattholieke Universiteit, Leuven, Belgium, 1987.

[3] *Горелик В. А., Ерохин В. И.* Оптимальная матричная коррекция несовместных систем линейных алгебраических уравнений по минимуму евклидовой нормы. — М: ВЦ РАН, 2004. — 193 с.

[4] *Горелик В. А., Ерохин В. И.* Оптимальная (по минимуму полиэдральной нормы) матричная коррекция несовместных систем линейных алгебраических уравнений и несобственных задач линейного программирования // Моделирование, декомпозиция и оптимизация сложных динамических процессов, М: ВЦ РАН, 2004. — С. 35–63.

[5] *Горелик В. А.* Матричная коррекция задачи линейного программирования с несовместной системой ограничений // ЖВМиМФ. — 2001. — Т. 41, № 11. — С. 1697–1705.

[6] *Горелик В. А., Ерохин В. И., Печенкин Р. В.* Оптимальная матричная коррекция несовместных систем линейных алгебраических уравнений с блочными матрицами коэффициентов // Дискретный анализ и исследование операций. Сер. 2. — 2005. — Т. 12, № 2. — С. 3–22.

[7] *Lemmerling P.* Structured total least squares: analysis, algorithms and applications // Ph.D. thesis, Kattholieke Universiteit, Leuven, Belgium, 1999.

[8] *Горелик В. А., Ерохин В. И., Печенкин Р. В.* Идентификация сигнала в виде суммы экспонент с помощью методов матричной коррекции несовместных линейных систем с матрицами Тёплица // Моделир., декомпоз. и оптимиз. сложных динамич. процессов, М: ВЦ РАН, 2003. — С. 74–88.

[9] *Тихонов А. Н.* О приближенных системах линейных алгебраических уравнений // ЖВМиМФ — 1980. — Т. 20, № 6. — С. 1373–1383.

Методы быстрого поиска ближайшего аналога в большой базе изображений

Загоруйко Н. Г., Борисова И. А., Дюбанов В. В., Кутненко О. А.

zag@math.nsc.ru

Новосибирск, Институт математики СО РАН

Типичная задача при использовании больших БД состоит в поиске объекта, который является ближайшим аналогом нового (контрольного) объекта. Для решения этой задачи разработано семейство алгоритмов направленного поиска «Локатор» [1], основанных на пошаговом сокращении количества конкурирующих объектов и фокусировании внимания на тех объектах, которые имеют наибольшие шансы стать победителями в этой конкуренции.

Для снижения размерности пространства за счет исключения малоинформативных признаков разработан алгоритм направленного поиска GRAD [2]. Нами предложен новый критерий информативности, основанный на функции конкурентного сходства (FRiS-функции) [3]. Эта же функция применяется и в новом алгоритме кластеризации FRiS-Cluster [4].

Эффективность разработанных алгоритмов исследовалась в модельных тестах и при решении реальных задач. Для проведения исследований была разработана программа LOCATEST [5].

Содержание задач поиска аналога и методов их решения зависит от ответов на следующие вопросы: *Описываются ли образы значениями признаков X или парными расстояниями r между всеми образами? Если решается задача распознавания в признаковом пространстве , то какова метрика этого пространства:* L_∞, L_2 *или* L_1*? Задается ли порог d допустимых различий между объектом Z и аналогом, или нет, т. е. ищется d-аналог или abs-аналог? Требуется ли найти все t d-аналогов или один самый близкий?* Сочетания разных ответов на эти вопросы порождает 12 различных задач поиска аналога. Алгоритмы их решения описаны в [1]. Здесь для примера приведены краткие характеристики некоторых из них.

Для поиска всех аналогов, которые удовлетворяют условию $R_\infty \leqslant d$, используется алгоритм **Локатор-1**. Все K объектов БД и контрольный объект Z проецируются на одну из N координат. Если расстояние от Z до объекта O_i больше d, то i–й объект из списка конкурентов на роль ближайшего вычеркивается. Для оставшихся $k < K$ объектов та же процедура повторяется с использованием проекции на вторую координату, и т. д. Данный алгоритм сокращает время поиска приблизительно в 6 раз.

Если объекты описаны не признаками, а матрицей M парных расстояний между ними, и если требуется найти одного abs-аналога объекту Z, то применяется алгоритм **Локатор-8**. Для исключения тех объектов, которые не могут быть ближайшими аналогами, используется следующее легко доказуемое

Утверждение 1. *Объекты, удаленные от объекта* O_i *на расстояние* $R > 2R(z_i)$*, находятся по отношению к объекту Z дальше, чем объект* O_i*.*

Для выбора следующего локатора по i–й строке матрицы M находим объект O_j, для которого величина $F = |R(z_i) - R(j_i)|,\ j = 1, \ldots, k_l$, минимальна. Знание расстояний от точки Z до двух объектов-локаторов позволяет более точно пеленговать позицию точки Z среди оставшихся объектов. Такие шаги повторяются до тех пор, пока в списке претенден-

тов не останется один объект. На ряде реальных задач затраты времени сокращались на 2–3 порядка.

Процесс поиска ускоряется, если в качестве локаторов выбирать не случайные объекты, а такие, которые находятся в центре локальных сгустков. Для поиска таких локаторов применяется алгоритм кластеризации FRiS-Cluster.

Недостаток существующих *мер сходства* состоит в том, что сходство рассматривается в качестве абсолютной характеристики, в то время как ответы на вопросы типа «близко–далеко?» или «похож–не похож?» зависят от ответа на вопрос «по сравнению с чем или кем?». Это свойство относительных понятий сходства и различия может быть выражено функцией F конкурентного сходства (FRiS-функцией). Если расстояния от объекта Z до двух ближайших объектов a и b равны R_a и R_b, то сходство Z с объектом a равно $F(a) = (R_b - R_a)/(R_a + R_b)$. Значения F меняются в пределах от $+1$ до -1. Если контрольный объект Z совпадает с объектом a, то $R_a = 0$ и $F(a) = 1$, а $F(b) = -1$. При расстояниях $R_a = R_b$ значения $F(a) = F(b) = 0$, что указывает на границу между образами.

При использовании алгоритма FRiS-Cluster на его первых шагах все M объектов заданного множества A принадлежат одному кластеру. В связи с этим вводится конкурирующее множество из виртуальных объектов, удаленных от каждого объекта множества A на расстояние R^*. Произвольный объект O_i множества A назначается центром первого кластера, оцениваются расстояния R_j до него от всех остальных объектов O_j и определяются значения функции $F(i) = (R^* - R_j)/(R^* + R_j)$. Вычисляется сумма значений функций сходства F_s всех объектов первого кластера со своим центром. Затем в качестве центра второго кластера выбирается объект, набравший наибольшее значение F_s в конкуренции с центром первого кластера. Процесс увеличения числа кластеров k останавливается, когда достигается первый локальный максимум функции $F_s(k)$.

Если объекты разделены на классы, то для выбора признаков может быть использован алгоритм GRAD. На его первом этапе создаются вторичные признаки в виде «гранул» — наиболее информативных комбинаций из двух и трёх зависимых признаков. На множестве гранул выполняется последовательность процедур добавления (Addition) наиболее информативных и исключения (Deletion) наименее информативных признаков. Их информативность оценивается по среднему значению функции сходства F_s всех объектов обучающей выборки со своими эталонами. Преимущество этого F_s критерия по сравнению с распространенным критерием U минимума ошибок на обучении подтверждена на большом

числе задач. Применение критерия F позволило решить «проблему пригодности признаков», поставленную А. Н. Колмогоровым [6].

В докладе приводятся примеры решения задач выбора ближайшего аналога в базе генетических данных и базе микрофотографий. Показано, что для ускорения поиска ближайшего аналога следует использовать различные средства: направленный перебор претендентов (алгоритмы Локатор), предварительное сокращение размерности пространства (алгоритм GRAD) и предварительный выбор числа сгустков объектов и их центров (алгоритм FRiS-Cluster).

Работа выполнена при поддержке РФФИ, грант № 05-01-00241.

Литература

[1] *Загоруйко Н. Г., Дюбанов В. В.* Семейство алгоритмов «Локатор» для быстрого поиска ближайшего аналога // СибЖИМ. — 2006. — Т. 38, № 5. — С. 54–62.

[2] *Загоруйко Н. Г., Кутненко О. А.* Алгоритм GRAD для выбора признаков // Труды VIII Межд. конф. «Применение многомерного статистического анализа в экономике и оценке качества», Москва: Изд. МЭСИ, 2006. — С. 81–89.

[3] *Загоруйко Н. Г.* Методы интеллектуального анализа данных, основанные на функции конкурентного сходства // Автометрия (в печати).

[4] *Борисова И. А.* Алгоритм таксономии FRiS-Tax // Научный вестник НГТУ (в печати).

[5] `www.dvv-2.gorodok.net`

[6] *Колмогоров А. Н.* К вопросу о пригодности найденных статистическим путем формул прогноза // Завод. лаб. — 1933. — № 1. — С. 164–167.

Выявление групп объектов, описанных набором многомерных временных рядов

Ивахненко А. А., Каневский Д. Ю., Рудева А. В., Стрижов В. В.

`strijov@ccas.ru`

Москва, Вычислительный центр РАН

Далеко не во всех прикладных задачах стандартные способы оценивания корреляции между временными рядами имеют «разумную» содержательную интерпретацию. В данной работе предполагается, что исходная информация представлена матрицей «объект–показатель–время», и предлагается методика выделения групп объектов, демонстрирующих сходное поведение относительно заданного показателя, либо относительно заданного набора показателей.

Методика основана на построении матриц сходства между всеми парами временных рядов. Сходство (или расстояние) определяется числом

размеченных интервалов совместного роста или спада. Используется понятие «разметки», определенное в [1], и алгоритм автоматической разметки временных рядов [2]. По матрице сходства однозначно определяются группы объектов со сходным поведением. Группы объектов задаются матрицей смежности и представляются в виде множества графов [3].

Описан алгоритм, который отыскивает все группы сходных объектов. Для этого алгоритм последовательно выполняет следующие операции: разметку основных рядов, построение матрицы сходства временных рядов, нахождение групп объектов для фиксированного показателя, нахождение групп объектов для произвольного набора показателей.

Разметка временных рядов

Дано декартово произведение множества объектов, множества показателей и множества моментов времени $\{x_{ijt}\}_{i,j,t=1}^{I,J,T}$. Элементом x_{ijt} этого произведения является значение j-ого показателя на i-ом объекте в момент времени t.

Временной ряд является размеченным, если каждому его элементу поставлен в соответствие знак из алфавита $\mathcal{A}$. Примером алфавита может служить множество $\mathcal{A} = \{\mathrm{U}, \mathrm{D}, \mathrm{N}\}$, где символы U и D интерпретируются как увеличение и уменьшение значения показателя данного объекта в данный момент времени, символ N соответствует отказу от классификации. Таким образом каждому временному ряду $\{x_{ijt}\}_{t=1}^{T}$ поставлены в соответствие ряды $\{u_{ijt}\}_{t=1}^{T}$ и $\{d_{ijt}\}_{t=1}^{T}$, где $u_{ijt} \in \{\mathrm{U}, \mathrm{N}\}$, $d_{ijt} \in \{\mathrm{D}, \mathrm{N}\}$, интерпретируемые как ряды, размеченные интервалами роста и падения значений показателя на данном объекте.

Приведем пример условий, определяющих разметку временного ряда. Элементу из упорядоченной последовательности $x_{ijt_1}, \ldots, x_{ijt_N}$ ставится в соответствие знак U, если выполнено:

$$\begin{cases} \bar{N} \leqslant N \in \mathbb{Z}; \\ x_{ijt_1} < x_{ijt_2}; \\ x_{ijt_{N-1}} < x_{ijt_N}; \\ \bar{D} < \left((x_{ijt_N} - x_{ijt_1}) - \sum\limits_{\forall \tau \in T_D} (x_{ijt_\tau} - x_{ijt_{\tau+1}}) \right) (x_{ijt_N} - x_{ijt_1})^{-1}; \\ T_D = \{\tau \colon x_{ijt_{\tau+1}} < x_{ijt_\tau}\}; \\ x_{ijt_1} \leqslant x_{ijt_\tau}, \quad \tau = 1, \ldots, N; \\ \bar{L} \leqslant (x_{ijt_N} - x_{ijt_1}). \end{cases},$$

где $\bar{N}$, $\bar{D}$, $\bar{L}$ — заданные параметры. Каждому элементу из упорядоченной последовательности $x_{ijt_1}, \ldots, x_{ijt_N}$ ставится в соответствие

знак D, если выполнены вышеприведенные условия с учетом замены x_{ijt} на $(-x_{ijt} - \max\limits_{t} x_{ijt})$. В противном случае ставится знак N.

Вычисление функции сходства временных рядов

Для каждой пары рядов $\{u_{ijt}\}_{t=1}^{T}$ и $\{d_{kjt}\}_{t=1}^{T}$ вычисляется функция сходства ρ_{ikj}^{U}, где $i, k = 1, \ldots, I$ — номера объектов, а значение $j \in \{1, \ldots, J\}$ — номер показателя. Значение этой функции равно числу интервалов, удовлетворяющих следующим условиям.

Рядам с индексами i, k поставлены в соответствие наборы $\{\mathfrak{T}_{i\xi j}^{\mathrm{U}}\}$ интервалов $\{\mathfrak{T}_{i\xi j}^{\mathrm{U}}\}, \{\mathfrak{T}_{k\zeta j}^{\mathrm{U}}\}$, составленные из последовательно идущих индексов знаков U соответствующих рядов, где порядковые номера интервалов $\xi, \zeta \in \Xi = \{1, \ldots, \lfloor \frac{T}{2} \rfloor\}$, а элементы $\mathfrak{T}_{i\xi j}^{\mathrm{U}}, \mathfrak{T}_{k\zeta j}^{\mathrm{U}} \subset= \{1, \ldots, T\}$. Значение функции сходства ρ_{ikj}^{U} равно числу интервалов, удовлетворяющих условию

$$\rho_{ikj}^{\mathrm{U}} = \#\Big\{(\xi, \zeta) \;\Big|\; \big|\mathfrak{T}_{i\xi j}^{\mathrm{U}} \cap \mathfrak{T}_{k\zeta j}^{\mathrm{U}}\big| + \bar{\Delta} \geqslant \big|\mathfrak{T}_{i\xi j}^{\mathrm{U}} \cup \mathfrak{T}_{k\zeta j}^{\mathrm{U}}\big|\Big\},$$

где $\bar{\Delta}$ — заданный параметр.

Функция сходства ρ_{ikj}^{D} для каждой пары рядов $\{d_{ijt}\}_{t=1}^{T}$ и $\{d_{kjt}\}_{t=1}^{T}$ вычисляется так же.

Полученные функции сходства ρ_{ikj}^{U} и ρ_{ikj}^{D} задают трехиндексную матрицу сходства $S^{\mathrm{UD}} = \{\rho_{ikj}^{\mathrm{U}} + \rho_{ikj}^{\mathrm{D}}\}_{i,k,j=1}^{I,I,J}$. Каждый элемент этой матрицы является значением функции сходства, определенной на всех возможных парах объектов и на всех показателях.

Поиск группы объектов для одного показателя

Все действующие группы, включающие объекты, действующие сходным образом и имеющие на фиксированном показателе расстояние более $\bar{R}$, однозначно заданы полносвязным графом на матрице смежности S. Эта матрица получается путем порогового отсечения значений элементов матрицы S^{UD}:

$$S = \{\rho_{ikj}\}_{i,k,j=1}^{I,I,J}, \text{ где } \rho_{ikj} = \begin{cases} 1, & \text{если } (\rho_{ikj}^{\mathrm{U}} + \rho_{ikj}^{\mathrm{U}}) \geqslant \bar{R}; \\ 0, & \text{в другом случае.} \end{cases}$$

Полносвязные графы, полученные на трехиндексной матрице сходства S, представимы в виде множества $\{G_{j\gamma}\}_{j,\gamma=1}^{J,\Gamma_j}$, где сходно действующая на j-й акции группа с номером $\gamma \in \{1, \ldots, \Gamma_j\}$ есть набор объектов $G_{j\gamma} = \{g_{j\gamma 1}, \ldots, g_{j\gamma l_{j\gamma}}\}$, где $g \in \{1, \ldots, I\}$ — номер объекта.

Группы объектов для нескольких показателей

Группа объектов P_s сходно действует по нескольким показателям, если существует непустое пересечение элементов групп $G_{j\gamma}$ такое, что

выполняются условия

$$P_s = \bigcap_{\substack{j\in\mathcal{J}\subseteq\{1,\ldots,J\},\\ \gamma\in\{1,\ldots,\Gamma_j\}}} G_{j\gamma},$$

где $|\mathcal{J}| \to \max$, $|P_s| \to \max$, s — номер полученной группы.

Примером использования вышеприведенной методики может служить задача поиска группы магазинов торговой сети, в которых набор товаров продается схожим образом. Другим примером может служить поиск групп участников биржевых торгов, действующих синхронно на наборе финансовых инструментов.

Работа выполнена при поддержке РФФИ, проекты № 07-07-00181, № 07-07-00372.

Литература

[1] *Рудаков К. В., Чехович Ю. В.* Алгебраический подход к проблеме синтеза обучаемых алгоритмов выделения трендов // Доклады РАН. — 2003. — Т. 388, № 1. — С. 33–36.

[2] *Васин Е. А., Костенко В. А., Коваленко Д. С.* Автоматическое построение алгоритмов, основанных на алгебраическом подходе, для распознавания предаварийных ситуаций динамических систем // Искусственный интеллект. — 2006. — № 2. — С. 130–134.

[3] *Tamm У. Т.* Теория графов. — Москва: Мир, 1988. — 314 с.

Поиск оптимальной метрики в задачах классификации с порядковыми признаками

Иофина Г. В., Кропотов Д. А.
giofina@gmail.com, dkropotov@yandex.ru
Москва, МФТИ, ВЦ РАН

Рассматривается класс задач распознавания, в которых объекты описываются n порядковыми признаками, представленными набором целых чисел от 0 до $N-1$. Для решения таких задач предполагается применять метрические алгоритмы классификации, такие как метод ближайших соседей, метод потенциальных функций или алгоритмы вычисления оценок [1]. Для этого необходимо вводить метрики как на отдельных признаках, так и на векторах признаковых описаний. В работе находится наилучшая метрика, при которой взвешенная разность между межклассовым и средним внутриклассовым расстояниями максимальна. Таким образом, найденная метрика отображает структуру задачи, т. е. дает маленькие расстояния для объектов из одного класса и большие — для объектов из разных классов.

Пусть на множестве $\tilde{N} = \{0, 1, \ldots, N-1\}$ допустимых значений признаков задано естественное отношение порядка $0 \leqslant 1 \leqslant \ldots \leqslant N-1$. Произвольная метрика $\rho(i,j)$ на множестве $\tilde{N}$ задается матрицей $\{c_{ij}\}$ размера $N \times N$, симметричной, с нулевой диагональю, элементы которой, находящиеся выше главной диагонали, не убывают по строкам и не возрастают по столбцам (удовлетворяют отношению порядка).

Пусть даны два конечных множества объектов:
$a_u = (a_u^1, \ldots, a_u^n) \in \tilde{N}^n$, $u = 1, \ldots, m$ — объекты класса K_1;
$b_v = (b_v^1, \ldots, b_v^n) \in \tilde{N}^n$, $v = 1, \ldots, l$ — объекты класса K_2.

Под расстоянием между объектами $a_u \in K_1$ и $b_v \in K_2$ будем понимать величину $\rho(a_u, b_v) = \sum\limits_{i=1}^{n} \rho_i(a_u^i, b_v^i)$, где ρ_i — метрика, заданная на i-ом признаке.

Под внутриклассовыми и межклассовым расстояниями будем понимать величины

$$\alpha_1 = 1/N_1 \sum_{u=1}^{m} \sum_{v=1}^{m} \rho(x_u, x_v), \quad x_u, x_v \in K_1;$$

$$\alpha_2 = 1/N_2 \sum_{u=1}^{l} \sum_{v=1}^{l} \rho(x_u, x_v), \quad x_u, x_v \in K_2;$$

$$\beta = 1/M \sum_{u=1}^{m} \sum_{v=1}^{l} \rho(x_u, x_v), \quad x_u \in K_1, x_v \in K_2;$$

где N_1, N_2 и M — нормировочные множители.

Задачу максимизации взвешенной разности между межклассовым и средним внутриклассовым расстояниями можно представить в виде следующей оптимизационной задачи:

$$\beta - 0.5\lambda(\alpha_1 + \alpha_2) \to \max_{\rho_i, i=1,\ldots,n}, \tag{1}$$

где λ можно рассматривать как отношения весов межклассового и среднего внутриклассового расстояний.

Считается, что на каждом признаке задана своя функция расстояния, и признаки не зависят друг от друга, поэтому функции расстояний для разных признаков можно искать независимо. Далее в работе рассматривается один признак и, следовательно, ищется одна функция расстояния.

Обозначим количество нулей, единиц, двоек, троек, и т. д. среди значений признака у объектов первого класса через $\xi_0, \xi_1, \ldots, \xi_{N-1}$, а у объектов второго класса — через $\eta_0, \eta_1, \ldots, \eta_{N-1}$. При попарном сравнении объектов из класса K_1 количество сравнимаемых пар (i,j) можно представить в виде матрицы внутриклассовых расстояний первого класса

$A_1 = \{a_1^{ij}\}$, где $a_1^{ij} = \xi_i\xi_j,\ i,j = 0,\dots,N-1,\ i \neq j;\ a_1^{ii} = \xi_i(\xi_i - 1)/2,$ $i = 0,\dots,N-1$. Для объектов из второго класса матрица внутриклассовых расстояний $A_2 = \{a_2^{ij}\}$, где $a_2^{ij} = \eta_i\eta_j,\ i,j = 0,\dots,N-1,\ i \neq j;$ $a_2^{ii} = \eta_i(\eta_i - 1)/2,\ i = 0,\dots,N-1$. Аналогично, матрица межклассовых расстояний $B = \{b^{ij}\}$, где $b^{ij} = \eta_i\xi_j + \xi_i\eta_j,\ i,j = 0,\dots,N-1,\ i \neq j;$ $b^{ii} = \eta_i\xi_i,\ i = 0,\dots,N-1$.

Матрица расстояний $\{c_{ij}\}$ в пространстве $\tilde{N}$ определяется $N(N-1)/2$ числами. Поэтому её можно представить вектором $x = (x_1,\dots,x_{N(N-1)/2})$, где $x_k = c_{ij}$, $k = (2N-1-i)i/2 + j - i,\ i \leqslant j$.

Критерий оптимизации задачи (1) запишется в виде следующей задачи целочисленного линейного программирования [2]:

$$\begin{cases} \sum\limits_{k=1}^{N(N-1)/2} \gamma_k x_k \to \max\limits_{\{x_k\}}; \\ 1 \leqslant x_k \leqslant N-1, \quad k = 1,\dots,N(N-1)/2; \\ x_k \text{— целые и удовлетворяют отношению порядка;} \end{cases} \tag{2}$$

где $\gamma_k = \frac{1}{M}\left(\xi_i\eta_j + \eta_i\xi_j\right) - \frac{0.5\lambda}{N_1}\xi_i\xi_j - \frac{0.5\lambda}{N_2}\eta_i\eta_j,\ k = (2N-1-i)i/2 + j - i,$ $i \leqslant j$ — соответствующие коэффициенты.

Следующая теорема сильно упрощает решение задачи.

Теорема 1. *Решением оптимизационной задачи*

$$\begin{cases} \sum\limits_{k=1}^{N(N-1)/2} \gamma_k x_k \to \max\limits_{\{x_k\}}; \\ X_{\min} \leqslant x_k \leqslant X_{\max}, \quad k = 1,\dots,N(N-1)/2; \\ x_k \text{ — действительные и удовлетворяют отношению порядка;} \end{cases} \tag{3}$$

могут являться только векторы $b = (b_1,\dots,b_{\frac{N(N-1)}{2}})$, *в которых* $b_k = X_{\min}$ *или* $b_k = X_{\max},\ k = 1,\dots,\frac{N(N-1)}{2}$.

Теорема верна и для целочисленных $x_k,\ k = 1,\dots,N(N-1)/2$. Верно и обратное — решение задачи целочисленного линейного программирования будет также решением задачи линейного программирования в непрерывном случае с теми же ограничениями на x_k. Поэтому вначале можно решить задачу линейного программирования в бинарном случае, а затем элементарным линейным преобразованием координат получить решение исходной задачи.

Теорема 2. *Число матриц расстояний размерности N можно представить следующей рекуррентной формулой:*

$$f(N) = \sum_{i=0}^{N-1} f(i)f(N-1-i);$$
$$f(0) = 1, \quad f(1) = 1.$$

Из данной рекуррентной формулы можно получить явное выражение для числа допустимых матриц $f(N) = \frac{C_{2N}^{N}}{N+1}$ [3].

Так как алгоритмы решения задач целочисленного линейного программирования имеют сложность $O(2^{N^2})$ в худшем случае, теорема (2) позволяет уменьшить сложность задачи по крайней мере до $O(2^{(NlogN)/2})$, даже если решать задачу простым перебором допустимых метрик.

Работа выполнена при поддержке РФФИ, проект № 05-01-00332.

Литература

[1] *Журавлев Ю. И.* Избранные научные труды— М.: Магистр, 1998— 420 с.

[2] *Галеев Э. М., Тихомиров В. М.* Краткий курс теории экстремальных задач. — М.: Изд. Московского университета, 1989. — 204 с.

[3] *Садовничий В., Григорьян А., Конягин С.* Задачи студенческих математических олимпиад МГУ. — М.: Изд. Московского университета, 1987. — 310 с.

[4] *Иофина Г. В.* Выбор наилучшей метрики в алгоритме распознавания по ближайшему соседу // Труды 49-й научной конференции МФТИ, Москва-Долгопрудный, 2006 — С. 266–267.

Алгоритмы распознавания, основанные на оценке взаимосвязанности признаков

Камилов М. М., Фазылов Ш. Х., Мирзаев Н. М.

kamilov@yandex.ru, shavkat-faz@mail.ru, mnm2005@rambler.ru

Ташкент, Институт математики и информационных технологий АН Руз

В данной работе рассматривается вопрос построения алгоритмов распознавания на основе анализа взаимосвязанности признаков, которые являются логическим продолжением работ Ю. И. Журавлева и его учеников [1–4]. Корректность, и другие не приведенные здесь определения понятий, а также обозначения, можно найти в [1, 2]. Задание этих алгоритмов включает следующие основные этапы.

1. Задание меры парных связей между признаками. Пусть S — совокупность допустимых объектов. В пространстве признаков $X = (x_1, \ldots, x_n)$ любому объекту $S \in \{S\}$ соответствует вектор $\bar{a} = (a_1, \ldots, a_n)$. Каждый признак x_i имеет своё множество допустимых

значений D_i, $i = 1, \ldots, n$. Введём функцию $\mu(x_i, x_j)$, характеризующую силу парной связи между признаками x_i и x_j, и удовлетворяющую условиям:

1) $\mu(x_i, x_j) \geqslant 0$;
2) $\mu(x_i, x_j) = \mu(x_j, x_i)$;
3) $\mu(x_i, x_i) > \mu(x_i, x_j), \quad i \neq j$.

В зависимости от выбора $\mu(x_i, x_j)$ можно получить разнообразные алгоритмы определения взаимозависимости признаков.

2. Определение «независимых» множеств сильносвязанных признаков. Пусть A_k, $k = 1, \ldots, k_0$ — множества сильносвязанных признаков. Меру близости $L(A_p, A_q)$ можно задать различными способами, например:

$$L(A_p, A_q) = \frac{1}{N_p N_q} \sum_{x_i \in A_p} \sum_{x_j \in A_q} \mu(x_i, x_j),$$

где N_p, N_q — число признаков, входящих, соответственно, в множества A_p, A_q. В зависимости от способа задания меры близости $L(A_p, A_q)$ между A_p и A_q можно получить разнообразные алгоритмы выделения независимых групп сильносвязанных признаков.

3. Определение моделей функциональной зависимости в каждой группе признаков для каждого класса K_j, $j = 1, \ldots, l$. Пусть x_i — произвольный признак из группы A_q, и $x_{i_0} = \arg\max\limits_{x_i} \sum\limits_{x_i \in A_q} \mu(x_i, x_j)$. Тогда модели функциональной зависимости в A_q зададим в виде

$$x_{i_0} = F(\bar{c}, \bar{y}), \quad \bar{y} \in A_q \setminus \{x_{i_0}\},$$

где $\bar{c}$ — вектор неизвестных параметров, F — функция из некоторого заданного класса $\{F\}$.

Вычисленные значения вектора неизвестных параметров $\bar{c}$ определяют модель функциональной зависимости. В зависимости от задания параметрического вида $F(\bar{c}, \bar{x})$ и метода определения $\bar{c}$ получим разнообразные алгоритмы распознавания.

4. Определение элементарных пороговых правил принятия решений. Сформулируем элементарные пороговые правила принятия решений, основанные на задании порогов. Они характеризуют допустимые отклонения в рассматриваемой модели функциональной зависимости. Для простоты рассмотрим два класса объектов. Обозначим через K_j и $\complement K_j$ множества объектов, данных для обучения.

Определим множество A_q всех функциональных моделей во множестве признаков A_q. Обозначим через δ_i все элементарные пороговые пра-

вила принятия решений в A_q:

$$\delta_i(K_j, S) = \begin{cases} 1, & \text{если } \left|x_{i0} - F(\bar{c}, x_i)\right| < \varepsilon_i; \\ 0, & \text{в противном случае.} \end{cases}$$

где ε_i — заданный порог близости в рамках модели $x_{i_0} = F(\bar{c}, x_i)$.

5. Задание функции близости $B_q(K_j, S)$ между K_j и S по множеству признаков A_q. Функцию близости $B_q(K_j, S)$ можно задавать различными способами, например:

$$B_q(K_j, S) = \sum_{i=1}^{N_q} \rho_i \delta_i(K_j, S),$$

где N_q — число используемых элементарных проговых правил. ρ_i — параметр алгоритма.

6. Оценка принадлежности объекта к классу. Оценка принадлежности объекта к классу K_j, $j = 1, \ldots, l$, вычисляется оператором $B(S) = (\eta_1(S), \ldots, \eta_l(S))$, где

$$\eta_j(S) = \sum_{u=1}^{k_0} \gamma_u B_u(K_j, S),$$

γ_u — параметр алгоритма, $u = 1, \ldots, k_0$; k_0 — число независимых множеств.

7. Решающее правило. Решение принимается поэлементно [1], т. е.

$$\beta_i = \begin{cases} 0, & \text{если } \eta_j(S_i) < c_1; \\ 1, & \text{если } \eta_j(S_i) > c_2; \\ \Delta, & \text{если} c_1 \leqslant \eta_j(S_i) \leqslant c_2. \end{cases}$$

Нами определены параметрические модели распознающих алгоритмов, основанные на оценке взаимосвязанности признаков. Всякий алгоритм A из этой модели полностью определяется заданием набора параметров $\pi = \langle \bar{c}, \{\varepsilon_i\}, N_q, \{\rho_i\}, k_0, \{\gamma_u\} \rangle$. Множество всех алгоритмов распознавания из рассмотренной модели обозначим через $A(\pi, S)$. Очевидно, что существенным недостатком таких эвристических алгоритмов распознавания, проверяемых лишь на некотором количестве практических задач, является то, что алгоритмы, оптимальные для решения одной задачи из заданного класса, не всегда оптимальны (или приемлемы) для решения другой задачи того же класса. Поэтому возникает задача исследования корректности рассмотренных алгоритмов. В связи с этим доказана следующая

Теорема 1. *Пусть множество* $\{Z\}$ *удовлетворяет условиям:*
1) объекты, изоморфные относительно J_0*, отсутствуют в* $\tilde{S}^q$*;*
2) оператор $B(S)$ *ограничен;*
3) $J_0 \cap \tilde{S}^q = \varnothing$*.*

Тогда в рамках алгебраического замыкания расмотренных алгоритмов существует корректный алгоритм A^* для задачи $Z \in \{Z\}$.

Литература

[1] *Журавлев Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов // Кибернетика. — 1977. № 4. — С.14–20.

[2] *Журавлев Ю. И., Исаев И. В.* Построение алгоритмов распознавания, корректных для заданной контрольной выборки // ЖВМиМФ. — Москва, 1979. — Т. 19. — № 3. — С. 726–738.

[3] *Журавлев Ю. И., Камилов М. М., Туляганов Ш. Е.* Алгоритмы вычисления оценок и их применение. — Ташкент: Фан, 1974. — 119 с.

[4] *Фазылов Ш. Х., Мирзаев Н. М., Жуманазаров С. С.* О корректности алгоритмов распознавания, основанных на взаимосвязанности между признаками // «Проблемы информатики и энергетики». — Ташкент, 1997. — № 1. — С. 19–24.

Статистический подход к оцениванию зависимых признаков в интеллектуальных системах

Колесникова С. И., Янковская А. Е.
skolesnikova@yandex.ru, yank@tsuab.ru
Томск, Томский госуниверситет систем управления и радиоэлектроники,
Томский архитектурно-строительный университет

Приводятся два метода определения весовых коэффициентов зависимых признаков, используемых в интеллектуальных системах поддержки принятия решений, основанных на теоретико-информационном понятии энтропии и методе главных компонент.

Введение

Одной из наиболее важных проблем при создании интеллектуальных систем выявления закономерностей и поддержки принятия решений является проблема анализа признакового пространства на предмет выделения наиболее значимых признаков и оценивания величины их значимости [1, 2, 4] в виде весовых коэффициентов признаков (ВКП), используемых при принятии решений в интеллектуальных тестовых распознающих системах с матричным представлением данных и знаний [2, 3].

Матрица описаний Q и матрица различений R задают описание объектов в пространстве характеристических признаков объектов и в про-

странстве классификационных признаков, соответственно. Элемент q_{ij} матрицы Q задает значение j-го признака для i-го объекта. Множество всех неповторяющихся строк матрицы R сопоставлено множеству выделенных образов.

Задача распознавания состоит в определении по матрицам Q и R образа, которому принадлежит заданный совокупностью признаков исследуемый объект, как правило, не входящий в обучающую выборку.

Признаки называются зависимыми, если имеется хотя бы одна пара объектов из разных образов, ими различаемая.

Совокупность признаков, различающих все пары объектов из разных образов (классов), назовем диагностическим тестом (далее по тексту просто тестом).

Строки матрицы тестов T соответствуют тестам, а столбцы — признакам Z, каждый из которых содержится хотя бы в одном тесте. Два объекта считаются различимыми (при $q_{ij} = 0, 1, \Delta$, где $q_{ij} = \Delta$ означает, что признак может принимать как нулевые, так и единичные значения), если хотя бы один признак в описании одного из них принимают значение 1 (0), а в описании другого — инверсное значение 0 (1).

Два предложенных метода базируются на представлении совокупности всех различимых пар объектов из разных образов для каждого признака в виде мультимножества [4, 5], использовании матриц парных сравнений признаков на основе специальным образом выбранных мер относительной важности признаков, учитывающих их взаимозависимость [4]: $\frac{\delta(|P_i - P_j|)}{\delta(|P_j - P_i|)}, \frac{\delta(/P_i - P_j/)}{\delta(/P_j - P_i/)}$, где $|P_i|$ — мощность i-го мультимножества, соответствующего признаку z_i ; $/P_i/$ — размерность (количество элементов, встречающихся один раз) i-го мультимножества; $P_i - P_j$ — разность мультимножеств, соответствующих признакам z_i и z_j; $\delta(x) = x$, если $x \neq 0$, и $\delta(x) = 1$ иначе.

Постановка задачи

Пусть по матрицам Q и R построены все (или некоторые) безусловные безызбыточные диагностические тесты, представленные матрицей тестов , строки которой сопоставлены тестам, а столбцы — характеристическим признакам, и определено число различающих пар «образ–образ» по каждому (характеристическому) признаку. Требуется определить весовые коэффициенты признаков, входящих в объединение всех (или некоторых) диагностических тестов [2, 5] с учетом их зависимости.

Метод определения ВКП на основе главных компонент

Исходными данными для метода на основе главных компонент является матрица описаний Q, содержащая информацию о частотах встречаемости признаков или вероятностях «проявления» признака для каждо-

го из образов (классов). Под «проявлением» признака здесь понимается различение этим признаком пар типа «образ–образ», образущих соответствующее мультимножество.

Перечислим кратко основные этапы метода:

1) вычисление матрицы ковариации признаков;
2) вычисление собственных значений и собственных векторов матрицы ковариации признаков;
3) вычисление вектора главных компонент;
4) определение относительных дисперсий каждой главной компоненты;
5) определение первых главных компонент, обеспечивающих достаточную вариабельность признаков;
6) интерпретация главных компонент для конкретной задачи.

Метод на основе теоретико-информационного понятия энтропии

Предлагаемый метод включает следующие основные шаги:

1) отдельному тесту $T_1 = (z_1, \ldots, z_M)$, где M — количество признаков в тесте, сопоставляется объединение мультимножеств $\mathbf{P}_{T_1} = \bigcup_{j=1}^{M} P_j$, порожденных признаками, входящих в тест T_1;
2) мультимножество $\mathbf{P}_{T_1}$ представляется в виде объединения попарно непересекающихся мультимножеств $\mathbf{P}_{T_1} = \bigcup_{j=1}^{N} P_j,\ N = 2^M - 1$;
3) определяется дискретное распределение вероятностей $Pr(T_1)$ «проявления» совокупности признаков в тесте и распределение вероятностей «проявления» каждого признака $Pr(z_i)$, входящего в тест, на подмножествах $S_1, \ldots, S_N$;
4) вычисляется содержащаяся информация (энтропия) в тесте $I(T_1)$ и в признаке $I(z_i)$;
5) определяются значения ВКП w_i и теста w_{T_1}, в качестве которых принимаются величины $w_i = I_0 - I(z_i),\ w_{T_1} = I_0 - I(T_1)$, соответственно, где $\log_2 K$ — исходная неопределенность относительно образов, K — количество образов.

Заключение

Отметим, что в реальных данных зависимость между признаками наблюдается очень часто, и в этом случае оценками их индивидуальной «информативности» руководствоваться некорректно. Предложенные методы позволяют: во-первых, представить «вес» теста не в виде суммы ВКП признаков, что не корректно [3, 5] в силу возможной взаимозависимости этих признаков, а в виде суммы «весов» непересекающихся мультимножеств, составляющих тест; во-вторых, учитывать пары «образ–образ» («объект–образ») без дополнительного дублирования (в силу пе-

ресечения соответствующих мультимножеств), что вносило искажение в значение ВКП и, соответственно, неточность в решающее правило.

Работа выполнена при поддержке РФФИ, проект №07-01-00452, и РГНФ, проект №06-06-12603в.

Литература

[1] *Журавлев Ю. И., Гуревич И. Б.* Распознавание образов и анализ изображений // Искусственный интеллект в 3-х кн. Кн 2. Модели и методы: Справочник под ред. Д. А. Поспелова. Москва: Радио и связь, 2005. — С. 149–190.

[2] *Янковская А. Е.* Логические тесты и средства когнитивной графики в интеллектуальной системе // Новые информационные технологии в исследовании дискретных структур. Доклады 3-ей Всероссийской конф. с междунар. участием. Томск: Изд-во СО РАН. 2000. — С. 163–168.

[3] *Yankovskaya A. E.* Test Pattern Recognition with the Use of Genetic Algorithms // Patt. Recog. and Image Anal. 1999. — Vol. 9.— No. 1. — P. 121–123.

[4] *Петровский А. Б.* Упорядочивание и классификация объектов с противоречивыми признаками // Новости искусственного интеллекта. — 2003. — №4. — С. 34–43.

[5] *Янковская А. Е., Колесникова С. И.* О применении мультимножеств к задаче вычисления весовых коэффициентов признаков в интеллектуальных распознающих системах // Искусственный интеллект, Украина. Донецк: «Наука і освіта», 2004. — №2. — С. 216–220.

Синдромальные процедуры распознавания для исследования фазового пространства конкретных многомерных динамических систем

Котельников И. В.

neymark@pmk.unn.ru

Нижний Новгород, НИИПМК ННГУ

Общая концепция компьютерной статистической модели исследования конкретных многомерных динамических систем (ДС) с применением методов распознавания образов предложена в [1]. Основным аргументом в пользу предлагаемой концепции выступает положение о том, что методы распознавания образов слабо зависят от размерности исследуемых объектов, в то время как на пути классических методов исследования непреодолимой стеной встает «проклятие размерности» уже при $n \geqslant 4$, где n — размерность ДС. Ниже под ДС подразумеваются ДС порядка $n \geqslant 4$, задаваемые системой обыкновенных дифференциальных уравнений. Рассматриваются вопросы исследования фазового пространства ДС при фиксированных параметрах. Входной информацией метода является выборка траекторий ДС в ограниченном фазовом пространстве.

Начальные условия формируются случайным образом с помощью стандартных программ равномерного распределения. Траектории строятся с помощью известных технологий, имеющихся, например, в программных средах MATLAB, MAPLE. Траектории строятся с постоянным шагом интегрирования и разделяются на две равных по числу точек части. Первые части траекторий образуют траектории области притяжения аттрактора, вторые — траектории самого аттрактора.

Решающее правило для отдельного аттрактора

Входной информацией для построения решающего правила (РП) является выборка траекторий аттрактора одного и того же типа (состояние равновесия, предельный цикл, хаотический аттрактор) [2]. Построение РП производится следующей процедурой. Первая точка траектории аттрактора принимается за центр, а вторая за вершину n-мерного параллелепипеда синдрома и строится такой синдром. Затем вторая точка принимается за центр, а третья за вершину для построения второго синдрома, и так для всех последовательных точек траектории. Объектом, поступающим на вход РП, является траектория аттрактора. Если более заданного значения P процентов точек траектории находится внутри или на поверхности синдромов РП, траектория относится к аттрактору, для которого строится РП. Траектория относится к другому аттрактору, если ни одна из точек траектории не принадлежит синдромам РП. Алгоритм адаптивный. Если траектория относится к аттрактору РП, то на цепочках точек траектории, которые не попали в синдромы РП, аналогичным образом строятся цепочки синдромов, присоединяемые к имеющемуся РП. РП, помимо своего прямого назначения, применяется для разделения выборки траекторий на выборки траекторий различных аттракторов, а также, для существенного сжатия информации об этих выборках. Множество точек траекторий, образующих рассматриваемое РП, используется позднее как обучающая выборка (ОВ) класса конкретного аттрактора при построении разделяющего РП аттракторов ДС.

Разделяющее решающее правило аттракторов

Построение разделяющего РП аттракторов производится в полном соответствии с [3] с той лишь разницей, что последним этапом является этап преобразования всех его k-мерных синдромов в n-мерные. Это необходимо из-за ограниченности изучаемого фазового пространства, за пределами которого располагается безграничная область траекторий ДС, которые не представлены в ОВ. Алгоритм принятия решения по этому РП претерпевает тоже существенные изменения. На вход РП в качестве объекта подается не отдельный n-мерный объект, а траектория. РП определяет класс каждой точки траектории, а затем подводит общий итог по

всем классам РП. Получается РП, построенное на принципе голосования. Траектория относится к тому конкретному аттрактору, за который проголосовало большинство точек. Благодаря предварительному отбору при формировании ОВ это большинство, как правило, оказывается подавляющим. Алгоритм адаптивный. После определения класса траектории все неправильно классифицированные точки траектории дополняются в ОВ с признаком своего класса. После небольшого числа шагов адаптации все траектории ОВ классифицируются 100-процентным большинством. Процедура адаптации особенно необходима для построения областей притяжения аттракторов, что должно быть сделано с очень малой наперед заданной вероятностью ошибки.

Построение областей притяжения аттракторов

Ясно, что построение областей притяжения производится на первых частях траекторий ОВ, или на траекториях областей притяжения. Для них, по существу, делается все то же, что и для траекторий аттракторов. Строится РП одного аттрактора для каждого из аттракторов по алгоритму первого раздела. Однако это производится с единственной целью возможно большего сжатия информации, поскольку остальные вопросы уже решены на траекториях аттракторов. Строится разделяющее РП предыдущего раздела со 100-процентной адаптацией для получения оптимальных синдромов для траекторий области притяжения. Из полученного РП выбирается синдром с наименьшей вероятностью ошибки того класса, для которого строится область притяжения. Затем формируется контрольная выборка (КВ) траекторий выбранного класса с начальными условиями внутри или на поверхности выбранного синдрома. Полученная КВ пропускается через разделяющее РП областей притяжения. Вполне вероятно, что полученная вероятность ошибки будет меньше заданного значения. Тогда следует перейти для построения аналогичным образом области притяжения для очередного аттрактора. Если же это не так, то та же КВ вторично подается на вход РП, с той же процедурой принятия решения, но со следующей дополнительной функцией. Из всех точек траекторий КВ, находящихся внутри выбранного синдрома, РП формирует новую ОВ в соответствии с принятым по этим точкам решением. Получаем ОВ, расположенную внутри и на поверхности выбранного синдрома. На этой ОВ строится новое разделяющее РП для области притяжения, выбирается новый синдром, формируется новая КВ из начальных условий внутри выбранного синдрома. Важно отметить, что новый синдром будет находиться всегда внутри первого, т. к. именно на его точках построено новое РП. Построением нового РП из первого синдрома будут удалены те области, начальные условия в которых ведут к чужому аттрактору. Как видно, процедура повторяется с перено-

сом области построения РП в область выбранного синдрома. Указанная процедура продолжается в цикле до тех пор, пока не будет достигнута допустимая вероятность ошибки для выделенной области притяжения.

Предлагаемая методика апробирована на 7-мерной ДС турбулентности с двумя аттракторами цикла и двумя хаотическими аттракторами при заданных параметрах. Полученные результаты полностью подтверждают справедливость и эффективность предложенной в [1] концепции.

Работа выполнена при поддержке РФФИ, проект № 05-01-00391.

Литература

[1] *Неймарк Ю. И.* Компьютерная концепция исследования конкретных динамических систем // 7 Всеросс. конф. «Нелинейные колебания механических систем», Нижний Новгород: Издательство ННГУ, 2005. — С. 17–18.

[2] *Котельников И. В.* Определение типа фазовых траекторий динамических систем на основе оптимальных тупиковых нечетких тестов и синдромов // Всеросс. конф. ММРО-12. — Москва: Макс Пресс, 2005. — С. 144–147.

[3] *Kotel'nikov I. V.* A Sindrome Recognition Method Based on Optimal Irreducible Fuzzy Tests // Patt. Rec. and Image Anal. — 2001. — V. 11, № 3. — Pp. 553–559.

Адаптивный нестационарный регрессионный анализ

Красоткина О. В., Моттль В. В., Марков М. Р., Мучник И. Б.

krasotkina@uic.tula.ru

Тула, ТулГУ; Москва, ВЦ РАН;

США, Markov Processes International; США, университет Rutgers

Задача восстановления числовой регрессионной зависимости $y_t\colon T \to \to \mathbb{R}$ в некотором множестве наблюдений $t \in T$ является одной из ключевых в интеллектуальном анализе данных. В качестве наиболее проблематичного аспекта этой задачи обычно рассматривается выбор наиболее подходящего подмножества $\hat{I} \subseteq I$ в множестве доступных числовых признаков объектов $\left\{x_t^{(i)}, i \in I\right\}$, относительно которых искомая зависимость ищется, чаще всего, в классе линейных моделей.

Для многих приложений типичны ситуации, когда совокупность наблюдений $t \in T$ естественно рассматривать как упорядоченную последовательность $T = \{1, \ldots, N\}$, допуская, что коэффициенты регрессии могут принимать разные значения в разные моменты времени:

$$y_t = \sum\nolimits_{i \in \hat{I}} \beta_t^{(i)} x_t^{(i)} + e_t, \quad t = 1, \ldots, N, \tag{1}$$

где e_t — последовательность независимых ошибок наблюдения с нулевым средним. Принципиальная специфика модели нестационарной регрессии

заключается в том, что число переменных $(\beta_t^{(i)},\ i \in \hat{I},\ t = 1,\ldots,N)$, подлежащих оцениванию, всегда во много раз превышает число наблюдений N. В результате оказывается, что скрытые коэффициенты регрессии невозможно оценить без принятия некоторых априорных предположений об их последовательности, т. е. без дополнительной регуляризации задачи.

Задача оценивания модели нестационарной регрессии (1) интенсивно изучалась в мировой литературе. Популярным инструментом ее решения является метод FLS — Flexible Least Squares [1]:

$$J\big(\beta_t^{(i)},\ t = 1\ldots,N,\ i \in \hat{I} \mid \hat{I}, \rho\big) = \\ = \sum_{t=1}^{N}\Big(y_t - \sum_{i\in\hat{I}} \beta_t^{(i)} x_t^{(i)}\Big)^2 + \rho \sum_{t=2}^{N}\sum_{i\in\hat{I}}\Big(\beta_t^{(i)} - \beta_{t-1}^{(i)}\Big)^2 \to \min\,. \quad (2)$$

Подмножество регрессоров $\hat{I} \subseteq I$ и коэффициент $\rho > 0$ являются параметрами критерия. Здесь первое слагаемое отвечает за аппроксимацию наблюдений, а второе слагаемое регулирует изменчивость искомых коэффициентов регрессии во времени. Чем больше ρ, тем более плавной будет последовательность оценок, что уменьшает фактическую «размерность» задачи, делая ее промежуточной между $|\hat{I}|$ и $N|\hat{I}|$. При $\rho \to \infty$ критерий сводится к обычному методу наименьших квадратов $\hat{\beta}_1^{(i)} = \ldots = \hat{\beta}_N^{(i)}$.

Если рассматривать априорную модель последовательности коэффициентов регрессии как совокупность независимых скрытых случайных процессов $\beta_t^{(i)} = \beta_{t-1}^{(i)} + \xi_t^{(i)}$, каждый из которых порождается нормальным белым шумом $\xi_t^{(i)}$, дисперсия которого в ρ раз меньше дисперсии шума e_t в модели наблюдения (1), то критерий FLS (2) максимизирует апостериорную плотность распределения вероятности на множестве реализаций скрытого случайного процесса.

Для выбранного множества регрессоров $\hat{I}$ и фиксированного значения коэффициента сглаживания ρ минимизация квадратичной функции (2) сводится к решению, вообще говоря, очень большой системы линейных уравнений относительно $N|\hat{I}|$ переменных, имеющей однако блочно-трехдиагональную матрицу с блоками $|\hat{I}|\times|\hat{I}|$. Эта особенность допускает применение метода прогонки, обеспечивающего линейную вычислительную сложность решения системы относительно длины временного ряда N. Методу прогонки эквиваленты квадратичное динамическое программирование [2, 3] и фильтр-интерполятор Калмана-Бьюси [4].

Как правило, эффективное подмножество регрессоров $\hat{I} \subset I$ выбрать априори невозможно, и уже хотя бы поэтому задачу нестационарного регрессионного анализа следует рассматривать как задачу интеллекту-

ального анализа данных. Более того, нестационарность модели порождает проблему, совершенно новую для интеллектуального анализа данных, а именно, неизбежную необходимость выбирать степень изменчивости во времени коэффициентов регрессии. Часто по физической природе исследуемого явления нестационарными являются коэффициенты лишь при некоторых регрессорах, что порождает также проблему разделения множества регрессоров на стационарные и нестационарные.

В данной работе мы рассматриваем модификацию критерия FLS (2), позволяющую, во-первых, автоматически выбирать подмножество эффективных регрессоров, во-вторых, автоматически отбирать, далее, еще меньшее подмножество нестационарных регрессоров, и, в-третьих, определять для них индивидуальные коэффициенты сглаживания последовательности коэффициентов регрессии. В отличие от (2), адаптивный критерий применяется к полному множеству доступных регрессоров I:

$$
\begin{aligned}
&J(\beta_t^{(i)},\ t=1\ldots,N,\ \delta^{(i)},\ \lambda^{(i)},\ i\in I \mid \rho) = \\
&= \sum_{t=1}^{N}\Big(y_t - \sum_{i\in I}\beta_t^{(i)}x_t^{(i)}\Big)^2 + \rho\sum_{t=2}^{N}\sum_{i\in I}\tfrac{\delta^{(i)}+\lambda^{(i)}}{\delta^{(i)}\lambda^{(i)}}\Big(\beta_t^{(i)} - \tfrac{\delta^{(i)}}{\delta^{(i)}+\lambda^{(i)}}\beta_{t-1}^{(i)}\Big)^2 \to \min, \\
&\prod_{i\in I}\tfrac{\delta^{(i)}\lambda^{(i)}}{\delta^{(i)}+\lambda^{(i)}} = 1, \quad \text{т. е.} \quad \sum_{i\in I}\log\tfrac{\delta^{(i)}\lambda^{(i)}}{\delta^{(i)}+\lambda^{(i)}} = 0.
\end{aligned}
\tag{3}
$$

Функцию адаптации выполняют вспомогательные переменные $\delta^{(i)} \geqslant 0$ и $\lambda^{(i)} \geqslant 0$, определяющие априорную модель случайных последовательностей коэффициентов регрессии, порождаемых независимыми реализациями нормального белого шума с нулевым средним значением:

$$
\beta_t^{(i)} = \tfrac{\delta^{(i)}}{\delta^{(i)}+\lambda^{(i)}}\beta_{t-1}^{(i)} + \xi_t^{(i)}, \quad \mathsf{E}(\xi_t^{(i)})^2 = \tfrac{\delta^{(i)}\lambda^{(i)}}{\delta^{(i)}+\lambda^{(i)}}. \tag{4}
$$

Если $\lambda^{(i)} \to 0$, то $\mathsf{E}\big(\xi_t^{(i)}\big)^2 \to 0$, и последовательность коэффициентов при i-м регрессоре всегда будет оставаться постоянной $\beta_t^{(i)} = \beta_{t-1}^{(i)}$ с некоторым априори неизвестным значением. Если же $\delta^{(i)} \to 0$, то $\mathsf{E}\big(\xi_t^{(i)}\big)^2 \to 0$ вместе с $\frac{\delta^{(i)}}{\delta^{(i)}+\lambda^{(i)}} \to 0$, и i-я последовательность коэффициентов превращается в нулевую константу. Совокупность ненулевых переменных $\delta^{(i)}$ образует множество активных регрессоров $\hat{I} = \{i : \delta^{(i)} > 0\} \subseteq I$, а ненулевые переменные $\lambda^{(i)}$ выделяют среди них подмножество регрессоров с нестационарными коэффициентами $\hat{I}_{\text{var}} = \{i : \lambda^{(i)} > 0\} \subseteq \hat{I}$.

Произведение дисперсий возмущающего шума $\mathsf{E}\big(\xi_t^{(i)}\big)^2$ по всем регрессорам определяет объем эллипсоида рассеяния случайного вектора $(\beta_t^{(i)}, i \in I)$ вокруг его условного математического ожидания (4) — чем меньше этот объем, тем интенсивнее подавляется общее отклонение

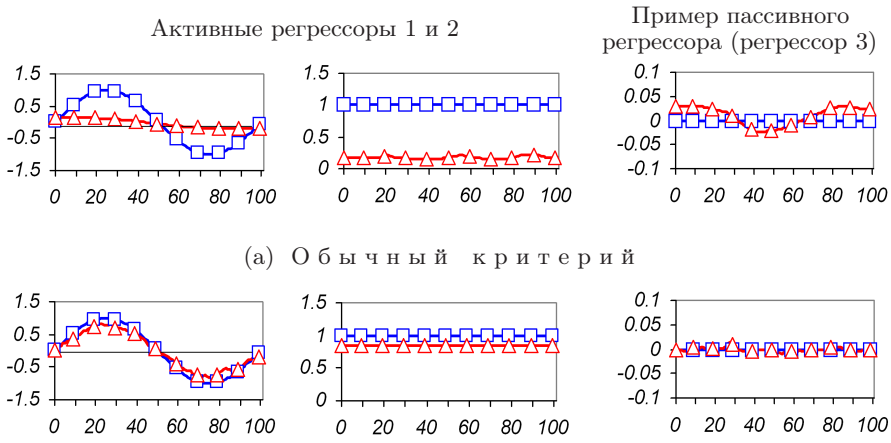
всех коэффициентов регрессии как друг от друга во времени, так и от нуля. Роль переменных $\delta^{(i)}$ и $\lambda^{(i)}$ заключается в управлении соотношением между уровнями вариабельности коэффициентов при разных регрессорах, а не их общей вариабельностью в нестационарной модели, поэтому объем эллипсоида рассеяния зафиксирован в критерии (3) ограничением-равенством $\prod_{i\in I}\frac{\delta^{(i)}\lambda^{(i)}}{\delta^{(i)}+\lambda^{(i)}}=1$.

Эту роль берет на себя параметр ρ адаптивного критерия (3), не входящий в число варьируемых переменных и показывающий, во сколько раз дисперсия шума наблюдения $\mathsf{E}\big(e_t^{(i)}\big)^2$ в (1) предполагается большей, чем среднегеометрическое значение дисперсий шума в компонентах модели вариабельности коэффициентов регрессии (4).

При таких предположениях минимум критерия (2) соответствует максимуму апостериорной плотности распределения скрытой последовательности коэффициентов регрессии относительно заданного временного ряда $(y_t, x_t^{(i)},\ i\in I)$. Точку минимума легко найти, применяя итерационный алгоритм Гаусса-Зайделя к двум группам переменных $(\delta^{(i)},\lambda^{(i)},\ i\in I)$ и $(\beta_t^{(i)},\ i\in I,\ t=1,\ldots,N)$, начиная с некоторых значений $\delta^{(i),0}$ и $\lambda^{(i),0}$, удовлетворяющих ограничению в (2).

Заметим, что при фиксированных значениях $\delta^{(i)}$ и $\lambda^{(i)}$, в частности, $\delta^{(i),k}$ и $\lambda^{(i),k}$ на k-й итерации, критерий является квадратичной функцией блочно-трехдиагональной структуры относительно последовательности векторных переменных $(\beta_t^{(i)},\ i\in I),\ \ t=1,\ldots,N$, и его минимизация осуществляется эквивалентными методами прогонки, квадратичного динамического программирования, либо фильтрации-интерполяции Калмана-Бьюси [1, 2, 3, 4] за время, пропорциональное длине временного ряда N. Нетрудно доказать, что после того, как последовательность $(\beta_t^{(i),k},\ i\in I,\ t=1,\ldots,N)$ найдена, очередные значения $\delta^{(i),k+1}$ и $\lambda^{(i),k+1}$, минимизирующие (3), вычисляются по формулам, в которых $a^{(i)}=\frac{\delta^{(i)}}{\delta^{(i)}+\lambda^{(i)}}$ и $0\lessapprox a_0<a_1\lessapprox 1$:

$$
\begin{aligned}
0<a^{(i),k+1}&=\left(a_0,\frac{\sum_{t=2}^N\beta_{t-1}^{(i),k}\beta_t^{(i),k}}{\sum_{t=2}^N\big(\beta_{t-1}^{(i),k}\big)^2},a_1\right)<1;\\
\delta^{(i),k+1}&=\frac{a^{(i),k+1}}{1-a^{(i),k+1}}\lambda^{(i),k+1};\\
\lambda^{(i),k+1}&=\frac{1}{a^{(i),k+1}}\frac{\sum_{t=1}^N\big(\beta_t^{(i),k}-a^{(i),k+1}\beta_{t-1}^{(i),k}\big)^2}{\left[\prod_{j\in I}\sum_{t=1}^N\big(\beta_t^{(j),k}-a^{(j),k+1}\beta_{t-1}^{(j),k}\big)^2\right]^{1/|I|}},\quad i\in I.
\end{aligned}
$$



(а) О б ы ч н ы й к р и т е р и й

(б) А д а п т и в н ы й к р и т е р и й

Рис. 1. Результаты экспериментов по оцениванию нестационарной регрессии: □ — модельные последовательности коэффициентов, △ — оцененные последовательности.

Итерационный процесс обычно сходится за 10–15 итераций, проявляя явную тенденцию к практическому обнулению вспомогательных переменных, отвечающих за подавление части регрессоров $\delta^{(i),k} \to \hat{\delta}^{(i)} \gtrapprox 0$, при этом переменные, ответственные за нестационарность коэффициентов регрессии, также почти обнуляются $\lambda^{(i),k} \to \hat{\lambda}^{(i)} \gtrapprox 0$. Предельные значения, оставшиеся существенно ненулевыми $\delta^{(i),k} \to \hat{\delta}^{(i)} > 0$, выделяют подмножество эффективных регрессоров $\hat{I} \subset I$. Однако переменные $\lambda^{(i),k}$ стремятся к нулю, как правило, для большего числа регрессоров, подавляя нестационарность коэффициентов регрессии для части регрессоров, выделенных как эффективные $i \in \hat{I}$. Остальные предельные значения $\hat{\lambda}^{(i)} > 0$ указывают подмножество эффективных регрессоров с нестационарными коэффициентами $\hat{I}_{\text{var}} \subseteq \hat{I}$.

Значение параметра ρ не может быть определено путем дополнительной оптимизации критерия (2) и подбирается с помощью процедуры скользящего контроля, особенности использования которой в задаче оценивания нестационарной регрессии рассмотрены в работе [3].

Эффект выделения регрессоров, адекватных анализируемому временному ряду, и среди них регрессоров, входящих в модель с нестационарными коэффициентами, иллюстрирует следующий модельный эксперимент.

Модельный временной ряд $(y_t, x_t^{(i)}, i \in I)$ вида (1) построен как последовательность ста зашумленных линейных комбинаций, $t = 1, \ldots, 100$, ста регрессоров $I = \{1, \ldots, 100\}$, в качестве которых использовались независимые реализации нормального белого шума. Среди ста последовательностей коэффициентов регрессии две приняты отличными от нуля, одна из которых образована одним периодом синусоиды $x_t^{(1)} = \sin\left(\frac{2\pi t}{100}\right)$, а вторая является единичной константой $x_t^{(2)} \equiv 1$, остальные же коэффициенты регрессии тождественно равны нулю $x_t^{(3)} = \ldots = x_t^{(100)} \equiv 0$.

Таким образом, если неизвестно, какие именно регрессоры являются активными, то оцениванию подлежат десять тысяч коэффициентов регрессии при наличии всего лишь ста наблюдений. Естественно, что обычный критерий FLS (2), хотя в нем параметр сглаживания ρ подбирался процедурой скользящего контроля, оказался не способен в этом примере даже приближенно восстановить модель нестационарной регрессии, «размывая» вклад двух активных регрессоров на все сто регрессоров, что хорошо видно на Рис. 1(а). В то же время адаптивный критерий (3), как показывает Рис. 1(б), практически полностью подавляет пассивные регрессоры. Вся нестационарность модели оказывается сконцентрированной в изменении коэффициента только при первом регрессоре, а коэффициент при втором активном регрессоре идентифицирован как стационарный.

Работа выполнена при поддержке РФФИ, проект № 06-01-00412.

Литература

[1] *Kalaba R., Tesfatsion L.* Time-varying linear regression via flexible least squares. — International Journal on Computers and Mathematics with Applications. — Vol. 17. — Pp. 1215–1245.

[2] *Костин А. А., Красоткина О. В., Марков М. Р., Моттль В. В., Мучник И. Б.* Алгоритмы динамического программирования для анализа нестационарных сигналов. — ЖВМиМФ, 2004, — Т. 44, № 1. — С. 70–86.

[3] *Markov M., Krasotkina O., Mottl V., Muchnik I.* Time-varying regression model with unknown time-volatility for nonstationary signal analysis // 8th IASTED Int. Conf. on Signal and Image Processing, Honolulu, USA. — Pp. 14–16.

[4] *Wells C.* The Kalman Filter in Finance. — Kluwer Academic Publishers, 1996.

Алгоритм обобщения, работающий с зашумлёнными данными

Куликов А. В., Фомина М. В.
m_fomina2000@mail.ru
Москва, МЭИ (ТУ)

Обнаружение знаний в базах данных является стремительно увеличивающейся областью, развитие которой вызвано большим интересом к настоятельным практическим, социальным и экономическим нуждам. Современные базы данных содержат так много данных, что практически невозможно вручную проанализировать их для извлечения ценной информации, помогающей принимать важные решения. Отсюда следует, что люди нуждаются в помощи интеллектуальных систем для повышения своих аналитических возможностей.

Шум в обучающей выборке

Индуктивные экспертные системы, обрабатывающие реальные массивы данных, обычно работают в условиях наличия шума во входных данных. Шум возникает из-за таких причин, как, например, некорректное измерение входного параметра, неверное описание значения параметра экспертом, использование испорченных измерительных приборов, потеря данных при пересылке и хранении информации.

Шум вызывает две проблемы: сначала при построении обобщённых правил, а затем при классификации объектов с использованием этих правил.

Мы исследуем две модели шума:

1. Шум связан с исчезновением значений атрибутов.
2. Шум связан с искажением некоторых значений атрибутов в обучающей выборке. При этом истинное значение заменяется на одно из допустимых, но ошибочных значений (значения перемешаны).

Зашумлённые обучающие выборки должны обрабатываться алгоритмом обобщения в соответствии с процедурой «обучения с учителем» Бонгарда [1].

Предсказание неизвестных значений методом ближайшего соседа

Пусть дана выборка с шумом, K', причём искажениям подвергаются атрибуты, принимающие как дискретные, так и непрерывные значения. Рассмотрим проблему использования объектов обучающей выборки K' при построении решающего дерева T и при проведении экзамена с использованием решающего дерева T.

Пусть $X \in K'$ — очередной объект выборки; $X = (x_1, \ldots, x_n)$. Среди всех значений его атрибутов имеются атрибуты со значением N

(Not known). Это могут быть как дискретные, так и непрерывные атрибуты [2].

Наличие неизвестных значений в примерах обучающей выборки затрудняет как обучение, так и экзамен, поскольку часть примеров может быть отвергнута, либо при классификации получен неоднозначный результат. Предлагается восстановить эти неизвестные значения, используя аналог метода «ближайшего соседа» [3, 4].

Основная идея алгоритма в следующем. Если пример X обучающей выборки K' содержит неизвестные значения, определяем на основе введенной в [3, 4] метрики p ближайших к нему примеров, не имеющих неизвестных значений. На основе анализа этих примеров, имеющих максимальное сходство с X, восстанавливаем значения признаков этого объекта.

Рассмотрим стратегию определения неизвестного значения.

1. Признак — количественный. Определяем неизвестное значение как среднее арифметическое значений его ближайших p примеров, для которых определены значения признака.
2. Признак — качественный. Определить неизвестное значение признака как наиболее часто встречающееся среди ближайших p примеров.

Использование процедуры восстановления при построении дерева решений

Рассмотрим возможность использования процедуры восстановления для решения задач индуктивного формирования понятий. Предлагается алгоритм IDTUV (Induction of Decision Tree with restoring Unknown Values), который включает процедуру восстановления неизвестных значений при наличии в обучающей выборке примеров, содержащих шум. Когда неизвестные значения атрибутов восстановлены, используется один из алгоритмов построения деревьев решений. Примеры, для которых не удалось восстановить неизвестные значения, удаляются из обучающей выборки.

Ниже приводится псевдокод алгоритма IDTUV.

Результаты классификации примеров с шумом

Был проведен ряд экспериментов на следующих четырех группах данных из известной коллекции тестовых наборов данных Machine Learning Repository кафедры информатики и вычислительной техники Калифорнийского университета UCI.

Поскольку главной задачей эксперимента было оценить влияние шума на результаты построения классификационных правил и распознавания тестовых примеров, было принято решение отказаться от хаотичного

Алгоритм 1. IDTUV

Вход: $K = K^+ \cup K^-$;
Выход: дерево решений T;
1: получение $K = K^+ \cup K^-$;
2: **для всех** информативных атрибутов K
3: **пока** имеется неизвестное значение атрибута
4: применить алгоритм ВОССТАНОВЛЕНИЕ;
5: **если** информативные атрибуты имеют непрерывные значения **то**
6: применить алгоритм C4.5;
7: **иначе**
8: применить алгоритм ID3;
9: **вернуть** T — дерево решений.

внесения шума в поля любых признаков в тестовых таблицах. Для внесения искажений был использован наиболее информативный признак таблицы, который размещается в корне дерева решений. Были рассмотрены и проанализированы ситуации полного отсутствия шума и наличия шума в 5%, 10% и 20% по выбранному признаку. На каждом тестовом множестве проводился ряд экспериментов. Затем результаты усреднялись.

Влияние шума на построение дерева решений. Первая группа опытов предназначалась для проверки того, как шум в обучающей выборке влияет на построение дерева решений. Опыты проводились по следующей схеме. В обучающую выборку вносится некоторое количество неизвестных значений. Затем производится восстановление таких значений по методу ближайшего соседа. На полученной обучающей выборке вновь строится дерево решений, которое необходимо сравнить с деревом, построенным на основе выборки без шума.

Полученные результаты показали, что в некоторых случаях, даже при отсутствии до 20% значений наиболее информативного признака, не происходило изменения дерева решений, и, следовательно, не изменялись правила классификации (либо эти изменения были крайне незначительны). Это свидетельствует о высокой эффективности метода восстановления.

Влияние шума на классификацию примеров. Второй этап проверки заключался в установлении того, как шум влияет на успешность классификации примеров. Были использованы две модели шума: шум, как отсутствие значений, и шум, заключающийся в перепутывании определенного количества значений в тестовой выборке. В обоих случаях для классификации использовались правила, полученные на обучающей выборке, не содержащей шума. Полученные результаты показали, что алго-

ритм IDTUV в сочетании с алгоритмами восстановления позволяет повысить точность классификации примеров с отсутствующими значениями признаков в 3–4 раза по сравнению с классическими алгоритмами ID3 и C4.5 [5, 6]. При использовании шума типа «перемешивание значений» алгоритм IDTUV повышает точность классификации примеров в 2 раза. Из проведенных опытов можно сделать вывод, что алгоритм IDTUV способен успешно работать с зашумлённой информацией.

Работа выполнена при поддержке РФФИ, проект № 05-01-00818.

Литература

[1] *Бонгард М. М.* Проблема узнавания. — М.: Наука, 1967. — 320 с.

[2] *Вагин В. Н., Куликов А. В., Фомина М. В.* Методы теории приближенных множеств в решении задачи обобщения понятий // Известия РАН. Теория и системы управления. — 2004. — № 6. — с. 52–66.

[3] *Бериша А. М., Вагин В. Н.* Использование алгоритма построения деревьев решений для зашумлённых данных // Международный форум информатизации — 2004: Труды международной конференции «Информационные средства и технологии». Т. 1. — М.: Янус-К, 2004. — с. 171–174.

[4] *Бериша А. М., Вагин В. Н., Куликов А. В., Фомина М. В.* Методы обнаружения знаний в «зашумленных» базах данных // Известия РАН. Теория и системы управления. — 2005. — № 6. — с. 143–158.

[5] *Quinlan J. R.* C4.5: Programs for Machine Learning. — San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[6] *Quinlan J. R.* Induction of Decision Trees // Machine Learning, 1986. — № 1. — Pp. 1–81.

Исследование стратегий обучения ранговой классификации по методу опорных векторов

Куракин А. В., Татарчук А. И., Моттль В. В.
aleksеyvk@yandex.ru, aitech@yandex.ru, mottl@yandex.ru
Москва, МФТИ, ВЦ РАН

При решении практических задач распознавания образов типичной является ситуация, когда на конечном множестве значений $Y = \{0, \ldots, m\}$ некоторой целевой характеристики объектов $y(\omega) \in Y$, $\omega \in \Omega$, подлежащей восстановлению, задано отношение порядка. Задачу обучения с ранжированными классами (рангами) принято называть *задачей ранговой классификации* или *задачей восстановления ранговой регрессии*. Для двух рангов задача представляет собой классической задачу распознавания образов с двумя классами.

В основе большинства подходов к решению такой задачи лежит естественное предположение о модели ранговой зависимости, состоящее

в том, что объективно существует скрытая характеристика объектов, выраженная некоторой действительнозначной функцией $f\colon \Omega \to \mathbb{R}$. Однако, в отличие от задачи восстановления регрессии, значение такой скрытой характеристики известно только через сравнение его с некоторым набором порогов $h_1 < \cdots < h_m$, разделяющих действительную ось на $m+1$ интервалов. Таким образом, для каждого объекта конечной обучающей совокупности $\omega \in \Omega^* = \{\omega_1, \ldots, \omega_N\} \subset \Omega$ известно лишь, в какой из промежутков $y(\omega) \in Y = \{0, \ldots, m\}$ на оси он попал. При этом все признаковое пространство $\mathbb{R}^n$, в котором объекты представлены векторами своих признаков $\mathbf{x}(\omega) \in \mathbb{R}^n, \omega \in \Omega$, разделяется на упорядоченные области набором параллельных гиперплоскостей с общим направляющим вектором $\mathbf{a} \in \mathbb{R}^n$:

$$\begin{cases} f(\mathbf{x}(\omega)) < h^{(1)} & \to \hat{y}(\omega) = 0; \\ h^{(1)} \leqslant f(\mathbf{x}(\omega)) < h^{(2)} & \to \hat{y}(\omega) = 1; \\ \ldots & \ldots \\ h^{(m)} \leqslant f(\mathbf{x}(\omega)) & \to \hat{y}(\omega) = m; \end{cases}$$

где $f(\mathbf{x}(\omega)) = \mathbf{a}^{\mathbf{T}}\mathbf{x}(\omega)$.

При такой интерпретации природы ранговости целевой характеристики, задача обучения сводится к поиску направляющего вектора $\mathbf{a} \in \mathbb{R}^n$ и набора числовых порогов $h_1 < \cdots < h_m \in \mathbb{R}$.

Основой метода опорных векторов [1], получившего широкое распространение при решении двухклассовой задачи распознавания, является концепция оптимальной разделяющей гиперплоскости, максимизирующей зазор между объектами двух классов. Однако в случае нескольких гиперплоскостей зазоров тоже несколько, и по этой причине возможны различные стратегии обучения [2, 3, 4].

В настоящее время существует два основных обобщения метода опорных векторов для решения задачи ранговой классификации.

Идея так называемой стратегии обучения с фиксированным зазором (Fixed Margin Strategy) состоит в максимизации зазора между объектами ближайших соседних рангов в обучающей совокупности. Альтернативная стратегия обучения с суммированием зазоров (Sum of Margins Strategy) основана на максимизации суммы зазоров между всеми рангами [2]. В данной работе рассматривается только первая стратегия обучения с фиксированным зазором, представляющая собой прямое обобщение метода опорных векторов.

В работе [2] был предложен оптимизационный критерий выбора направляющего вектора $\mathbf{a} \in \mathbb{R}^n$ и набора порогов $h_1, \ldots, h_m \in \mathbb{R}$, которые в соответствующем признаковом пространстве дают наилучшее разделение объектов $\omega \in \Omega$ только двух смежных рангов $y(\omega) = i-1$ и $y(\omega) = i$,

относительно каждой границы $h^{(i)}$. Такая постановка приводит к задаче квадратичного программирования с числом переменных, немногим меньшим удвоенного количества объектов обучающей совокупности $2|\Omega^*|$.

Предложенный подход является привлекательным с вычислительной точки зрения, однако, хотя и в относительно редких случаях, может приводить к некорректной упорядоченности оптимальных величин порогов $\hat{h}^{(i-1)} \geqslant \hat{h}^{(i)}$. Такую постановку будем называть сокращенной стратегией обучения с фиксированным зазором (Truncated Fixed Margin Strategy или TFM).

В работе [3] был предложен другой подход, основанный на идее оптимального разделения каждой границей $h^{(i)}$ объектов из обучающей совокупности, принадлежащих всем низшим рангам $y(\omega) \leqslant i-1$, от объектов всех высших рангов $y(\omega) \geqslant i$. Такой подход автоматически гарантирует корректность упорядоченности оптимальных величин порогов $\hat{h}^{(0)} < \cdots < \hat{h}^{m}$. Будем называть такую постановку полной стратегией обучения с фиксированным зазором (Full Fixed Margin Strategy или FFM).

Ценой гарантированной корректности получаемого решения в полной постановке задачи обучения является необходимость решать задачу квадратичного программирования с числом переменных, равным произведению числа объектов и количества искомых границ $m|\Omega^*|$, что в $(m/2)^3$ раз превышает вычислительную сложность сокращенной постановки.

Следует заметить, что в простейшем случае двух рангов $m = 1$ обе рассматриваемые стратегии обучения вырождаются в классическую постановку задачи обучения распознавания с двумя классами по методу опорных векторов.

В работе [5] определены условия, при которых решение оптимизационной задачи квадратичного программирования сокращенной стратегии обучения в точности совпадает с решением соответствующей оптимизационной задачи полной стратеги обучения.

Предложен новый подход к обучению ранговой классификации, основанный на последовательном применении вначале сокращенной стратегии обучения, затем проверке корректности полученного результата и, в заключении, применении полной стратегии обучения, но только в случае, если условия корректности решения для сокращенной стратегии не выполняются.

Другим отличием рассмотренных обобщений метода опорных векторов от его классической постановки состоит в том, что по решениям соответствующих оптимизационных задач квадратичного программирования удается найти оптимальные величины только для одного активного по-

рога, соответствующего минимальному зазору между рангами. Для нахождения оптимальных значений остальных порогов необходимо решать дополнительную задачу линейного программирования. В данной работе указаны способы явного вычисления оптимальных значений порогов разделяющих гиперплоскостей по решению соответствующих двойственных оптимизационных задач полной и сокращенной стратегий обучения ранговой классификации.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00679, № 06-01-08042, № 06-07-89249, а также INTAS, проект № 04-77-7347.

Литература

[1] *Vapnik V.* Statistical Learning Theory. — New York: John-Wiley & Sons, Inc., 1998. — 732 с.

[2] *Shashua A., Levin A.* Ranking with large margin principle: two approaches // Advances in Neural Information Processing Systems, 2003. — С. 937–944.

[3] *Chu W., Keerthi S. S.* New approaches to support vector ordinal regression // 22nd Int. Conf. on Machine Learning, Bonn, Germany, 2005. — Pp. 145–152.

[4] *Waegeman W., Boullart L.* An ensemble of weighted support vector machines for ordinal regression // Transactions on Engineering, Computing and Technology, 2006. — Vol. 12. — Pp. 71–75.

[5] *Mottl V., Tatarchuk A., Kurakin A.* Support vector machines for ranking learning: the Full and the Truncated fixed margin strategies // Int. Conf. on Machine Learning and Cybernetics, Hong Kong, China, August 19-22, 2007.

Синтез и анализ гибридных алгоритмов распознавания образов

Лапко А. В., Лапко В. А.

lapko@icm.krasn.ru

Красноярск, Институт вычислительного моделирования СО РАН

Совместное использование в едином решающем правиле разнотипных моделей является перспективным способом наиболее полного учёта априорной информации. В данном направлении получены успешные результаты исследований, к которым можно отнести методы локальной аппроксимации [1], гибридные модели стохастических зависимостей [2], полупараметрические и частично линейные модели [3], непараметрические коллективы решающих правил [4]. При этом особое внимание уделяется алгоритмам восстановления стохастических зависимостей, обеспечивающих учет частичных сведений об их виде и данных обучающих выборок.

Для решения проблемы эффективного использования априорной информации предлагаются гибридные системы распознавания образов, которые обеспечивают сочетание в обобщенном решающем правиле класси-

фикации преимущества параметрических и локальных методов аппроксимации, основанных на оценках плотности вероятности типа Розенблатта-Парзена [5].

Синтез гибридных алгоритмов распознавания образов

Пусть исходную информацию при решении двухальтернативной задачи распознавания образов составляют обучающая выборка $V = \left(x^i, \sigma(x^i),\ i = 1, \ldots, n\right)$ и априорные сведения $F_{12}(x, \alpha)$ о виде уравнения разделяющей поверхности $f_{12}(x)$ между классами $\Omega_1,\ \Omega_2$ в пространстве $x \in R^k$.

Информация обучающей выборки V формируется на основании данных о значениях признаков x классифицируемых объектов и соответствующим им «указаний учителя» $\sigma(x) = -1$, если $x \in \Omega_1$; 1, если $x \in \Omega_2$.

Для использования в полном объёме априорной информации $(F_{12}(x, \alpha), V)$ воспользуемся принципами гибридного моделирования.

Для этого определим параметры α уравнения разделяющей поверхности $F_{12}(x, \alpha)$ из условия минимума эмпирической ошибки распознавания образов.

По результатам вычислительного эксперимента сформируем выборку расхождений $V_1 = \left(x^i, q(x^i),\ i = 1, \ldots, n\right)$ между «решениями» $\bar{\sigma}(x^i)$ алгоритма классификации, использующего уравнение разделяющей поверхности $F_{12}(x, \bar{\alpha})$, и «указаниями учителя» $\sigma(x^i)$ из обучающей выборки V. При этом значения функции расхождений

$$q(x^i) = \begin{cases} 0, & \sigma(x^i) = \bar{\sigma}(x^i); \\ F_{12}(x^i, \bar{\alpha}) + \Delta, & \bar{\sigma}(x^i) = -1 \text{ и } \sigma(x^i) = 1; \\ -(F_{12}(x^i, \bar{\alpha}) + \Delta), & \bar{\sigma}(x^i) = 1 \ \text{ и } \sigma(x^i) = -1. \end{cases}$$

При наличии ошибки функция расхождения принимает значение, обратное по знаку уравнения разделяющей поверхности $F_{12}(x, \bar{\alpha})$ и превышает его на величину параметра Δ.

Восстановим функцию $q\,(x)$ по выборке V_1 на основе непараметрической регрессии [6]

$$\bar{q}(x) = \frac{\sum\limits_{i=1}^{n} q(x^i)\beta_i(x)}{\sum\limits_{i=1}^{n} \beta_i(x)}, \quad \beta_i(x) = \prod_{\nu=1}^{k} \Phi\left(\frac{x_\nu - x_\nu^i}{c_\nu}\right),$$

где $\Phi(\cdot)$ — ядерная функция, удовлетворяющая условиям положительности, симметричности и нормированности.

Тогда гибридный алгоритм классификации запишется в виде

$$\bar{m}_{12}(x): \begin{cases} x \in \Omega_1 & \text{если } \bar{f}_{12}(x) \leqslant 0, \\ x \in \Omega_2 & \text{если } \bar{f}_{12}(x) > 0, \end{cases}$$
$$\bar{f}_{12}(x) = F_{12}(x, \bar{\alpha}) + \bar{q}(x).$$

Меняя вид функции $q(x)$, обеспечивающей коррекцию $F_{12}(x, \bar{\alpha})$, получено семейство новых гибридных решающих правил.

Асимптотические свойства гибридных решающих функций

Рассмотрены асимптотические свойства гибридных моделей уравнений разделяющих поверхностей $\bar{f}_{12}(x)$. В частности, при достаточно большом объёме обучающей выборки, смещение

$$\lim_{n\to\infty} \mathsf{E}\left(f_{12}(x) - \bar{f}_{12}(x)\right) = \lim_{n\to\infty} \left[\frac{\left(q(x)p(x)\right)''}{2p(x)} c^2 + \Delta(n) + 0(c^4)\right] = 0,$$

при $\Delta(n) \to 0$ и $c(n) \to 0$. Здесь E — знак математического ожидания; $\left(q(x)p(x)\right)''$ — вторая производная по x произведения $q(x)p(x)$.

Асимптотическое выражение среднеквадратического отклонения имеет вид

$$\mathsf{E}\left(f_{12}(x) - \bar{f}_{12}(x)\right)^2 \sim (ncp(x))^{-1} q^2(x) \left\|\Phi(u)\right\|^2 +$$
$$+ c^4 \left((q(x)p(x))^{(2)}\right)^2 \left(4p^2(x)\right)^{-1} + \frac{(q(x)p(x))''}{p(x)}\, c^2 \Delta(n) + \Delta^2(n),$$

где $\left\|\Phi(u)\right\|^2 = \int \Phi^2(u)\, du$.

Отсюда следует, что $\bar{f}_{12}(x)$ сходится в среднеквадратическом, если $\Delta(n) \to 0$, $c(n) \to 0$ и $nc \to \infty$ при $n \to \infty$.

Асимптотическая несмещённость и сходимость в среднеквадратическом $\bar{f}_{12}(x)$ определяют свойство её состоятельности.

Установлена зависимость скоростей сходимости $\bar{f}_{12}(x)$ от объёма обучающей выборки, вида корректирующей функции и её параметров. На этой основе, исследуя отношение среднеквадратических отклонений модификаций гибридных решающих функций, установлены условия их эффективного использования.

Работа выполнена при поддержке фонда «Научный потенциал».

Литература

[1] *Катковник В. Я.* Линейные и нелинейные методы непараметрического регрессионного анализа // Автоматика, 1979. — № 5. — С. 165–170.

[2] *Лапко А. В.* Имитационные модели неопределённых систем. — Новосибирск: Наука, 1993.

[3] *Хардле В.* Прикладная непараметрическая регрессия. — М.: Мир, 1993.

[4] *Лапко В. А.* Непараметрические коллективы решающих правил. — Новосибирск: Наука, 2002.

[5] *Parzen E.* On the estimation of probability density function and mode // Ann. Math. Stat., 1962. — Vol. 33. — P. 1065–1076.

[6] *Надарая Э. А.* Непараметрические оценки кривой регрессии // Тр. ВЦ АН ГССР, 1965. — Вып. 5. — С. 568.

Комбинированные системы распознавания образов

Лапко А. В., Лапко В. А.

lapko@icm.krasn.ru

Красноярск, Институт вычислительного моделирования СО РАН

В работе с позиции последовательных процедур принятия решений и принципов коллективного оценивания предлагаются статистические модели распознавания образов, представляющие собой семейство частных решающих функций, организация которых в нелинейном решающем правиле осуществляется с помощью методов непараметрической статистики. Частные решающие функции формируются на основе однородных частей обучающей выборки, которые удовлетворяют одному или нескольким требованиям: наличие однотипных признаков, пропусков данных, возможность декомпозиции исходных признаков на группы в соответствии со спецификой решаемой задачи. Это порождает широкий круг постановок задач синтеза непараметрических решающих правил. При интеграции частных решающих функций используются непараметрические оценки оптимальных байесовских решающих правил.

Синтез структуры системы

1. Пусть $V = \left(x_1^i, x_2^i, \ldots, x_k^i,\ \sigma(x^i),\ i = 1, \ldots, n\right)$ — обучающая выборка объёма n, составленная из значений признаков классифицируемых объектов и соответствующих «указаний учителя» об их принадлежности к одному из двух классов, причём $\sigma(x^i) = -1$ для всех $x^i \in \Omega_1$ и $\sigma(x^i) = 1$ для всех $x^i \in \Omega_2$.

Осуществим декомпозицию исходной выборки V на T однородных выборок в соответствии со спецификой задачи

$$V(t) = \left(x^i(t),\ \sigma(x^i),\ i = 1, \ldots, n\right), \quad t = 1, \ldots, T,$$

где $x\,(t)$ имеет размерность k_t, а $\sum_{t=1}^{T} k_t = k$.

2. На основе каждой выборки $V(t)$ построим непараметрическое решающее правило

$$\bar{m}_t\big(x(t)\big) : \begin{cases} x \in \Omega_1, & \bar{f}_{12}\big(x(t)\big) \leqslant 0; \\ x \in \Omega_2, & \bar{f}_{12}\big(x(t)\big) > 0; \end{cases} \qquad t = 1, \ldots, T, \tag{1}$$

где $\bar{f}_{12}\big(x(t)\big)$ — непараметрические оценки решающих функций

$$\bar{f}_{12}\big(x(t)\big) = \left(n \prod_{v \in I_t} c_v\right)^{-1} \sum_{i=1}^{n} \sigma\left(x^i\right) \prod_{v \in I_t} \Phi\left(\frac{x_v - x_v^i}{c_v}\right);$$

I_t — множество номеров признаков, входящих в группу $x(t)$; $\Phi(\cdot)$ — ядерные функции, удовлетворяющие условиям положительности, симметричности, нормированности, и имеющие конечные центральные моменты [1].

Оптимизация частных решающих правил (1) по коэффициентам размытости ядерных функций $c_v,\ v \in I_t$ осуществляется в режиме «скользящего экзамена» из условия минимума статистической оценки вероятности ошибки распознавания образов.

3. Используя непараметрические оценки решающих функций $\bar{f}_{12}(x(t))$, сформируем обучающую выборку

$$\big(\bar{f}_{12}(x^i(t)),\ t = 1, \ldots, T,\ \sigma(x^i),\ i = 1, \ldots, n\big)$$

и построим комбинированное решающее правило в пространстве значений $\bar{f}_{12}(x) = \big(\bar{f}_{12}(x(t)),\ t = 1, \ldots, T\big)$

$$\bar{m}\left(\bar{f}_{12}\left(x\right)\right) : \begin{cases} x \in \Omega_1, & \bar{F}_{12}\left(\bar{f}_{12}\left(x\right)\right) \leqslant 0; \\ x \in \Omega_2, & \bar{F}_{12}\left(\bar{f}_{12}\left(x\right)\right) > 0; \end{cases}$$

где непараметрическая оценка обобщённой решающей функции между классами имеет вид

$$\bar{F}_{12}(\bar{f}_{12}(x)) = \left(n \prod_{v=1}^{T} c_v\right)^{-1} \sum_{i=1}^{n} \sigma(x^i) \prod_{v=1}^{T} \Phi\left(\frac{\bar{f}_{12}(x(t)) - \bar{f}_{12}(x^i(t))}{c_v}\right).$$

На первом уровне структуры рассматриваемой системы классифицируемая ситуация x преобразуется в значения непараметрических оценок $\bar{f}_{12}\big(x(t)\big),\ t = 1, \ldots, T$, в пространстве которых принимается решение $\bar{\sigma}\left(x\right)$ правилом $\bar{m}\left(\bar{f}_{12}\left(x\right)\right)$ о принадлежности ситуации x к тому или иному классу.

Предлагаемый алгоритм классификации обеспечивает не только эффективное решение задач распознавания образов в условиях малых выборок, но и позволяет учитывать априорные сведения о виде частных решающих функций.

Анализ результатов вычислительного эксперимента

На основании данных вычислительного эксперимента сравнивается эффективность комбинированных решающих правил с хорошо зарекомендовавшим себя на практике традиционным непараметрическим алгоритмом распознавания образов в пространстве признаков $x = (x_1, \ldots, x_k)$ [2].

Исследования осуществлялись при решении двухальтернативной задачи распознавания образов в k-мерном пространстве признаков со сложной нелинейной границей, $k = 4, \ldots, 20$.

Достоверность различия эмпирических оценок вероятности ошибки распознавания образов сравниваемых методов рассчитывалась в соответствии с критерием Смирнова. При этом установлено достоверное преимущество предлагаемого алгоритма над традиционным. Данная закономерность сохраняется для различных объемов обучающих выборок.

Обнаружен экстремальный характер зависимости показателей эффективности комбинированного классификатора от количества T частных решающих правил. Причём с ростом размерности k признаков классифицируемых объектов его преимущество при оптимальных значениях T над традиционным непараметрическим алгоритмом возрастает. Отношение средних значений их оценок вероятности ошибки достигает трёх на контрольных выборках, что особенно проявляется при малых объёмах экспериментальных данных.

Заключение

Нелинейные непараметрические алгоритмы распознавания образов являются эффективным средством решения задач классификации в условиях малых обучающих выборок. Их применение обеспечивает значительное снижение ошибки распознавания образов на контрольных выборках (в 1.5–3 раза) по сравнению с традиционным непараметрическим классификатором.

Перспективы развития предлагаемого подхода связаны с его применением в задачах классификации в условиях разнотипной информации и неоднородных выборок, получаемых в результате заполнения пропусков данных.

Работа выполнена при поддержке РФФИ, проект № 07-01-00006.

Литература

[1] *Лапко А. В.* Непараметрические системы классификации / А. В. Лапко, В. А. Лапко, М. И. Соколов, С. В. Ченцов. — Новосибирск: Наука, 2000. — 240 с.

[2] *Лапко А. В.* Обучающиеся системы обработки информации и принятия решений / А. В. Лапко, С. В. Ченцов, С. И. Крохов, Л. А. Фельдман. — Новосибирск: Наука, 1996. — 296 с.

Адаптивные методы построения логических решающих функций в задачах распознавания образов, регрессионного анализа и оптимизации

Лбов Г. С.

lbov@math.nsc.ru

Новосибирск, Институт математики СО РАН

В рамках единого подхода рассматривается задача адаптивного построения решающих функций в области распознавания, регрессионного анализа и многоэкстремальной оптимизации. Единый подход заключается в использовании класса логических решающих функций от разнотипных переменных [1]. Необходимость в адаптивном построении решающих функций возникает при больших затратах либо на физический эксперимент, либо на подбор значений параметров сложной математической модели изучаемого объекта. При этом естественно предполагается, что активный эксперимент возможен.

В распознавании образов наиболее полно изучен классический случай пассивного эксперимента на основе анализа таблиц данных. Идея изменения решающей функции распознавания по мере поступления объектов обучающей выборки также достаточно известна (например, эта идея используется в нейронных сетях и в методе растущих пирамидальных сетей). В регрессионном анализе также существуют методы, позволяющие проводить последовательную адаптацию регрессионной функции по отношению к выборке. Также к настоящему времени существуют адаптивные методы для решения задач регрессионного анализа и многоэкстремальной оптимизации (например, генетические алгоритмы).

В данной работе предлагаются методы не последовательного, а адаптивного планирования экспериментов для решения трех указанных задач в случае разнотипных переменных. В докладе приводятся результаты исследований адаптивного планирования экспериментов для регрессионного анализа.

Также для распознавания образов разработан метод адаптивного согласования экспертных вероятностных логических высказываний, при

этом возможен еще один вариант адаптивной обработки информации: при наличии противоречивых или недостаточно достоверных данных запрашивается дополнительная информация от экспертов. В докладе также приводятся результаты исследования эффективности адаптивного алгоритма многоэкстремальной оптимизации. Исследования проведены на пяти известных тестовых функциях. При этом рассматривается связь между точностью решения, числом испытаний и коэффициентом адаптации.

Работа выполнена при поддержке РФФИ, проект № 07-01-00331а.

Литература

[1] *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: Издательство Института математики, 1999. — 211 с.

Адаптивное планирование эксперимента в распознавании образов и в регрессионном анализе с использованием класса логических решающих функций

Лбов Г. С., Бериков В. Б.

lbov@math.nsc.ru, berikov@math.nsc.ru

Новосибирск, Институт математики им. С. Л. Соболева СО РАН

В достаточно широком круге задач интеллектуального анализа данных, возникающих в различных трудноформализуемых областях исследований, имеется возможность активного влияния на выборку. Так, например, при изучении нового способа лечения можно целенаправленно проводить отбор пациентов с интересующей формой патологии; при исследовании мутационных спектров ДНК имеется возможность выбора генетической последовательности, подвергающейся действию мутагена, и т. д. Сбор экспериментальной информации и анализ полученных данных связаны с большими затратами, которые определяются числом изучаемых объектов. Выбор объектов может быть организован так, чтобы добиться наилучшего качества при заданном числе наблюдений.

Одним из возможных методов решения такого рода задач является предлагаемый метод, основанный на адаптивном планировании случайного эксперимента с использованием класса логических решающих функций. При адаптивном планировании проводится последовательный случайный отбор объектов с учетом уже выявленных закономерностей в структуре данных. Логические решающие функции, наиболее удобная форма представления которых — деревья решений, позволяют строить легко интерпретируемые модели, одновременно проводя отбор наиболее

информативных показателей. Основная задача состоит в том, чтобы построить дерево решений, наилучшим образом приближающее оптимальную байесовскую решающую функцию (либо соответствующую регрессионную функцию) f_0. Предположим, что имеется экспертная информация о том, что вероятность ошибки для f_0 (дисперсия помехи, в случае регрессионного анализа), ограничена некоторой малой величиной. Для задачи распознавания это означает, что образы достаточно хорошо «разделены» в пространстве переменных.

Пусть задано максимально возможное число N объектов анализа, каждый из которых описывается некоторыми показателями, среди которых могут быть показатели как количественной, так и качественной природы. Имеется некоторый прогнозируемый показатель, который может быть либо качественным (в случае задачи распознавания), либо количественным (для задачи регрессионного анализа). Зададим некоторое число L этапов планирования. При проведении l-го этапа планирования используется эмпирическая информация, полученная на основе анализа всех экспериментов предыдущих $l-1$ этапов.

На первом этапе расстановка планируемых точек в пространстве переменных осуществляется случайным образом с использованием равномерного распределения (так как на данном этапе отсутствует информация о поведении прогнозируемой переменной). По сформированной таким образом выборке строится дерево решений. Для этого может использоваться, например, рекурсивный $\mathfrak{R}$-метод [1].

Рассмотрим разбиение области планирования на M подобластей $E_l^{(1)}, \ldots, E_l^{(m)}, \ldots, E_l^{(M)}$, соответствующее листьям дерева решений, построенного по наблюдениям предыдущих этапов (рис. 1; задача распознавания образов).

Проведение l-й группы экспериментов следует организовать так, чтобы в максимальной степени уменьшить степень расхождения с оптимальной функцией. В качестве оценки степени расхождения можно использовать байесовскую оценку вероятности ошибки [2], в которой учитывается число имеющихся подобластей разбиения и экспертная оценка степени «пересечения» образов. Можно также использовать и обычную частотную оценку. В случае задачи регрессионного анализа оценкой может служить величина среднеквадратического отклонения прогнозируемого показателя. Пусть фиксирован способ расстановки планируемых на данном этапе N_l точек, $\sum\limits_{l=1}^{L} N_l = N$, в каждой из подобластей (например, в соответствии с равномерным распределением). При этом адаптация будет заключаться в изменении набора вероятностей $\{P_l^{(1)}, \ldots, P_l^{(m)}, \ldots, P_l^{(M)}\}$ попадания планируемой точки в подобласти. Это изменение должно от-

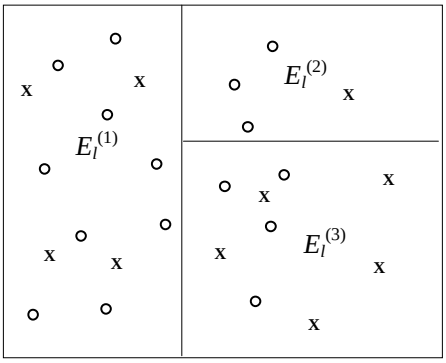


Рис. 1. Пример дерева разбиения и планируемых объектов (x — первый образ; o — второй образ)

ражать накопленную на данном этапе информацию о поведении прогнозируемой переменной. Например, если выяснилось, что прогнозируемая количественная переменная в некоторой подобласти изменяется относительно мало, то и вероятность попадания в эту подобласть должна быть уменьшена.

Рассмотрим класс стратегий планирования, для которых выполняется: $P_l^{(m)} = g_l\big(\delta^{(m)}, |E^{(m)}|\big)$, где $\delta^{(m)}$ — оценка степени расхождения с оптимальной функцией в области $E^{(m)}$, $|E^{(m)}|$ — мощность или объем данной области, $g_l(\cdot,\cdot)$ — некоторая заданная неотрицательная функция, монотонно возрастающая по каждому из аргументов, причем должно выполняться $\sum\limits_{m=1}^{M} P_l^{(m)} = 1$. Таким образом, вероятность попадания точки в подобласть увеличивается при возрастании, с одной стороны, ошибки для данной подобласти, а с другой стороны, объема подобласти. Конкретный вид функции g_l может задаваться по разному. Например, можно положить, что эта функция линейна: $g_l(a,b) = \frac{1}{Z_l}\big(\varkappa(l)a + b\big)$, где Z_l — нормировочный коэффициент, $\varkappa(l)$ — коэффициент адаптации, растущий с увеличением номера этапа планирования l («адаптация» означает, что с увеличением объема информации растет степень доверия к соответствующей оценке).

После проведения планирования для данного этапа строится новое дерево решений по сформированной выборке, и т. д.

В докладе будут продемонстрированы результаты экспериментального исследования алгоритма адаптивного планирования.

Работа выполнена при поддержке РФФИ, проект № 07-01-00331а.

Литература

[1] *Лбов Г. С., Бериков В. Б.* Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. — Новосибирск: Изд-во Ин-та математики, 2005. — 220 с.

[2] *Бериков В. Б., Лбов Г. С.* Байесовские оценки качества распознавания по конечному множеству событий // Доклады РАН — 2005. — Т. 402. — № 1. — С. 1–4.

Построение решающей функции распознавания на основе экспертных высказываний

Лбов Г. С., Герасимов М. К., Толстик А. А.

lbov@math.nsc.ru, max_post@ngs.ru

Новосибирск, Институт математики СО РАН

Введение. Рассматривается задача адаптивного согласования экспертных знаний, представленных в виде логико-вероятностных высказываний [1]. При этом высказывания могут быть частично или полностью повторяющимися (у одного и того же эксперта), подтверждающими друг друга (у разных экспертов), дополняющими или противоречивыми. Предполагается, что со временем экспертные знания могут изменяться, а также возможны добавления новых знаний от других экспертов.

Постановка задачи. Пусть Γ — некоторая совокупность объектов. Каждому объекту $a \in \Gamma$ поставлены в соответствие номер образа $Y(a) = k,\ k \in \{1, \ldots, K\}$, и набор значений $X(a) = \big(X_1(a), \ldots, X_n(a)\big)$, где $X_j(a)$ — значение переменной X_j для объекта a. Набор $X = \{X_1, \ldots, X_n\}$ может одновременно содержать как количественные, так и качественные переменные. Обозначим через D_j множество возможных значений переменной X_j. Декартово произведение $D = \prod_{j=1}^{n} D_j$ задает многомерное пространство разнотипных переменных.

Назовем $E \subseteq D$ *прямоугольной областью*, если $E = \prod_{j=1}^{n} E_j$, где $E_j = [\alpha_j, \beta_j]$, если X_j — количественная переменная; E_j — некоторый список имен, если X_j — номинальная переменная.

В данной статье рассматриваются экспертные высказывания S следующего вида: «если $X(a) \in E$, то $Y(a) = k$ с вероятностью p», где E — прямоугольная область. При этом предполагается, что эксперт делает высказывание, используя, как правило, лишь некоторые переменные из набора X. В этом случае, с нашей точки зрения, эксперт допускает все возможные значения для остальных переменных, т. е. для них $E_j = D_j$. Пусть имеется набор высказываний $\Omega = \{S^1, \ldots, S^M\}$. Обозначим через Ω^k высказывания о принадлежности объектов к образу k. Предполагается, что каждому высказыванию S^i приписан некоторый вес w^i в соответствии

с компетентностью эксперта, сделавшего высказывание, и уверенностью эксперта в достоверности высказывания. При отсутствии априорной информации всем высказываниям приписываются одинаковые веса. Высказыванию S^i можно поставить в соответствие четверку $\langle E^i, k^i, p^i, w^i\rangle$. Требуется из исходного набора высказываний Ω построить другой набор Ω' из минимального числа согласованных друг с другом высказываний таким образом, чтобы набор Ω' был максимально согласован с набором Ω с точки зрения качества распознавания.

Для согласования высказываний предлагается метод, использующий расстояния в многомерном разнотипном пространстве. Можно использовать, например, меры близости, предложенные в [2]–[4].

Рассмотрим сначала по отдельности высказывания каждого эксперта по каждому образу k.

Пусть имеются прямоугольные области E^{i_1} и E^{i_2}. Определим множество $E^{i_1 i_2}$ следующим образом: $E^{i_1 i_2} := E^{i_1} \oplus E^{i_2} = \prod_{j=1}^{n}(E_j^{i_1} \oplus E_j^{i_2})$, где $E_j^{i_1} \oplus E_j^{i_2} = E_j^{i_1} \cup E_j^{i_2}$, если X_j — номинальная; и минимальный интервал, такой что $E_j^{i_1} \cup E_j^{i_2} \subseteq E_j^{i_1} \oplus E_j^{i_2}$, если X_j — количественная переменная.

Введем меру «незначительности» множества $E^{i_1 i_2} \setminus \big(E^{i_1} \cup E^{i_2}\big)$. Можно использовать, например, величину

$$r^{i_1 i_2} := \max_{E'} \frac{\operatorname{diam}(E')}{\operatorname{diam}(E^{i_1 i_2})},$$

где $E' \subseteq E^{i_1 i_2} \setminus (E^{i_1} \cup E^{i_2})$ — прямоугольная область в D.

Рассмотрим множества $E^{i_1}, \ldots, E^{i_q}$ такие, что $r^{i_u i_v} \leqslant \varepsilon$ для всех $u, v \in \{1, \ldots, q\}$, где ε — параметр $(0 \leqslant \varepsilon \leqslant 1)$, $2 \leqslant q \leqslant Q$, Q — число высказываний данного эксперта по образу k. Пусть нет такого множества E^l, что $r^{l i_u} \leqslant \varepsilon$ для всех $u \in \{1, \ldots, q\}$.

Обозначим $J_q = \{i_1, \ldots, i_q\}$, $E^{J_q} = E^{i_1} \oplus \cdots \oplus E^{i_q}$, $c^{i J_q} = 1 - \rho(E^i, E^{J_q})$, где $\rho(E, F)$ — расстояние между прямоугольными областями E и F [3]. Объединим высказывания $S^{i_1}, \ldots, S^{i_q}$ в высказывание $S^{J_q} = \langle E^{J_q}, k, p^{J_q}, w^{J_q}\rangle$, где

$$p^{J_q} = \frac{\sum_{i \in J_q} c^{i J_q} w^i p^i}{\sum_{i \in J_q} c^{i J_q} w^i}, \quad w^{J_q} = \left(1 - d\Big(E^{J_q}, \bigcup_{i \in J_q} E^i\Big)\right) \frac{\sum_{i \in J_q} c^{i J_q} w^i}{\sum_{i \in J_q} c^{i J_q}}.$$

Процедура согласования высказываний одного эксперта по образу k состоит в построении всех высказываний S^{J_q}, $q = 2, \ldots, Q$.

После согласования высказываний для каждого образа экспертов по отдельности, мы можем построить скоординированное решающее правило по каждому образу. Процедура аналогична вышеописанной, за исключением весов: $w^{J_q} = \sum_{i \in J_q} c^{i J_q} w^i$ (чем больше экспертов делают похожие

высказывания, тем они достовернее). Обозначим $\Omega_1 = \bigcup_{k=1}^{K} \Omega_1^k$, где Ω_1^k — набор согласованных высказываний для образа $k \in \{1, \ldots, K\}$.

Решение противоречий. Рассмотрим высказывания $S^i \in \Omega_1^i$, $i \in I$, $I = \{i_1, \ldots, i_m\} \subseteq \{1, \ldots, K\}$, $i_u \neq i_v$ при $u \neq v$. Назовем их *противоречивыми* на множестве $E \subseteq D$, если $E \subseteq E^i$ для всех $i \in I$ и $\sum_{i=i_1}^{i_m} p^i > 1$.

Зададим для каждого $k \in \{i_1, \ldots, i_m\}$ множество

$$I(k) = \left\{ t \;\middle|\; S^t \in \Omega^k \text{ и } \rho(E^t, E) < \varepsilon^* \right\}, \text{ где } \varepsilon^* \text{ — параметр.}$$

Обозначим

$$c^t = 1 - \rho(E^t, E); \qquad p(k) = \frac{\sum_{t \in I(k)} c^t w^t p^t}{\sum_{t \in\ I(k)} c^t w^t}; \quad w(k) = \frac{\sum_{t \in I(k)} c^t w^t}{\sum_{t \in I(k)} c^t};$$

$$\tilde{p}^k = p(k) - \sum_{l \in I} \big(p(l) - 1\big) \frac{\sum_{l \in I} w(l) - w(k)}{(m-1) \sum_{l \in I} w(l)}; \quad \tilde{w}^k = \frac{w(k)}{1 + |\sum_{l \in I} p(l) - 1|}.$$

Сформулируем новые высказывания $\tilde{S}^k = \langle E, k, \tilde{p}^k, \tilde{w}^k \rangle$.

Обозначим через Ω_2 набор согласованных таким способом высказываний.

Результирующее решающее правило строится на основе согласованных наборов высказываний Ω_1 и Ω_2. Рассмотрим множество $E \subset D$. Если найдётся $S^i \in \Omega_2$ такое, что $E \cap E^i \neq \varnothing$, то для $E \cap E^i$ строится решение по высказываниям из набора Ω_2, в противном случае строится решение для E по высказываниям из набора Ω_1.

Заключение. Заметим, что возможен еще один вариант адаптивной обработки информации: при наличии противоречивых или недостаточно достоверных данных в определенных областях запрашивается дополнительная информация от экспертов.

Работа выполнена при поддержке РФФИ, проект № 07-01-00331а.

Литература

[1] *Lbov G., Gerasimov M.* Constructing of a Consensus of Several Experts Statements // Proc. of XII Int. Conf. «Knowledge — Dialogue — Solution», Varna, Bulgaria, 2006. — С. 193–195.

[2] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Издательство Института математики, 1999. — 268 с.

[3] *Лбов Г. С., Герасимов М. К.* Введение расстояния между логическими высказываниями в задачах прогнозирования // Искусственный интеллект. — 2004. — № 2. — С. 105–108.

[4] *Викентьев А. А.* Расстояние на высказываниях экспертов и мера опровержимости (информативности) высказываний с помощью моделей некоторых теорий // 12-я Всероссийская конференция «Математические методы распознавания образов», Москва, 2005. — С. 60–63.

Построение решающих деревьев минимальной стоимости для попарного сравнения объектов

Майсурадзе А. И.
useraim@mail.ru
Москва, Московский государственный университет

Рассматривается задача построения коллективной функции сходства для пар объектов распознавания при заданном наборе функций сходства. При этом считается, что функции сходства из набора для пар объектов еще не вычислены и имеют определенную стоимость вычисления, а результирующая коллективная функция сходства должна удовлетворять заданным прецедентным ограничениям.

Постановка задачи синтеза функции сходства

При решении многих задач интеллектуального анализа данных нередко предлагается проводить сравнение исследуемых объектов или вводить на них некоторые функции сходства. Во многих прикладных областях такие функции сходства возникают естественным путём, превосходя по качеству попытки признакового описания объектов. В других случаях функции сходства отражают экспертные знания о предментной области. Но практически всегда естественные или экспертные функции сходства невозможно или малоэффективно непосредственно использовать для решения поставленной задачи интеллектуального анализа данных.

В то же время в указанных ситуациях может быть доступна прецедентная информация — известно желаемое значение функции сходства для некоторых пар объектов. В частности, искомая функция сходства может быть формализована как бинарное отношение на объектах, для которого известна истинность или ложность на некоторых парах объектов. Кроме того, иногда на искомую функцию сходства налагаются универсальные ограничения. В рассматриваемом ниже прикладном примере искомая функция сходства должна быть симметричным бинарным отношением, значения которого известны для всех пар объектов из фиксированного списка.

Таким образом, в описанных выше ситуациях может быть поставлена задача синтеза функции сходства по заданному набору естественных или экспертных функций сходства, удовлетворяющей заданным прецедентным и универсальным ограничениям. Указанная задача практически яв-

ляется классической задачей обучения по прецедентам, для решения которой могут быть привлечены многочисленные алгоритмические модели восстановления регрессии или классификации. Существенной особенностью рассматриваемой задачи является тот факт, что, в отличие от традиционных признаков, вычисление функции сходства из набора во многих прикладных областях является весьма трудоёмкой операцией.

Указанная трудоёмкость создает сложности как на этапе настройки, так и в ходе использования настроенной функции сходства. В данной работе (исходя из содержания прикладного примера) выбор делается в пользу функций, которые в среднем как можно быстрее обрабатывают новые объекты, но могут потребовать существенных затрат времени при первоначальной настройке.

Описание модели на базе решающих деревьев

Большинство стандартных моделей классификации объектов по признаковым описаниям устроены следующим образом: отказ от использования некоторого признака может произойти только в ходе настройки, а при обработке каждого очередного объекта распознавания требуются значения всех оставленных признаков. В отличие от этого, основываясь на приведенной выше формализации задачи, требовалось выбрать такую модель алгоритмов классификации, которая при обработке очередного объекта распознавания запрашивает значения лишь некоторых признаков, причем при обработке разных объектов могут потребоваться разные признаки. Одной из наиболее распространенных и изученных моделей, обладающих указанным свойством, является модель решающих деревьев.

Отметим, что в рассматриваемом в работе подходе «объектом распознавания» является пара исходных объектов. Таким образом, на вход модели подается «признаковое описание» пары исходных объектов — вектор значений экспертных функций сходства из заданного набора. При этом реальное вычисление значений происходит тогда, когда значение действительно потребуется (lazy evaluation). Каждому вычислению экспертной функции сходства можно приписать стоимость (например, пропорционально тродоёмкости вычислений). Традиционные решающие деревья производят ветвление, сравнивая в узле значение только одной экспертной функции сходства с фиксированным порогом. Следовательно, каждой обработке настроенной моделью очередного объекта распознавания можно приписать суммарную стоимость вычисления. (Возможны два подхода: дублировать стоимость повторно вычисляемой экспертной функции сходства, либо кэшировать значения.) Кроме того, стоимость можно приписать ошибкам распознавания на прецедентах (штрафы). Таким образом, при заданном наборе прецедентов (пар исходных объек-

тов) каждое решающее дерево описывается парой стоимостей: затраты на вычисления и штраф за ошибки. Если стоимости заданы в сравнимых единицах измерения, то можно перейти к суммарной стоимости дерева. Это позволяет поставить задачу поиска решающего дерева минимальной стоимости для заданного набора прецедентов.

К положительным сторонам описанного подхода следует отнести тот факт, что минимизируемый функционал ограничивает рост дерева. Полученное дерево как бы не нуждается в «обрезке», т. к. размер дерева и точность распознавания уже согласованы. К отрицательным сторонам следует отнести высокую трудоёмкость настройки.

Прикладной пример

В работе [1] была рассмотрена задача биометрической идентификации личности по форме ладони. Аппроксимация силуэта ладони многоугольной фигурой и построение её скелета позволили разработать для сравнения ладоней целый ряд функций близости, описываемых полуметриками. Вновь поступающее для идентификации изображение требовалось сравнить со всей базой эталонов (по несколько эталонов для каждого человека). К сожалению, вычисление наилучших по качеству идентификации функций сходства по парам изображений требовало больших затрат времени, неприемлемых в реальных прикладных системах идентификации. Полный же отказ от таких функций приводил к недопустимой деградации качества распознавания. В то же время, правдоподобной выглядела гипотеза, что многие пары можно оценить, пользуясь только функциями сходства, не требующими существенных затрат времени. Предложенный в настоящей работе подход позволит существенно повысить качество идентификации за приемлемое для реальных прикладных систем время.

Работа выполнена при поддержке РФФИ, проект № 07-01-00211-а.

Литература

[1] *Местецкий Л. М., Мехедов И. С.* Комбинированное правило ближайших соседей при классификации формы ладоней // Тез. докл. межд. конф. Интеллектуализация обработки информации (ИОИ-2006). — Симферополь: Крымский научный центр НАН Украины, 2006. — С. 142–144.

Применение обобщенного метода наименьших квадратов к задаче построения разделяющей гиперплоскости

Матросов В. Л., Горелик В. А., Жданов С. А., Муравьева О. В.
gorelik@ccas.ru, saj@mpgu.edu.ru, muraveva@mpgu.edu.ru
Москва, МПГУ

Пусть задан набор объектов в n-мерном пространстве признаков. Требуется определить разрешающее правило, разделяющее множество точек на два класса. На практике признаки объектов меряются приближенно, поэтому границы между классами (выборками) имеют весьма причудливую форму и даже размыты. Вместе с тем, желательно, чтобы решающие правила были попроще. Особенно удобны линейные правила, когда границы между классами представляют собой гиперплоскости. Покажем, как можно использовать методы коррекции данных для построения разделяющих гиперплоскостей в пространстве признаков.

Класс K_1 представлен объектами с векторами признаков $x^1, \ldots, x^k$, класс K_2 — выборкой $y^1, \ldots, y^l$, $m = k + l$. Требуется найти коэффициенты линейной решающей функции $F(x) = (a, x) - b$, т. е. найти коэффициенты $a \in \mathbb{R}^n$ и $b \in \mathbb{R}$ такие, что выполняется система

$$\begin{cases} (a, x^i) \leqslant b, \ i = 1, \ldots, k; \\ (a, y^j) \geqslant b, \ j = 1, \ldots, l. \end{cases}$$

Полагая, что эта система неравенств относительно a, b несовместна, рассмотрим задачу минимальной коррекции всех объектов выборки по критерию суммы квадратов расстояний от заданных точек до их образов при коррекции.

Обозначим матрицу входных данных $X = \left[x^1, \ldots, x^k, -y^1, \ldots, -y^l\right]^{\mathrm{T}}$; $X' = \left[x'^1, \ldots, x'^k, -y'^1, \ldots, -y'^l\right]^{\mathrm{T}}$ — матрица скорректированных значений признаков, $H = X' - X$ — матрица коррекции, вектор $(b, \ldots, b, -b, \ldots, -b)^{\mathrm{T}} = bp$, где $p = (1, \ldots, 1, -1, \ldots, -1)^{\mathrm{T}} \in \mathbb{R}^m$. Получим задачу коррекции несовместной системы линейных неравенств

$$v = \inf_{H,a,b} \left\{ \|H\|^2 \colon (X + H)a \leqslant bp \right\}, \quad \text{где } \|H\|^2 = \sum_{i,j=1}^{m,n} h_{ij}^2.$$

Обозначим $I' \subset I = \{1, \ldots, m\}$, $\bar{X}$ — подматрица матрицы X, состоящая из части строк X с номерами из I', $\hat{X}$ — подматрица X, состоящая из остальных строк, $\bar{p}$, $\hat{p}$ — вектора с координатами p, и с номерами, соответственно, I' и $I \setminus I'$, $P_{\bar{p}} = \frac{\bar{p}\bar{p}^{\mathrm{T}}}{(\bar{p}, \bar{p})}$ — матрица проектирования на векторное пространство с базисом $\bar{p}$. Можно показать, что значение задачи

определяется формулой [1]:

$$v = \min_{I' \subset I\colon \hat{X}e - b\hat{p} \leqslant 0} \lambda_{\min}(\bar{X}^{\mathrm{T}}(E - P_{\bar{p}})\bar{X}),$$

где E — единичная матрица, $\lambda_{\min}(\bar{X}^{\mathrm{T}}(E - P_{\bar{p}})\bar{X})$ — минимальное собственное число матрицы $\bar{X}^{\mathrm{T}}(E - P_{\bar{p}})\bar{X}$, e — соответствующий собственный вектор. Решение задачи коррекции, в том числе коэффициенты a, можно выразить через e.

В результате получим следующий метод построения решающей функции. Рассматриваются все подсистемы неравенств системы $Xa - bp \leqslant 0$, в порядке возрастания мощности. Для рассматриваемой подсистемы $\bar{X}a - b\bar{p} \leqslant 0$ определяется минимальная матрица коррекции H такая, что совместна система уравнений $(\bar{X} + H)a - b\bar{p} = 0$. Если, кроме того, выполняются остальные неравенства, $\hat{X}a - b\hat{p} \leqslant 0$, то 1) полученное значение $\|H\|$ сравнивается с текущим наилучшим значением задачи коррекции; 2) все подмножества, содержащие данное, можно исключить из рассмотрения.

Геометрически алгоритм можно интерпретировать следующим образом: для выбранного подмножества I' точек строится аппроксимирующая гиперплоскость по полному методу наименьших квадратов, т. е. координаты этих точек корректируются так, чтобы через них можно было провести гиперплоскость.

Работа выполнена при поддержке Программы Федерального агенства по образованию «Развитие потенциала высшей школы».

Литература

[1] *Матросов В. Л., Горелик В. А., Жданов С. А., Муравьева О. В.* Применение методов коррекции несобственных задач линейного программирования к задаче классификации // Научные труды Мос. пед. гос. ун-та. Серия: Естественные науки, М: Прометей, 2005. — С. 55–60.

Об одном методе оценок

Матросов В. Л., Угольникова Б. З.
Москва, МПГУ

Пусть имеется объект R, который надо распознать, т. е. отнести к одному из классов $K_1, \ldots, K_s$. Дана начальная оценка объекта, и задан набор признаков $j = 1, \ldots, n$, по которым объект может изучаться. При этом изучение (оценка) объекта проходит последовательно. Случайным образом выбирается первый признак. Объект изучается, меняется его оценка, и т. д. После выбора k признаков ($k < n$) делается попытка распознать объект, при этом все $k + 1$ оценка объекта держатся в памяти.

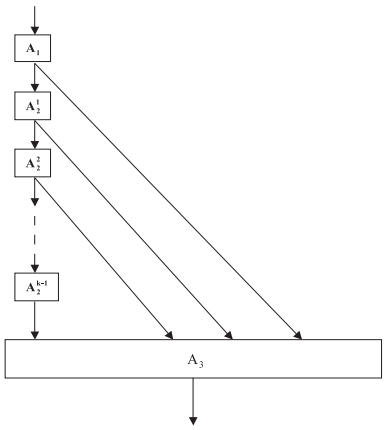


Рис. 1. Цепочка конечных автоматов.

Если распознать объект невозможно, то он изучается по следующему признаку. По последним k оценкам объекта и начальной оценке делается следующая попытка распознавания. При этом учитывается общая информация по всем предыдущим оценкам. Например, опрос студента на экзамене. Есть предварительная информация о работе в семестре. Далее случайным образом выбирается билет. Студент последовательно отвечает на вопросы, и в процессе ответа у экзаменатора или складывается окончательное мнение об оценке студента, или предлагаются дополнительные вопросы. Окончательная оценка ставится исходя из «общего впечатления» и ответов на некоторое количество последних вопросов. Для решения такой задачи предлагается использовать цепочку конечных автоматов, как показано на рисунке.

Автомат A_1 имеет p состояний ($p \geqslant s$). Начальное состояние выбирается по начальной оценке объекта. Входной алфавит — это результаты изучения объекта по признакам. Выходной алфавит — это номера классов (или состояний). Автоматы A_2^i — это автоматы-задержки с нулевым начальным состоянием $i = 1, \ldots, k-1$. На вход автомата A_2^1 подается выход автомата A_1, далее автоматы соединены последовательно. Автомат A_3 имеет k входов — это выходы автоматов и A_1, и A_2^i, $i = 1, \ldots, k-1$. Таким образом, в момент времени t на вход автомата A_3 поступает информация о состоянии объекта в последние k моментов времени $t \geqslant k$. Автомат A_3 имеет $s+1$ состояние: s состояний соответствуют классам $K_1, \ldots, K_s$, и одно состояние конечное. Начальное состояние соответствует начальной оценке объекта. Если в момент времени t автомат может распознать объект, то он переходит в конечное состояние и выдает но-

мер класса. Если распознать не может — то остается в том же начальном состоянии и выдает 0.

Изучается ряд вопросов.

1. Задать функции переходов и выходов для автоматов A_1 и A_3 таким образом, чтобы после оценивания объекта не более чем по N признакам процесс распознавания завершался. N задаётся в зависимости от конкретной задачи. Например, для оценивания знаний на экзамене можно брать $k = 3$, $N = 6$.
2. Сравниваются возможности такой системы автоматов при различной глубине памяти, т. е. при изменении количества автоматов-задержек.
3. Рассматриваются случаи, когда автомат A_1 имеет кроме основных s состояний, соответствующих классам $K_1, \ldots, K_s$, еще некоторые промежуточные состояния. Например, при оценке ответа на экзамене ставятся промежуточные баллы (ответ между 3 и 4).

Комитетный бустинг: минимизация числа базовых алгоритмов при простом голосовании

Мацепов А. А.
amacenov@yandex.ru
Москва, МФТИ

Построение композиций алгоритмов является универсальным способом повышения качества классификации в сложных прикладных задачах. Вполне естественно требовать, чтобы композиция была не избыточной, т. е. состояла из минимального числа базовых алгоритмов.

Пусть X — множество допустимых описаний объектов, $Y = \{-1, 1\}$ — множество меток классов, $X^\ell = (x_i, y_i)_{i=1}^\ell \subset X \times Y$ — обучающая выборка, $b_t \colon X \to Y$ — базовые алгоритмы, $t = 1, \ldots, T$. *Комитетом большинства* [1] называется алгоритмическая композиция вида

$$a(x) = \operatorname{sign}\bigl(b_1(t) + \cdots + b_T(x)\bigr), \tag{1}$$

при условии, что она корректно классифицирует объекты обучающей выборки: $a(x_i) = y_i$. Комитет, состоящий из минимального возможного числа алгоритмов T, называется *минимальным*. В данной работе понятие комитета (1) обобщается сразу по нескольким направлениям.

Рассматриваются базовые алгоритмы $b_t \colon X \to \mathbb{R}$, которые могут выдавать как ± 1, так и числовую оценку принадлежности объекта классу $+1$. Для многих приложений гарантировать корректность не обязательно (в таких случаях комитет называется *грубым* [1]). Гораздо важнее, чтобы композиция обладала хорошей обобщающей способностью, т. е. допускала как можно меньше ошибок на независимых контрольных

данных. Минимальность T также не является жестким требованием; достаточно обеспечить значение T, близкое к минимальному. Но тогда можно отказаться от решения NP-трудной задачи построения минимального комитета [2] и применять разного рода эвристики для построения «почти минимального» грубого комитета. Такими эвристиками, по сути дела, являются бустинг [3] и бэггинг [4]. Однако в них задача минимизации T не ставится даже приближённо, что приводит к чрезмерно сложным композициям из десятков и сотен базовых алгоритмов.

Метод ComBoost

Степенью граничности или *отступом* (margin) объекта x_i называется величина $m_T(x_i) = y_i \sum_{t=1}^{T} b_t(x_i)$. Отступ отрицателен тогда и только тогда, когда $a(x)$ допускает ошибку на объекте x_i. Поэтому число ошибок композиции $a(x)$ на обучающей выборке X^ℓ определяется как

$$Q(a, X^\ell) = \sum_{i=1}^{\ell} \big[m_T(x_i) < 0\big].$$

Известно, что оптимальным с точки зрения понижения вероятности ошибки является такое распределение отступов, при котором все они принимают одинаковое и как можно большее значение [5]. В то же время, в реальных задачах всегда встречаются объекты-выбросы, которые плохо описываются как базовыми алгоритмами, так и их композицией. Ошибки на таких объектах неизбежны, а их отступы, соответственно, отрицательны. При этом заранее неизвестно, какие именно объекты являются выбросами. Изложенные соображения приводят к новому методу построения композиций, основанному на идентификации выбросов.

Базовый алгоритм b_1 строится по выборке X^ℓ некоторым стандартным методом. Рассмотрим t-й шаг, когда базовые алгоритмы $b_1, \ldots, b_{t-1}$ уже есть, и требуется построить алгоритм b_t. Упорядочим выборку X^ℓ по возрастанию отступов $m_{t-1}(x_i)$. Объекты с одинаковыми отступами договоримся располагать в случайном порядке. Обучим алгоритм b_t по подвыборке длины $(\ell_2 - \ell_1)$, с $\ell_1{+}1$-го по ℓ_2-й объекты включительно. Параметр ℓ_1 ограничивает количество выбросов; параметр ℓ_2 позволяет найти компромисс между качеством и различностью базовых алгоритмов.

Значения параметров ℓ_1, ℓ_2 могут либо фиксироваться, либо подбираться на каждом шаге t по критерию минимума числа ошибок композиции на всей выборке X^ℓ. В экспериментах второй вариант практически всегда давал лучший результат. Оптимизация ℓ_1, ℓ_2 может быть очень эффективной, если существует метод быстрой перенастройки базового алгоритма при добавлении одного объекта в обучающую выборку.

По мере увеличения числа алгоритмов в композиции повышается надёжность идентификации выбросов. При этом алгоритмы, которые были

Ошибка %	ionoshere	pima	bupa	votes
SVM	12,9	24,2	42	4,6
ComBoost$_0$[SVM]	12,6	23,1	34,2	4
ComBoost[SVM]	**12,3**	**22,5**	30,9	**3,8**
AdaBoost[SVM]	15	22,7	**30,6**	4
Parsen	6,3	25,1	41,6	6,9
ComBoost$_0$[Parsen]	6,1	25	38,1	6,8
ComBoost[Parsen]	**5,8**	**24,7**	30,6	**6,2**
AdaBoost[Parsen]	6	24,8	**30,5**	6,5

Таблица 1. Средняя частота ошибок на контроле по 50 случайным разбиениям в отношении «обучение : контроль» = 4 : 1.

обучены ранее, оказываются неоптимальными, так как их обучающие выборки содержали выбросы. Поэтому, начиная с некоторого шага t_0, перед обучением нового алгоритма «наиболее старый» удаляется.

Описанный алгоритм был назван ComBoost — *комитетный бустинг*, за очевидное сходство и с бустингом, и с методом комитетов.

Эксперименты и результаты

Исследование качества предложенного метода проводилось на реальных данных из репозитория UCI (в скобках указана длина выборки и число признаков): Ionosphere (351×34), Bupa (345×6), House-votes (435×16), Pima-diabets (768×8).

В качестве базовых алгоритмов использовались: (а) SVM с линейным ядром; (б) непараметрический байесовский классификатор с локальным оцениванием плотности по Парзену-Розенблатту, евклидовой метрикой на X и подбором ширины окна по скользящему контролю с одним отделяемым объектом (leave-one-out).

Сравнивались три метода построения композиций: ComBoost$_0$ — фиксация параметров ℓ_1, ℓ_2, подобранных для каждой задачи по критерию скользящего контроля; ComBoost — оптимизация параметров ℓ_1, ℓ_2 на каждом шаге t; AdaBoost — алгоритм бустинга [3]. Для всех трёх методов использовался один и тот же критерий останова — отсутствие существенного улучшения качества классификации обучающей выборки.

Результаты сравнения представлены в таблицах 1, 2.

Особенности ComBoost

ComBoost не накладывает ограничений ни на вид базовых алгоритмов, ни на метод их обучения. Базовые алгоритмы могут быть как бинарными (возвращать ± 1), так и вещественнозначными (возвращать числовую оценку). В отличие от бустинга, не требуется, чтобы метод обучения был чувствителен к весам объектов.

Число базовых алгоритмов	ionoshere	pima	bupa	votes
`ComBoost`$_0$`[SVM]`	**4**	**2**	**5**	**2**
`ComBoost[SVM]`	5	**2**	5	3
`AdaBoost[SVM]`	65	18	15	8
`Оценка` q	11	87	51	27

Таблица 2. Мощность композиций. Для сравнения приводится классическая верхняя оценка числа членов минимального комитета линейных алгоритмов $q = 2\left\lceil \frac{1}{n}\left\lfloor \frac{\ell-n}{2} \right\rfloor \right\rceil + 1$, где n — число признаков [6].

ComBoost демонстрирует достаточно высокую обобщающую способность даже для устойчивых базовых алгоритмов, коррекцию которых принято считать малоэффективной, в частности, для SVM.

В отличие от бустинга и бэггинга, ComBoost строит короткие композиции за счёт удаления наименее удачных алгоритмов из композиции.

Автоматически обнаруживаются объекты-выбросы. Обучающие объекты ранжируются по отступам, что позволяет выделить среди них шумовые, пограничные, типичные и эталонные объекты.

ComBoost легко обобщается на случай $|Y| > 2$, если воспользоваться понятием многоклассового геометрического отступа [7].

Работа выполнена при поддержке РФФИ, проект №05-01-00877, и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

[1] *Мазуров В. Д.* Метод комитетов в задачах оптимизации и классификации. — М.: Наука, 1990.

[2] *Хачай М. Ю.* О вычислительной сложности задачи о минимальном комитете и смежных задач // Доклады РАН. — 2006. — Т. 406, № 6. — С. 742-745.

[3] *Freund Y., Schapire R. E.* Experiments with a new boosting algorithm // International Conference on Machine Learning. — 1996. — Pp. 148–156.

[4] *Breiman L.* Bagging predictors // Machine Learning. — 1996. — Vol. 24, No. 2. — Pp. 123–140.

[5] *Mason L., Bartlett P., Baxter J.* Direct optimization of margins improves generalization in combined classifiers // Proc. of the 1998 conf. on Advances in Neural Information Processing Systems II. — MIT Press, 1999. — Pp. 288–294.

[6] *Hachai M. Y., Rybin A. I.* A new estimate of the number of members in a minimum committee of a linear inequalities system // Pattern Recognition and Image Analysis. — 1998. — Vol. 8, No. 4. — Pp. 491–496.

[7] *Allwein E. L., Schapire R. E., Singer Y.* Reducing multiclass to binary: A unifying approach for margin classifiers // 17th Int'l Conf. on Machine Learning. — Morgan Kaufmann, San Francisco, CA, 2000. — Pp. 9–16.

Комбинирование потенциальных функций в многомодальном распознавании образов

Моттль В. В., Татарчук А. И., Красоткина О. В., Сулимова В. В.

mottl@yandex.ru, aitech@yandex.ru, krasotkina@yandex.ru, vsulimova@yandex.ru

Москва, Вычислительный центр РАН, Тула, ТулГУ

Естественно, что физический объект не может быть непосредственно воспринят компьютером, поэтому в качестве посредника между объектами реального мира $\omega \in \Omega$ и процедурой анализа данных, в частности, распознавания образов, всегда выступает тот или иной конструктивный способ выражения доступной информации, который принято называть *модальностью* представления объектов. Например, набор биометрических характеристик, используемых при идентификации личности [1], как правило включает в себя фотопортрет, отпечатки пальцев, подпись, и другие. Специфика в анализе данных социологических опросов населения [2] состоит в том, что интересующие исследователя свойства людей как элементов популяции выражаются специальной формулировкой вопросов в анкете, каждый из которых определяет множество допустимых ответов.

В терминах выбранной модальности каждый объект $\omega \in \Omega$ отображается в пространство значений соответствующего *обобщенного признака* $x(\omega)\colon \Omega \to \mathbb{X}$, например, в виде сигнала, изображения, а в сравнительно простых случаях в виде действительного числа или вектора.

Невозможность обеспечить требуемое качество классификации объектов на основе какой-либо одной модальности привела к огромному разнообразию используемых способов их представления и к появлению концепции *многомодальных систем*, комбинирующих сразу все доступные представления объектов $(x_i(\omega) \in \mathbb{X}_i, i = 1, \ldots, n)$ в единой процедуре распознавания $\hat{y}\big(x_1(\omega), \ldots, x_n(\omega)\big)\colon \mathbb{X}_1 \times \cdots \times \mathbb{X}_n \to Y = \{1, \ldots, m\}$.

В таких ситуациях ключевым моментом является *уровень комбинирования модальностей*, на котором происходит слияние доступной информации перед принятием итоговых суждений о классах объектов. Наиболее естественной представляется идея формирования единого представления объектов непосредственно на *уровне сенсоров*, формирующих исходные представления объектов. Однако до последнего времени считалось, что комбинировать выходные сигналы сенсоров, часто имеющие различную физическую природу, проблематично, либо не представляется возможным в принципе, поэтому основное внимание уделялось комбинированию разнородной информации на *уровне классификаторов*, построенных независимо по каждой модальности.

Развитие беспризнаковой методологии распознавания образов [3], основанной на понятии *потенциальной функции*, позволило унифицировать представление объектов в виде элементов линейного пространства, а появление методов комбинирования нескольких разных потенциальных функций [4, 5, 6] дало возможность комбинировать модальности представления объектов фактически на уровне сигналов сенсоров.

В то же время по-прежнему остается открытым вопрос о соотношении комбинирования модальностей на уровне сенсоров и на уровне классификаторов. В работе [7] проведено исследование условий, при которых один из известных принципов комбинирования классификаторов получается как частный случай комбинирования потенциальных функций.

Специфика потенциальной функции $K(x', x'')\colon \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, заданной на множестве значений некоторого обобщенного признака $x(\omega) \in \mathbb{X}$, заключается в том, что ее значения могут интерпретироваться непосредственно как скалярное произведение объектов $x' = x(\omega')$, $x'' = x(\omega'')$ в воображаемом линейном пространстве $\tilde{\mathbb{X}} \supseteq \mathbb{X}$, в которое данная потенциальная функция погружает шкалу значений выходного сигнала соответствующего сенсора, минуя промежуточное понятие вектора числовых признаков.

Задав, как минимум, по одной потенциальной функции для каждой модальности $K_i(x'_i, x''_i)$, $x'_i, x''_i \in \mathbb{X}_i$, $i = 1, \ldots, n$, удобно рассматривать порождаемые этими функциями линейные пространства совместно как декартово произведение $\tilde{\mathbb{X}} = \tilde{\mathbb{X}}_1 \times \cdots \times \tilde{\mathbb{X}}_n$ с комбинированной потенциальной функцией $K(\mathbf{x}', \mathbf{x}'')$, $\mathbf{x}' = (x'_1, \ldots, x'_n)$, $\mathbf{x}'' \in \tilde{\mathbb{X}}$ в роли скалярного произведения. При таком подходе основная нагрузка по интерпретации исходных представлений объектов перекладывается на этап формирования потенциальных функций, но в данной работе этот аспект не рассматривается.

Большинство методов комбинирования потенциальных функций основано на формировании результирующей потенциальной функции как линейной комбинации исходных функций $K(\mathbf{x}', \mathbf{x}'') = \sum_{i=1}^{n} \alpha_i K(x'_i, x''_i)$ с коэффициентами $\alpha_i \geqslant 0$. Очевидно, что если некоторый коэффициент равен нулю, то соответствующая модальность не будет участвовать в принятии решений. Такой подход означает замену дискретной задачи перебора подмножеств модальностей с целью выбора подмножества, наиболее адекватного анализируемому массиву данных $\hat{I} \subseteq I = \{1, \ldots, n\}$, на непрерывную задачу поиска весов $\hat{I} = \{i \in I\colon \alpha_i > 0\}$. Специфика конкретного метода комбинирования заключена в выборе критерия оптимизации.

В работе [4] был предложен *метод опорных потенциальных функций* (Support Kernel Machines или SKM), особенность которого состоит

в стремлении процедуры комбинирования выбирать нулевые веса $\alpha_i = 0$ для большинства функций и оставлять ненулевыми $\alpha_i > 0$ только веса для активных (опорных) функций. Однако, такая стратегия обучения приводит к двойственной оптимизационной задаче квадратичного программирования при квадратичных ограничениях, решение которой вычислительно существенно сложнее стандартной задачи квадратичного программирования при линейных ограничениях, лежащей в основе классического метода опорных векторов [3].

Другая стратегия комбинирования [5] приводит к итерационной процедуре вычисления весов $(\alpha_1, \ldots, \alpha_n)$, но в этом случае веса неинформативных потенциальных функций лишь стремятся к нулю, но никогда его не достигают. Будем называть такую стратегию комбинирования *методом релевантных потенциальных функций* (Relevance Kernel Machines или RKM) по аналогии с методом релевантных векторов, предложенным в работе [8] для поиска линейных дискриминантных функций как альтернатива классическому методу опорных векторов.

Вообще говоря, методология обучения распознаванию объектов двух классов на основе концепции оптимальной линейной дискриминантной функции не предусматривает принятия каких-либо предположений о вероятностной модели генеральной совокупности. Такой концепции адекватна качественная модель в виде в линейного пространства признаков, в котором объективно существует некоторая гиперплоскость, такая, что объекты двух разных классов отображаются, в основном, по разные стороны от нее, без количественного уточнения, в какой степени это предположение может нарушаться. При этом выбор значений направляющего элемента $\hat{\vartheta} \in \tilde{\mathbb{X}}$ и порога $\hat{b} \in \mathbb{R}$ линейной дискриминантной функции

$$f(\mathbf{x}(\omega) \,|\, \vartheta, b) = K(\vartheta, \mathbf{x}(\omega)) + b \begin{cases} < 0 \to y(\mathbf{x}(\omega)) = -1 \\ > 0 \to y(\mathbf{x}(\omega)) = 1 \end{cases}$$

полностью задает некоторую классификацию множества объектов $\omega \in \Omega$ в комбинированном линейном пространстве обобщенных признаков $\tilde{\mathbb{X}} = \tilde{\mathbb{X}}_1 \times \cdots \times \tilde{\mathbb{X}}_n$ со скалярным произведением $K(\mathbf{x}', \mathbf{x}'')$, $\mathbf{x}', \mathbf{x}'' \in \tilde{\mathbb{X}}$.

В данной работе предлагается *квази-статистический подход* к решению задачи комбинирования потенциальных функций, который полностью охватывает методы опорных и релевантных потенциальных функций. В качестве модели генеральной совокупности объектов предлагается рассматривать два параметрических семейства плотностей распределения $\varphi_{-1}(\mathbf{x}(\omega) \,|\, \vartheta, b)$ и $\varphi_1(\mathbf{x}(\omega) \,|\, \vartheta, b)$, связанных с двумя классами объектов и сконцентрированных преимущественно по разные стороны гиперплоскости $f(\mathbf{x}(\omega) \,|\, \vartheta, b) = 0$. Однако невозможно выбрать эти плотности

распределения так, чтобы они количественно отражали возможность попадания объектов за пределы своего класса и не вносили бы при этом иной информации о значениях признака, поскольку такое требование связано с предположением о равномерном распределении в бесконечных областях пространства, что приводит к равенству нулю соответствующей плотности. В книге [9] рекомендуется задавать «безразличное» априорное распределение в виде ненулевой функции-константы, несмотря на то, что ее интеграл по бесконечной области не существует. Функции такого вида были названы *несобственными плотностями распределениями*.

В основу предлагаемого подхода положен принцип максимизации плотности апостериорного распределения в пространстве параметров модели генеральной совокупности, который приводит к стандартному байесовскому правилу обучения.

Предположение о нормальном распределении направляющего вектора $\vartheta = (\vartheta_1, \ldots, \vartheta_n) \in \tilde{\mathbb{X}}$ с независимыми компонентами немедленно приводит к критерию обучения, полностью идентичному критерию по методу опорных потенциальных функций. В то же время, принятие в качестве априорного распределения компонент направляющего вектора распределения Лапласа приводит к методу релевантных потенциальных функций.

Работа выполнена при поддержке INTAS, проект №04-77-7347, и РФФИ, проекты №05-01-00679, №06-01-08042, №06-07-89249.

Литература

[1] *Ross A., Jain A.* Multimodal biometrics: An overview // 12th European Signal Processing Conference, Vienna, Austria, 2004. — С. 1221.

[2] *Галицкий Е. Б., Моттль В. В., Татарчук А. И.* Обучение распознаванию образов в анализе данных опросов населения. // ММРО-12, Москва, 2005.

[3] *Vapnik V.* Statistical Learning Theory. John-Wiley & Sons, Inc. 1998.

[4] *Bach F. R., Lankriet G. R. G., Jordan M. I.* Multiple kernel learning, conic duality, and the SMO algorithm // 21th International Conference on Machine Learning, Banff, Canada, 2004.

[5] *Sonnenburg S., Rätsch G., Schäfer C.* A general and efficient multiple kernel learning algorithm // 19th Annual Conference on Neural Information Processing Systems, Vancouver, Canada, 2005.

[6] *Mottl V., Krasotkina O., Seredin O., Muchnik I.* Kernel fusion and feature selection in machine learning // 8th IASTED International Conference on Intelligent Systems and Control, Cambridge, USA, 2005.

[7] *Татарчук А. И., Елисеев А. П., Моттль В. В.* Комбинирование классификаторов и потенциальных функций в многомодальном распознавании образов // ММРО-13 (в настоящем сборнике). — 2007. — С. 220–222.

[8] *Bishop C. M., Tipping M. E.* Variational relevance vector machines // 16th Conf. on Uncertainty in Artificial Intelligence, Morgan Kaufmann, 2000. — Pp. 46–53.

[9] *Де Гроот М.* Оптимальные статистические решения. — Москва: Мир, 1974.

Применение методов распознавания образов к исследованию динамических систем

Неймарк Ю. И.

neymark@pmk.unn.ru

Нижний Новгород, НИИПМК ННГУ

Решение и исследование решений систем дифференциальных уравнений — одна из основных задач, возникающих в самых разнообразных приложениях и науке. Вместе с тем, исследование сколько-нибудь сложной многомерной динамической системы, описываемой обыкновенными дифференциальными уравнениями, до появления компьютеров было неразрешимой задачей, а сегодня стало возможным, но требует квалифицированной, длительной и трудоемкой работы. Подчас время и трудности исследования настолько велики, что исследование практически невыполнимо, прежде всего, в силу «проклятия размерности» как фазового пространства, так и пространства параметров. В наших работах была высказана идея автоматизации огрублённого численного исследования динамической системы на основе использования методов распознавания образов и статистического моделирования. При этом образы — это фазовые траектории, установившиеся движения и их области притяжения, а статистический подход позволяет преодолеть проклятие размерности, поскольку при этом объем необходимых вычислений мало зависит от размерности. Ниже рассказывается, что удалось осуществить на этом пути.

Постановка задачи

Прежде всего, ясно, что, несмотря на теоретическую возможность бесконечного размера фазового пространства, численное исследование осуществимо только в конечной его области. Эта конечная область должна быть указана исходя из реальной задачи. Далее в ней может быть выделена часть, которую фазовые траектории не покидают. Следующий этап состоит в алгоритмизации численного отыскания и описания установившихся движений: устойчивых равновесий, периодических движений и притягивающих, в целом неустойчивых, хаотических движений. Затем находятся их области притяжения, точнее, их некоторые части, прилегающие к соответствующему установившемуся движению. При этом в отношении областей притяжения может быть указана статистическая достоверность полученных результатов. Эти вероятности с ростом используемого времени численных расчетов приближаются к единице. Иссле-

дование не встречает затруднений, если исследуемая система достаточно грубая (под грубостью имеется в виду грубость отыскиваемой упрощенной структуры фазового пространства динамической системы).

Основные этапы исследования конкретной динамической системы методами распознавания образов

Численное исследование конкретной динамической системы состоит в построении огрубленного компьютерного фазового портрета и сводится к последовательному решению целого ряда задач анализа и распознавания данных, главными из которых являются распознавание фазовых траекторий и установившихся движений (аттракторов), а также определение областей притяжения для каждого из аттракторов. Основные подходы к решению этих задач базируются на исследовании одномерных временных рядов, синдромальном анализе данных и использовании универсальной рекуррентной формы метода наименьших квадратов — методов, обладающих широкими адаптивными возможностями по отношению к изменяющейся исследуемой выборке данных, что позволяет решать задачи распознавания с активным экспериментом. Ограничимся кратким описанием путей решения каждой из поставленных задач с помощью одного из предложенных методов.

Распознавание фазовых траекторий и установившихся движений на базе одномерных временных рядов

В результате исследований была установлена возможность решения этой задачи на множестве признаков, описывающих поведение двух одномерных временных рядов $\mathbf{y}_1(t)$ и $\mathbf{y}_2(t)$, которые ставятся в соответствие каждой траектории $\mathbf{x}(t)$: ряд $\mathbf{y}_1(t)$ описывает поведение траектории при приближении её к аттрактору, и представляет собой изменение со временем расстояний между её соседними точками в некоторой заданной метрике, а ряд $\mathbf{y}_2(t)$ описывает устойчивость исследуемой траектории, и представляет собой изменение со временем расстояний между двумя траекториями с начальными условиями из малой окрестности друг друга.

Решающее правило для распознавания типа фазовых траекторий приведено в наших работах. Оно позволяет определять траектории, стремящиеся к состоянию равновесия; траектории, стремящиеся к предельному циклу; траектории, представляющие хаотические или стохастические аттракторы. Кроме того, существуют признаки для определения типа состояния равновесия (узел, фокус), признаки наличия многообразия состояний равновесия или седловых точек, признаки дискриминации различного вида хаотических движений и др. Анализ траекторий с помощью одномерных временных рядов полностью алгоритмизирован

и может быть проведен в автоматическом режиме вплоть до принятия решения о необходимости пополнения существующей базы знаний для анализа результатов в некоторой определенной области фазового пространства.

Поиск областей притяжения с использованием синдромов

Для получения областей притяжения аттракторов эффективным инструментом являются два алгоритма на основе абсолютных синдромов — n-мерных параллелепипедов, внутри и на поверхности которых располагаются объекты какого-то одного и только этого класса. Первый алгоритм эффективно уменьшает объем обучающей выборки каждого из выделенных аттракторов практически без потери информации (в единицы, десятки, сотни раз, в зависимости от типа аттрактора) и формирует выборку для построения разделяющего решающего правила областей притяжения аттракторов на основе оптимальных синдромов. Второй алгоритм строит упомянутое разделяющее правило на основе выборки, сформированной первым алгоритмом. В некоторых случаях из-за сложности границ областей притяжения решающее правило может содержать достаточно большое число синдромов. Однако всегда можно ограничиться небольшим числом наиболее представительных синдромов для каждого из выделенных аттракторов. Такое описание области притяжения делает его доступным для понимания исследователем.

Заключение

Изложенные методы использовались для исследования конкретных математических моделей как известных, так и новых задач, среди которых четырехмерная модель с 14 параметрами, описывающая динамику ответной реакции организма на вторжение инфекции, семимерная дискретная модель турбулентности, известная система Лоренца и др. Естественным продолжением описанного исследования является изучение бифуркаций в виде одномерных и двумерных бифуркационных портретов.

Работа выполнена при поддержке РФФИ, проект № 05-01-00391.

Анализ фазовых траекторий многомерных динамических систем методами распознавания на основе одномерных временных рядов

Неймарк Ю. И., Теклина Л. Г.
neymark@pmk.unn.runnet.ru
Нижний Новгород, НИИ прикладной математики и кибернетики
Нижегородского государственного университета

Качественное исследование динамических систем, заданных системами дифференциальных уравнений достаточно высокого порядка — это, на первый взгляд, чисто научная математическая проблема, но с выходом на аналитические и числовые результаты, имеющие большое практическое значение для физики и техники. В настоящее время исследование конкретной динамической системы порядка $n \geqslant 3$ весьма трудоемко, требует нестандартного подхода и основывается на интуиции исследователя и анализе особенностей рассматриваемой системы. Для формализации, а, в конечном счете, и автоматизации процесса исследования динамических систем, предлагается новый подход, основанный на решении задачи построения фазового портрета динамической системы методами распознавания образов путем извлечения знаний из рассмотрения фактических данных, получаемых в ходе активного эксперимента. Этот подход представлен в работе [1], а в настоящем докладе рассматриваются возможности его реализации с помощью адаптивных методов анализа, описания и распознавания динамически изменяющихся объектов на основе универсальной рекуррентной формы метода наименьших квадратов [2].

Цели и задачи анализа

Решение проблемы исследования динамических систем методами распознавания требует решения задачи распознавания различных типов фазовых траекторий. Для диссипативных систем это: траектории, стремящиеся к состоянию равновесия, траектории, стремящиеся к предельному циклу, и хаотические и стохастические движения. Для сведения задачи определения вида фазовой траектории к классической задаче распознавания с учителем путем формирования единого пространства информативных признаков, описывающих поведение траектории любой длительности в фазовом пространстве любой размерности, и был проведен анализ различного типа фазовых траекторий для разных и по своей природе, и по размерности динамических систем. Обучающая выборка $\mathbf{X}$ состояла из отрезков фазовых траекторий — кривых, представляющих собой решение нескольких многомерных систем дифференциальных уравнений разного порядка при различных начальных условиях, и заданных значениями своих координат в фазовом пространстве в последовательные

моменты времени с постоянным для данной кривой шагом дискретизации Δt. Таким образом, обучающая выборка представляла собой множество конечных многомерных временных рядов разной размерности и длительности: $\mathbf{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$, где $\mathbf{x}^i = (x^i_{j1}, \ldots, x^i_{jn_i})$, $j = 1, \ldots, M_i$, n_i — размерность i-ой траектории, M_i — длина соответствующего ей временного ряда.

Проведенный анализ включал в себя:

— определение длительности переходного периода до попадания траектории в зону аттрактора;
— описание особенностей поведения фазовой траектории как многомерного временного ряда при приближении к аттрактору;
— исследование устойчивости фазовых траекторий;
— описание области локализации траектории в фазовом пространстве при приближении ее к аттрактору.

Распознавание вида фазовых траекторий

В результате исследований была установлена возможность решения задачи распознавания различных типов фазовых траекторий на множестве признаков, описывающих поведение одномерных временных рядов, представляющих собой изменение со временем расстояний между соседними точками исходного временного ряда:

$$\mathbf{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\} \Rightarrow \mathbf{Y} = \{\mathbf{y}^1, \ldots, \mathbf{y}^N\},$$

где $\mathbf{y}^\nu = (y^\nu_1, \ldots, y^\nu_{M_\nu - 1})$ и $y^\nu_l = \sqrt{\sum_{j=1}^{n_\nu} \left(x^\nu_{lj} - x^\nu_{l+1,j}\right)^2}$.

После перехода от многомерного временного ряда $\mathbf{x}$ к одномерному временному ряду $\mathbf{y}$ была решена первая из всего множества задач, а именно — проблема предварительного (грубого) определения длительности переходного периода t^*, когда при $t < t^*$ временной ряд изменяется произвольно, а при $t > t^*$ характеристики ряда (среднее, дисперсия, коэффициенты авторегрессии и др.) приобретают четкие специфические свойства, что, предположительно, свидетельствует о том, что траектория попадает в зону аттрактора. В дальнейшем при расширении знаний об исследуемой системе эта величина при необходимости корректируется.

После определения t^* для $\mathbf{x}(t)$ при $t > t^*$ строятся два одномерных временных ряда $\mathbf{y}_1(t)$ (для описания поведения траекторий при приближении их к аттрактору) и $\mathbf{y}_2(t)$ (для исследования на устойчивость), где $\mathbf{y}_1(t)$ — это ряд $\mathbf{y}(t)$ при $t > t^*$, а $\mathbf{y}_2(t) = |\mathbf{x}(t) - \tilde{\mathbf{x}}(t)|$, где $\tilde{\mathbf{x}}(t)$ — траектория с начальными условиями из малой окрестности траектории $\mathbf{x}(t)$ при $t = t^*$.

Результатом анализа методами распознавания расширенной обучающей выборки, в которой каждая траектория $\mathbf{x}^i(t)$ описывалась двумя соответствующими ей одномерными временными рядами $\mathbf{y}_1^i(t)$ и $\mathbf{y}_2^i(t)$, стало выделение ряда признаков, информативных для решения задачи дискриминации различных типов траекторий, а именно:

1. Поведение ряда $\mathbf{y}_1(t)$ при $t \to \infty$.
2. Периодичность ряда $\mathbf{y}_1(t)$. Исследование на периодичность и отыскание периода проводится путем скольжения отрезка ряда $\mathbf{y}_0$ длительности τ по ряду $\mathbf{y}_1(t)$ с определением степени близости $\delta(t) = \sum_{j=0}^{\tau/\Delta t} \big(\mathbf{y}_1(t + j\Delta t) - \mathbf{y}_0(j\Delta t)\big)^2$ при различных значениях t.
3. Поведение ряда $\mathbf{y}_2(t)$ при $t \to \infty$.

На базе этих признаков построено решающее правило для распознавания типа фазовых траекторий:

1. Траектория стремится к состоянию равновесия, если $\lim\limits_{t\to\infty} \mathbf{y}_1(t) = 0$ и $\lim\limits_{t\to\infty} \mathbf{y}_2(t) = 0$. Причем по характеру изменения $\mathbf{y}_1(t)$ можно определить тип состояния равновесия: в узле, начиная с некоторого t, $\mathbf{y}_1(t)$ монотонно убывает, а в фокусе $\mathbf{y}_1(t)$ совершает колебательные движения с уменьшающейся амплитудой.
2. Если $\lim\limits_{t\to\infty} \mathbf{y}_1(t) = 0$, но $\lim\limits_{t\to\infty} \mathbf{y}_2(t) = A \neq 0$, причем величина A зависит от расстояния $|\mathbf{x}(t^*) - \tilde{\mathbf{x}}(t^*)|$ (начальные условия для $\tilde{\mathbf{x}}(t)$), то имеет место многообразие состояний равновесия.
3. О наличии предельного цикла свидетельствуют периодичность временного ряда $\mathbf{y}_1(t)$ и выполнение условия $\lim\limits_{t\to\infty} \mathbf{y}_2(t) = 0$.
4. Если и $\lim\limits_{t\to\infty} \mathbf{y}_1(t)$, и $\lim\limits_{t\to\infty} \mathbf{y}_2(t)$ не существуют, то траектория представляет собой хаотическое или стохастическое движение.

Приведенное правило не исчерпывает возможностей исследования фазовых траекторий с помощью одномерных временных рядов. В частности, существуют признаки наличия седловых точек, признаки дискриминации различного вида хаотических движений и др.

Работа выполнена при поддержке РФФИ, проект № 05-01-00391.

Литература

[1] *Неймарк Ю. И., Котельников И. В., Теклина Л. Г.* Исследование структуры фазового пространства динамической системы как задача распознавания образов // Докл. конф. ММРО-12. — М.: Макспресс, 2005. — С. 177–180.

[2] *Неймарк Ю. И., Теклина Л. Г.* Новые технологии применения метода наименьших квадратов. — Нижний Новгород: Изд. Нижегородского госуниверситета, 2003. — 196 с.

Планирование эксперимента при исследовании конкретных динамических систем методами распознавания образов

Неймарк Ю. И., Теклина Л. Г.
neymark@pmk.unn.ru
Нижний Новгород, НИИПМК ННГУ

Теория динамических систем и дифференциальных уравнений является основой современной науки, а исследование динамических систем — это один из путей познания окружающего нас мира и совершенствования современной техники. Теория динамических систем, возникшая в трудах великих ученых А. Пуанкаре и Д. Биркгофа, получила мощное развитие, однако успехи ее в исследовании конкретных динамических систем достаточно скромны. Конечно, с того времени математическая теория динамических систем существенно расширилась, но ее возможности по-прежнему отстают от современных потребностей. Основные трудности связаны с «проклятием размерности» как фазового пространства, так и пространства параметров. Для преодоления их в исследовании конкретных динамических систем предлагается использовать методы теории распознавания образов, основанные на извлечении знаний из рассмотрения множества фактических данных [1]. Отличительная особенность этой задачи распознавания образов состоит в том, что это — задача с активным экспериментом, причем возможности проведения эксперимента практически не ограничены и не требуют ни больших временных, ни больших материальных затрат. И, как всякая задача с активным экспериментом, она требует планирования эксперимента. Планирование эксперимента предусматривает выбор начальных условий для построения траектории, а также выбор таких ее характеристик, как шаг дискретизации, длительность и точность счета. Цель планирования определяется конечной целью исследования и состоит в обеспечении высокой точности результатов: определение и описание установившихся движений и их областей притяжения.

Рассмотрим особенности планирования эксперимента на разных этапах исследования динамических систем методами распознавания.

Этап I. Формирование первичной обучающей выборки данных

Решение задачи начинается с создания первичной обучающей выборки Θ. Такая выборка формируется случайным выбором начальных условий в заданной ограниченной области фазового пространства, но с покрытием исследуемой области подобластями, занятыми уже построенными траекториями, так, чтобы вероятность попадания начальных условий в незанятую область была меньше заданной величины ε. Таким образом,

создается случайная выборка большого объема, но с ограниченными требованиями к длительности и точности счета, что облегчает и получение, и обработку больших массивов данных.

Этап II. Предварительный анализ данных

На базе выборки Θ проводится предварительный анализ данных с целью идентификации устойчивых подмножеств фазового пространства — аттракторов (состояний равновесия, предельных циклов, хаотических аттракторов). Результат анализа — предварительные данные о виде и числе аттракторов и множествах представляющих их траекторий. Это — отправная точка для построения компьютерного фазового портрета.

Этап III. Уточнение структуры фазового пространства

Задачи этапа III ставятся и решаются как задачи распознавания:

— уточнение вида и числа аттракторов;
— описание аттракторов и разделение их в фазовом пространстве;
— построение решающего правила для принятия решения о принадлежности произвольной траектории к определенному аттрактору;
— выделение областей притяжения для каждого из аттракторов.

Каждая из задач решается последовательно друг за другом в адаптивном режиме по следующей схеме:

1. Решение задачи распознавания или классификации.
2. Анализ результатов решения. Под анализом результатов подразумевается либо анализ кластеров на их взаимное расположение и мощность представляющих их множеств траекторий в задаче классификации, либо анализ результатов распознавания на контрольной выборке для задач распознавания с учителем. Результат такого анализа — вывод об окончании (переход к 5) или продолжении решения задачи распознавания (переход к 3).
3. Планирование и проведение эксперимента.
4. Уточнение решения задачи с переходом на 2.
5. Переход к следующей задаче.

Адаптивному методу решения задачи исследования структуры фазового пространства динамической системы соответствует и метод планирования эксперимента, а именно: последовательное планирование, когда каждый новый эксперимент проводится с учетом полученных результатов и текущих оценок. Отличительная особенность решения задач на этом этапе в сравнении с первым — целенаправленное формирование обучающей выборки данных. Планирование и проведение эксперимента проводится в зависимости от результатов анализа данных и распознавания областей, определяющих структуру фазового пространства.

Цель планирования — достижение высокой точности решения поставленных задач распознавания, а основной принцип планирования: «Добавляй информацию там, где ее не хватает». Изменение обучающей выборки проводится двумя путями: либо путем коррекции ограниченного числа данных из выборки, либо путем пополнения ее новыми данными. Коррекция данных осуществляется путем усиления требований к проведению эксперимента как по точности счета, так и по шагу дискретизации и длительности фазовой траектории. Цель коррекции — уточнение описания аттракторов. Пополнение выборки новыми данными предполагает либо получение траекторий с заданными свойствами (например, для аттракторов с малой мощностью представляющих их множеств проводится вычисление траекторий в обратном направлении), либо построение траекторий с начальными условиями из заданных областей (например, при построении решающего правила распознавания обучающая выборка пополняется траекториями из областей, где точность распознавания недостаточно высока). В частности, для каждой задачи распознавания проводится анализ результатов решения на контрольной выборке, формируемой случайным выбором начальных условий в заданной области фазового пространства. Все данные из контрольной выборки, на которых получены ошибочные ответы, пополняют обучающую выборку.

Таким образом, предварительные оценки и решающие правила, полученные на основе случайной выборки большого объема из траекторий относительно небольшой длительности, с достаточно большим шагом дискретизации и невысокой точностью счета, в дальнейшем корректируются путем уточнения характеристик для небольшого числа траекторий и пополнения обучающей последовательности новыми данными с целенаправленным выбором для получения более точного описания аттракторов исследуемой динамической системы и их областей притяжения.

Этап IV. Построение компьютерного фазового портрета

На базе уточненных данных о виде и числе аттракторов и решающих правил распознавания аттракторов и их областей притяжения строится компьютерный фазовый портрет исследуемой динамической системы.

Работа выполнена при поддержке РФФИ, проект № 05-01-00391.

Литература

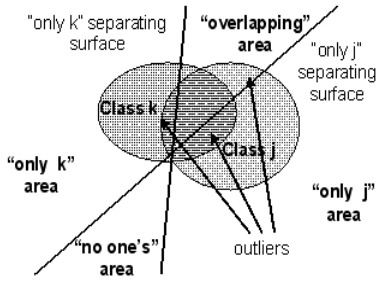
[1] *Неймарк Ю. И., Котельников И. В., Теклина Л. Г.* Исследование структуры фазового пространства динамической системы как задача распознавания образов // всеросс. конф. ММРО-12. — М.: МаксПресс, 2005. — С. 177–180.

Метод многотемной (multi-label) классификации на основе попарных сравнений с отсечением наименее релевантных классов

Петровский М. И., Глазкова В. В.
Москва, МГУ им. М. В. Ломоносова

Расширением традиционной постановки задачи классификации с несколькими классами (multi-class) является постановка задачи *много-темной* (multi-label) классификации, в которой классифицируемый объект может принадлежать нескольким классам одновременно, и сами классы являются не взаимоисключающими (возможно, даже вложенными). Задачи такого типа возникают при анализе, рубрикации и классификации текстов, изображений, цепочек ДНК, и т. д. В случае multi-label классификации в исходной обучающей совокупности $S = (x_i, y_i)_{i=1}^n$ для каждого примера x_i задан не единственный класс, а множество релевантных классов $y_i \subset \{1, \ldots, q\}$, и целью алгоритма машинного обучения является построение классификатора $f_S \colon X \to 2^q$, предсказывающего все релевантные классы, где X — исходное пространство признаков, а q — число классов. Традиционным подходом к решению таких задач является декомпозиция типа «*каждый против остальных*» исходной multi-label задачи в q задач бинарной классификации с целью независимо определить, является ли каждый из q классов релевантным x. Для каждого класса l формируется тренировочный набор, где x_i помечен как «положительный», если в S выполнялось $l \in y_i$, иначе x_i помечен как «отрицательный». Далее строится q бинарных классификаторов $f_S^l \colon X \to \{0, 1\}$. Этот подход критикуется за низкую точность, поскольку не учитываются корреляции между классами, и за ресурсоемкость, поскольку при обучении необходимо решать q задач размера n вместо одной.

В настоящей работе исследуется возможность использования для решения задачи multi-label классификации подхода на основе декомпозиции типа «*каждый против каждого*». Предлагается новый алгоритм, основанный на модифицированном *методе попарных сравнений* с помощью набора *бинарных классификаторов*. Следует отметить, что ранее этот подход не давал позитивных результатов для задач multi-label классификации, поскольку не удавалось построить точный классификатор, разделяющий два существенно перекрывающихся класса. Поэтому в нашем методе каждая пара возможно классов j и k разделяется с помощью двух бинарных классификаторов. Используя их, можно выделить четыре области: область «только класса j»; область «только класса k»; «перекрывающаяся область» (j и k); область, не принадлежащую «ни классу j, ни классу k»:



Для каждой пары классов j и k формулируется две задачи обучения, которые включают только примеры, помеченные в S либо j, либо k, либо обоими классами одновременно (число таких примеров, как правило, существенно меньше n). В первой подзадаче примеры, помеченные только классом k, рассматриваются как «положительные», все остальные — как «отрицательные». В результате, построенный классификатор предсказывает вероятность: $r^+_{kj}(x) = P(k \in f_S(x) \wedge j \notin f_S(x) | x \in k \cup j)$ и дополнительную вероятность: $r^-_{kj}(x) = 1 - r^+_{kj}(x) = P(j \in f_S(x) | x \in k \cup j)$. Вторая подзадача формулируется и решается аналогично, но для класса j: $r^+_{jk}(x) = P(j \in f_S(x) \wedge k \notin f_S(x) | x \in k \cup j)$ и $r^-_{jk}(x) = P(k \in f_S(x) | x \in k \cup j)$. Вероятности принадлежности x каждой из областей вычисляются так:

$$\begin{aligned} P(x \in \text{ overlapping}\, k, j) &= r^-_{kj}(x)\, r^-_{jk}(x); \\ P(x \in \text{ only}\, k) &= r^+_{kj}(x)\, r^-_{jk}(x); \\ P(x \in \text{ no one's}\, k, j) &= r^+_{kj}(x)\, r^+_{jk}(x); \\ P(x \in \text{ only } j) &= r^-_{kj}(x)\, r^+_{jk}(x). \end{aligned}$$

Важно отметить, что при такой формулировке оба бинарных классификатора разделяют взаимно исключающие суперклассы, и для решения и оценки попарных вероятностей могут быть использованы стандартные алгоритмы бинарной классификации, например, на основе Support Vector Machines или Kernel Fisher Discriminant. Формулируя и решая таким образом $q(q-1)$ задач бинарной классификации, каждая размера меньшего, чем n, мы получаем попарные вероятности сравнения $r^*_{jk}(x)$.

Далее необходимо, используя результаты попарных сравнений, оценить вероятности принадлежности $p_l(x)$ каждому из q невзаимоисключающих классов. Для этого мы предлагаем использовать обобщённую *модель ранжирования Бредли-Терри с «ничьёй»*, которую сформулировали Рао и Купер (1967). Учитывая, что в нашем случае вероятность «ничьи»

$$P(j \text{ ties } k) = P(x \in \text{ overlapping}\, k, j) + P(x \in \text{ no one's } k, j),$$

$P(k\,\text{beats}\,j) = P(x \in \text{only } k)$, $P(j\,\text{beats}\,k) = P(x \in \text{only } j)$, получаем оптимизационную задачу:

$$\min_{\bar{p},\theta} l(\bar{p},\theta), \quad p_k \geqslant 0;$$

$$l(\bar{p},\theta) = -\frac{1}{2}\sum_k\sum_j\left[2r^+_{kj}r^-_{jk}\ln\frac{p_k}{p_k+\theta p_j} + \right.$$
$$\left. + (r^-_{kj}r^-_{jk} + r^+_{kj}r^+_{jk})\ln\frac{(\theta^2-1)p_jp_k}{(\theta p_k+p_j)(p_k+\theta p_j)}\right].$$

Для решения данной задачи можно использовать итеративный minorization-maximization алгоритм [1], детально описанный для нашего случая в [2].

Решив сформулированную задачу оптимизации, мы получаем оценки релевантности классов $p_l(x)$, и теперь должны выбрать наиболее релевантные классы в качестве решения исходной multi-label задачи: $\{l|p_l(x) > t, l \in \{1,\ldots,q\}\}$. Порог отсечения t обычно зависит от классифицируемых объектов, поэтому возникает необходимость построения пороговой функции $t(x)$. Для этих целей используется подход, предложенный нами в [3] для сведения задачи ранжирования к задаче multi-label классификации. В существующих методах сложные пороговые функции строятся в пространстве признаков X, в нашем подходе предлагается строить простые линейные пороговые функции в пространстве релевантностей классов, используя результат работы алгоритмов ранжирования как новое множество признаков анализируемого объекта:

$$t(\bar{p}(x)) = \sum_{j=1}^{q} a_j p_j(x_i) + a_0.$$

Коэффициенты a_j определяются методом наименьших квадратов на примерах из обучающей совокупности. Детально подход описан в [3].

Все разработанные алгоритмы были экспериментально проверены на эталонных тестовых наборах данных Yeast 2K и Reuters-2000, где показали высокие результаты как по точности классификации так и скорости обучения [2, 3].

Работа выполнена при поддержке РФФИ, проект № 06-01-00691, гранта Президента РФ МК-4264.2007.9, а также в рамках госконтракта с Федеральным агенством по науке и инновациям № 02.514.11.4026.

Литература

[1] *Hunter D. R.* MM algorithms for generalized Bradley-Terry models. — Annals of Statistics, 32(1), 2004. — pp. 384–406.

[2] *Petrovskiy M.* Paired Comparisons Method for Solving Multi-label Learning Problem. — Hybrid Intelligent Systems, IEEE Press, 2006. — pp. 42–48.

[3] *Petrovskiy M., Glazkova V.* Linear Methods for Reduction from Ranking to Multilabel Classification. — Springer-Verlag, 2006. — LNAI, vol. 4304 — pp. 1152–1156.

Обучение композиций дипольных классификаторов на основе EM-алгоритма

Пустовойтов Н. Ю.
pustovoytov@forecsys.ru
Москва, МФТИ, ЗАО «Форексис»

Идея дипольного классификатора, предложенная в [2], так же, как и идея метода потенциальных функций [1], заимствована из физики. Диполем называется пара зарядов, имеющих одинаковый по модулю, но противоположный по знаку заряд, и расположенных на незначительном расстоянии друг от друга. Диполь разделяет пространство плоскостью на области положительного и отрицательного потенциала. Поэтому его можно рассматривать как линейный классификатор, заданный двумя точками (полюсами диполя), либо как алгоритм ближайшего соседа, построенный по выборке из двух объектов (полюсов). В данной работе рассматриваются композиции дипольных классификаторов.

Дипольный классификатор

Пусть $X \subset \mathbb{R}^n$ — множество *объектов*, $Y = \{-1, +1\}$ — множество *меток классов*. Требуется построить алгоритм классификации $a\colon X \to Y$, восстанавливающий неизвестную *целевую зависимость* $y^*\colon X \to Y$ по обучающей выборке $X^\ell = \{x_1, \ldots, x_\ell\}$, $y_i = y^*(x_i)$. Предполагается, что на множестве объектов определена функция расстояния $\rho(x, x')$.

Пусть $K\colon \mathbb{R}_+ \to \mathbb{R}_+$ — заданная невозрастающая функция.

Дипольным классификатором (или просто *диполем*) с *ядром* K, *полюсами* r^-, r^+ и *радиусом* d назовём функцию $a\colon X \to Y$ вида

$$a(x) = \operatorname{sign}\left(K\left(\frac{\rho(x, r^+)}{d}\right) - K\left(\frac{\rho(x, r^-)}{d}\right)\right). \qquad (1)$$

Диполь совпадает с байесовским классификатором, если предположить, что плотности классов имеют вид $p(x \,|\, y) = p(x; r^y, d) = \frac{1}{K_0} K\left(\frac{\rho(x, r^y)}{d}\right)$, где K_0 — нормировочный коэффициент; r^-, r^+, d — параметры диполя.

Введём также неотрицательную *функцию компетентности* диполя $C(x; r^-, r^+, R)$, значение которой непрерывно и монотонно убывает по мере удаления от полюсов диполя, а скорость убывания задаётся *радиусом*

функции компетентности R, вообще говоря, отличным от радиуса диполя d. Будем рассматривать только такие функции компетентности, которые симметричны относительно полюсов диполя. Существует три принципиально различных варианта задания таких функций:

$$C(x; r^-, r^+, R) = K'\left(\tfrac{\rho(x,r^+)}{R}\right) K'\left(\tfrac{\rho(x,r^-)}{R}\right); \tag{2}$$

$$C(x; r^-, r^+, R) = K'\left(\tfrac{\rho(x,r^+)+\rho(x,r^-)}{2R}\right); \tag{3}$$

$$C(x; r^-, r^+, R) = K'\left(\tfrac{1}{R}\rho\big(x, \tfrac{r^++r^-}{2}\big)\right); \tag{4}$$

где $K'\colon \mathbb{R}_+ \to \mathbb{R}_+$ — заданная невозрастающая функция, вообще говоря, отличная от ядра K. Три варианта соответствуют трём способам агрегирования расстояний до полюсов: на уровне ядер (2), на уровне функций расстояния (3), и на уровне самих полюсов (4).

Композиция дипольных классификаторов

Композицией из T диполей с параметрами $\Theta = \big(r_t^-, r_t^+, d_t, R_t, Q_t\big)_{t=1}^T$, будем называть алгоритм классификации вида

$$a(x;\Theta) = \operatorname{sign}\Bigg(\sum_{t=1}^{T} g_t(x,\Theta)\Big(p(x; r_t^+, d_t) - p(x; r_t^-, d_t)\Big)\Bigg), \tag{5}$$

где $g_t(x,\Theta)$ — *шлюзовая функция* (gate [3]), оценивающая компетентность t-го диполя в точке x. Если шлюзовые функции нормированы,

$$g_t(x;\Theta) = \frac{Q_t C(x; r_t^+, r_t^-, R_t)}{\sum_{s=1}^{T} Q_s C(x; r_s^+, r_s^-, R_s)},$$

то композицию диполей можно интерпретировать как *смесь экспертов* [3], а сами шлюзовые функции — как априорные вероятности того, что объект x порождён t-й компонентой смеси. Возвращаясь к физической аналогии, параметр Q_t можно интерпретировать как заряд диполя.

Построение смеси диполей

Композицию диполей удобно настраивать с помощью ЕМ-алгоритма. На каждой ЕМ-итерации сначала вычисляются «скрытые переменные» — апостериорные вероятности h_{ti} того, что объект x_i порождён t-й компонентой смеси (Е-шаг). Благодаря скрытым переменным задача максимизации правдоподобия распадается на T независимых подзадач, по одной для каждого диполя (М-шаг), которые решаются стандартными методами. Возможен вариант, когда полюса диполей выбираются только

из обучающих объектов. Если правдоподобие значительной доли объектов оказывается слишком низким, и среди них есть достаточное количество объектов обоих классов, то в композицию вводится новый диполь.

В докладе подробно рассматриваются различные альтернативные варианты построения дипольной композиции, а также вопросы ускорения сходимости. В экспериментах на модельных данных алгоритм показал хорошее качество классификации линейно разделимой выборки и выборки типа XOR, построив один и два диполя соответственно.

Достоинства дипольных композиций

Благодаря вероятностной модели, наряду с классификацией, могут быть выданы оценки принадлежности объекта каждому из классов, а также выделены *нетипичные объекты*, находящиеся в зоне низкой компетентности всех диполей.

Классификации легко интерпретируются в терминах сходства: «объект x отнесён к классу y потому, что он близок к эталонному объекту x_i класса y». Эталонными объектами являются полюса диполей.

В зависимости от целей применения в роли полюсов могут выступать либо обучающие объекты, либо произвольные точки пространства X.

Благодаря EM-алгоритму с динамическим добавлением компонент строится минимальное необходимое число диполей, в отличие от таких композиционных алгоритмов, как бустинг и бэггинг.

Для хранения алгоритма достаточно запомнить только полюса диполей (и ещё три числа на каждый диполь). В этом дипольная композиция аналогична методу опорных векторов, SVM [5]. Однако, в отличие от SVM, *опорными* являются не пограничные обучающие объекты (которые часто оказываются шумовыми выбросами), а полюса диполей, найденные оптимизационной процедурой. В этом дипольная композиция аналогична методу релевантных векторов, RVM [4]. Однако, в отличие от RVM, здесь не делается никаких априорных предположений о виде вероятностного распределения в пространстве параметров модели.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00877 и № 05-07-90410, а также программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

[1] *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — Р. 320.

[2] *Воронцов К. В.* О композициях дипольных классификаторов // Интеллектуализация обработки информации. — Симф., 2006. — С. 38–40.

[3] *Jordan M. I., Jacobs R. A.* Hierarchical mixtures of experts and the EM algorithm // Neural Computation. — 1994. — no. 6. — Pp. 181–214.

[4] *Tipping M.* The relevance vector machine // Advances in Neural Information Processing Systems, San Mateo, CA. — Morgan Kaufmann, 2000.

[5] *Vapnik V.* Statistical Learning Theory. — Wiley, New York, 1998.

Вероятностный выход для методов многоклассовой классификации на основе самокорректирующихся кодов

Соболев А. А., Вежневец А. П., Вежневец В. П.

neusobol@yandex.ru, {avezhnevets,dmoroz}@graphics.cs.msu.ru

Москва, МГУ им. М. В. Ломоносова,
Лаборатория машинной графики и мультимедиа

Одним из способов сведения задачи классификации с множеством классов к задаче бинарной классификации (с двумя классами) является семейство методов, основанных на самокорректирующихся кодах [1]. В этой статье рассматривается получение вероятностного выхода для данного семейства методов.

Введение

Пусть дана обучающая выборка $D = \{(x_n, y_n)\}_{n=1}^{N} \subset X \times Y$, где X — множество образов, а $Y = \{c_1, \ldots, c_n\}$ — множество меток классов. Пусть $M \in \{\pm 1\}^{C \times T}$ — кодовая матрица, где T — длина кодового слова. Финальный классификатор представляет собой комитет $f(x) = [f_1(x), \ldots, f_T(x)]^{\mathrm{T}}$, где $f_t \colon X \to \mathbb{R}$. В итоге настроенный классификатор работает по принципу минимального расстояния, то есть классом нового объекта x считается тот класс y^*, расстояние до кодового слова $M(y^*)$ которого минимально $y^* = \arg\min\limits_{y} \triangle(M(y), f(x))$. Обычно используется следующая формула расстояния: $\triangle(M(y), f(x)) = \sum_{t=1}^{T} \alpha_t \frac{1 - M(k,t) f_t(x)}{2}$.

Классический подход к получению вероятностей

Использование самокорректирующихся кодов дает существенный прирост в качестве классификации. Однако, для многих прикладных задач, например задач машинного зрения, требуется получение не просто наиболее вероятного класса для прецедента, но и вероятности принадлежности прецедента к тому или иному классу. Ранее для решения данной задачи предлагалось использовать подход, основанный на сведении задачи к решению системы линейных уравнений [2]. Пусть для каждого классификатора из комитета $f_t(x)$ можно вычислить вероятностный выход $P(f_t(x))$. Пусть $p = \big\langle P(f_1(x)), \ldots, P(f_T(x)) \big\rangle$. Например, если столбец t кодовой матрицы M равен $\langle 1, 0, 1 \rangle$, то $P(f_t(x)) = P(c_1|x) + P(c_3|x)$. Обозначим $z = \big\langle P(c_1|x), \ldots, P(c_k|x) \big\rangle$ — вектор искомых вероятностей.

Тогда можно записать матричное уравнение $M^{\mathrm{T}}z = p$. Решая эту систему, мы получим искомые вероятности. Система, вообще говоря, может быть несовместной, поэтому предлагается использовать метод наименьших квадратов. Данный метод имеет ряд недостатков:

— для каждого классифицируемого прецедента приходится решать систему заново, из-за чего метод становится вычислительно сложным;
— метод вычислительно неустойчив — зависит от обусловленности матрицы M.

Предлагаемый метод

Предлагается использовать подход, основанный на нормировке отступа, аналогично [3, 4]. Для этого требуется определить отступ для каждого конкретного класса и выбрать метод шкалирования значения отступа для аппроксимации условной вероятности класса. Отступ для класса c_i определим как

$$\rho(c_i, f(x)) = \min_{c \in Y, c \neq c_i} \triangle(M(c), f(x)) - \triangle(M(c_i), f(x)).$$

Для того, чтобы отступ максимально точно аппроксимировал апостериорную вероятность, применим алгоритм шкалирования [3, 4], в котором предлагается преобразовать отступ сигмоидальной функцией с параметрами, минимизирующими невязку предсказанной вероятности и реальной. Заметим, что при вычислении ответа на образ x вычисление отступов и вероятностей почти не требует дополнительных вычислений — расстояния до кодовых слов будут рассчитаны в любом случае во время классификации:

$$P(c_i|x) \approx \frac{1}{1 + \exp(A\rho(c_i, f(x)) + B)}.$$

Оценка параметров сигмиоды A и B производится на отдельной проверочной выборке, не являющейся ни частью обучающей, ни частью контрольной. В качестве альтернативы можно использовать скользящий контроль с глубиной 3, как предложено в [3].

Эксперименты

Мы сравнили работу своего метода и классического на нескольких выборках из репозитория задач UCI [5]. Мы использовали скользящий контроль глубины 3 для настройки параметров сигмоиды (для каждого класса отдельно). В качестве бинарных классификаторов мы использовали комитет деревьев глубины один (stumps), построенный методом AdaBoost.

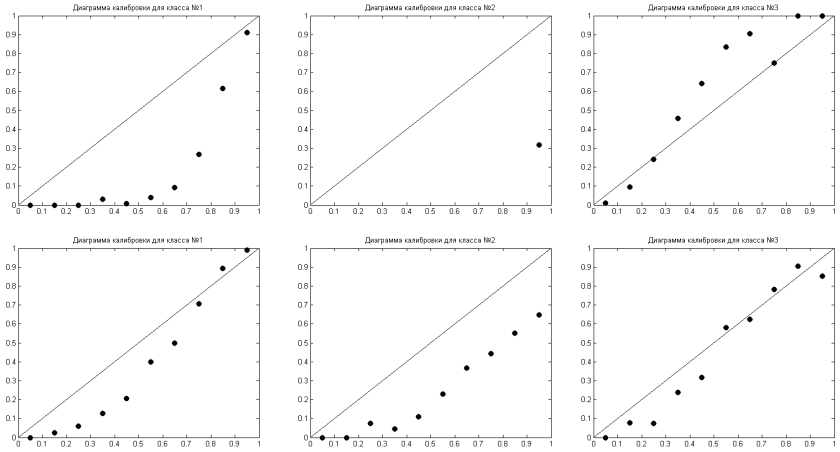


Рис. 1. Результаты экспериментов. Три верхних графика получены методом [2], нижние с помощью предложенного метода.

Ниже приводятся диаграммы калибровки [4] для различных классов из набора abalone для классического и предложенного метода (из-за ограниченного места мы, к сожалению, не можем привести больше графиков). Диаграммы строятся следующим образом. Весь диапазон предсказанных вероятностей делится на ячейки. Для каждой ячейки считается реальная доля прецедентов исследуемого класса. Чем ближе точки лежат к диагональной прямой, тем лучше откалиброван метод. Как видно из графиков, предложенный метод дает более адекватные результаты. Еще раз отметим, что предложенный метод вычислительно намного проще.

Литература

[1] *Dietterich T., Bakiri G.* Solving Multiclass Learning Problems via Error-Correcting Output Codes. // Journal of Artificial Intelligence Research. — 1995. — Pp. 263–286.

[2] *Kong E., Diettrich T.* Probability estimation via error-correcting output coding. // Int'l. Conf. of Articial Inteligence and soft computing. — 1997.

[3] *Platt J.* Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. // Advances in Large Margin Classifiers. — 1999. — Pp. 61–74.

[4] *Niculescu-Mizil A. and Caruana R.* Predicting good probabilities with supervised learning. // 22nd int'l conf. on Machine learning. — 2005. — Pp. 625–632.

[5] *Asuncion A., Newman D. J.* UCI Machine Learning Repository // University of California, Irvine. — 2007. — www.ics.uci.edu/~mlearn/MLRepository.html

Формирование и кластеризация понятий в задаче распознавания образов в пространстве знаний

Степанова Н. А., Емельянов Г. М.
StepanovaNadya@gmail.com
Великий Новгород, ГОУ ВПО НовГУ им. Ярослава Мудрого

Данная работа представляет метод, основанный на теории решеток, для извлечения лексических знаний из текста и последующего упорядочивания знаний. Извлечение знаний выполняется с целью пополнения лексического ресурса, базовым элементом которого является понятие, объединяющее толкование значения лексем с формами лексем.

Будем использовать расширение теории решеток — теорию Анализа Формальных Понятий (АФП) [1]. АФП является инструментом концептуальной кластеризации, так как Формальные Понятия (ФП) решетки являются классами с заданной в виде содержания понятий интерпретацией. При извлечении из текста лексемы, не содержащейся в лексиконе, требуется отнести данную лексему на основе ее признаков к одному из уже имеющихся в лексиконе ФП (классов) или образовать с помощью лексемы новый класс. Построенный, таким образом, лексикон необходимо обрабатывать с целью извлечения классов более высокого уровня абстракции, состоящих из нескольких ФП. Описанная выше задача кластеризации является классической задачей распознавания образов.

Приведем основные определения АФП. Пусть G и M — множества, называемые соответственно множествами объектов и признаков, а $I \subseteq G \times M$ — бинарное отношение. Если $g \in G$ и $m \in M$, то gIm имеет место, если g обладает признаком m. Тройка $\mathbb{K} = (G, M, I)$ называется формальным контекстом. Для произвольных $A \subseteq G$ и $B \subseteq M$ вводится пара отображений: $A' = \{m \in M | \forall g \in A : gIm\}$, $B' = \{g \in G | \forall m \in B : gIm\}$. Пара множеств (A, B), таких что $A' = B$ и $B' = A$, называется ФП с объемом A и содержанием B. ФП (A_1, B_1) называют подпонятием понятия (A_2, B_2), если $A_1 \subseteq A_2$, при этом (A_2, B_2) называют суперпонятием понятия (A_1, B_1), обозначается $(A_1, B_1) \leqslant (A_2, B_2)$. Множество всех ФП контекста $\mathbb{K}$ вместе с отношением порядка называют решеткой ФП $\underline{\mathfrak{B}}(G, M, I)$. Подмножество ФП, в котором каждые два элемента являются сравнимыми, называют цепочкой, а если каждые два элемента являются несравнимыми, называют антицепочкой.

Предлагается использовать Генитивные Конструкции (ГК) [2] в качестве базовой структуры обработки текста. При синтаксическом разборе

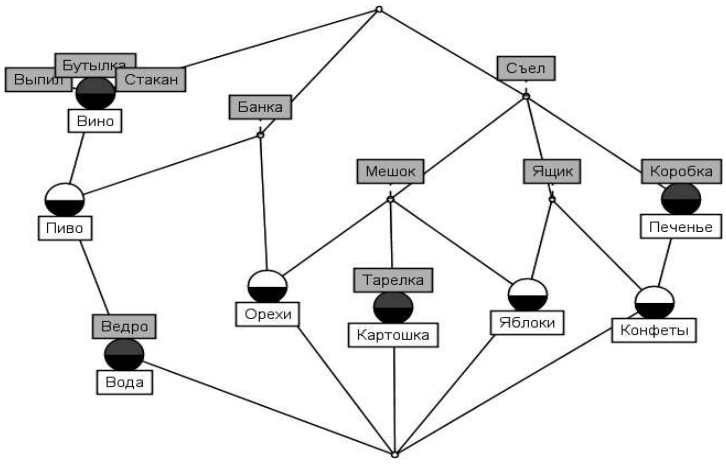


Рис. 1. Формальная решетка генитивных конструкций.

текста выделяются отдельные ГК с указанием Опорного Слова (ОС) и существительного Генитивной Группы (ГГ). Сорта — элементы «наивной картины мира», классы, к которым язык относит более конкретные реалии. Если две ГК относятся к одному сорту, то их ОС и/или ГГ также попарно относятся к одному сорту [2], а значит в их толкованиях должны содержаться общие для них свойства, определяемые их сортом. К одному сорту с некоторой вероятностью будем относить ГК при совпадении их форм ОС или ГГ.

Пусть V_s — множество форм ОС и $v_s \in V_s$, V_{gg} — множество ГГ и $v_{gg} \in V_{gg}$. Бинарным отношением $I \subseteq V_{gg} \times V_s$ назовем множество пар (v_{gg}, v_s) правильных генитивных конструкций. В $\mathbb{K} = (V_{gg}, V_s, I)$ ОС рассматриваются как признаки объектов (ГГ), означающие, что объекты имеют общие свойства. По формальному контексту $\mathbb{K}$ построим решетку $\underline{\mathfrak{B}}(V_{gg}, V_s, I)$. Для целей дополнительной спецификации значения ГК будем использовать глаголы, для модели управления которых ГК занимает место одного из актантов. В расширенном контексте в качестве множества признаков будем использовать объединение множеств $V_s \cup V_g$, где V_g — множество форм глаголов. Приведем пример решетки формальных понятий для набора ГК, извлеченных из текста (рис. 1). Все объекты из объема ФП обладают набором общих свойств A', которые описываются признаками из содержания понятия. Набор признаков будем рассматривать как толкования значений соответствующих лексем из объема ФП. Таким образом, через ГК из текста происходит извлечение знаний, представленных формальными понятиями.

Рассмотрим алгоритм сегментации решетки, выделяющий из первоначальной решетки L классы ФП, которые являются более высоким уровнем абстракции, чем отдельные ФП. Любое подмножество ФП будет обязательно иметь уникальное Наименьшее Общее Суперпонятие (НОСП). Областью в решетке называется набор ФП, связанных отношением порядка с одним НОСП. Задачей алгоритма сегментации решетки является конвертация решетки L в набор формальных решеток $\{L'\}$, где каждой решетке $L_i \in \{L'\}$ частично соответствует область первоначальной решетки L и каждое ФП, кроме вершинного и наименьшего ФП, принадлежит только к одной итоговой решетке из $\{L'\}$. Поскольку области в решетке могут пересекаться, то для выполнения условия принадлежности ФП только к одной итоговой решетке необходимо, чтобы классы L_i лишь частично соответствовали областям первоначальной решетки $\{L\}$. Спорным ФП первоначальной решетки L называют такое ФП C, что C принадлежит более чем к одной области решетки L, у которых НОСП являются несравнимыми ФП.

Для максимизации количества элементов в итоговых классах в качестве НОСП областей решетки L возьмем ФП, являющиеся непосредственными подпонятиями вершинного ФП. Критерием выделения решетки L_i из решетки L является условие, что каждое ФП $C \in L_i$ более схоже с другими ФП из решетки L_i, чем с ФП из решеток $L_j \in \{L'\}$, где $i \neq j$. Мера схожести между ФП вычисляется по формуле

$$\mathrm{spc}(C_i, C_j) = -\log\left(1 - \frac{D_c}{\mathrm{path}_C}\right) \times \frac{|B|}{|B_j \backslash B| + |B_j \backslash B| + |B|}, \qquad (1)$$

где ФП $C = (A, B)$ является НОСП для формальных понятий C_i и C_j; D_c — количество ФП в цепочке, в которой максимальным ФП является вершинное ФП, а минимальным — ФП C; path_C — минимальное количество ФП в цепочке, которой принадлежат вершинное, наименьшее ФП и ФП C. Мера схожести учитывает объем общей ($|B|$) и индивидуальной ($|B_i|, |B_j|$) информации ФП C_i и C_j, специфичность общей информации (D_c), а также неравномерность глубины иерархии решетки (path_C).

Алгоритм сегментации для каждого ФП C_i, являющегося непосредственным подпонятием вершинного ФП, включает C_i в решетку L_i. Далее выполняется поиск в решетке L всех подпонятий ФП C_i, и каждое из этих подпонятий C_j включается в решетку L_i, если все непосредственные суперпонятия ФП C_j являются подпонятиями ФП C_i или совпадают с C_i. В противном случае, ФП C_j является спорным ФП и относится алгоритмом к тому классу, к которому принадлежит то из его непосредственных суперпонятий, для которого значение меры схожести по формуле (1) с ФП C_j будет максимальным.

Работа выполнена при поддержке РФФИ, проект № 06-01-00028.

Литература

[1] *Ganter B., Wille R.* Formal Concept Analysis - Mathematical Foundations. — Berlin: Springer-Verlag, 1999. — 284 с.

[2] *Partee B. H.* Formal Semantics, Lectures, RGGU, 2003. — `people.umass.edu/partee/`.

Объективизация экспертных оценок, выставленных в ранговых шкалах

Стрижов В. В., Казакова Т. В.

`strijov@ccas.ru`

Москва, Вычислительный центр РАН

Предложен способ построения интегральных индикаторов качества сложных объектов с использованием экспертных оценок и измеряемых данных. Предполагается, что экспертные оценки выставлены в ранговых шкалах.

Ранее были предложены два подхода к построению индексов с использованием экспертных оценок. Первый подход состоит в нахождении параметров модели свертки данных, которые доставляют минимум невязки между вычисленным индексом и его экспертной оценкой [1]. Второй подход имеет целью согласовать экспертные оценки индексов, весов показателей, и заключается в поиске компромиссного решения [2, 3].

Предложен алгоритм вычисления индексов на основе экспертных оценок, выполненных в ранговых шкалах. Результатом работы алгоритма являются индексы, уточняющие экспертные оценки, и не противоречащие измеряемым данным. Индексы являются устойчивыми, то есть не зависящими от наличия объектов с неординарными описаниями. Измеряемые данные и экспертные оценки обобщаются в непротиворечивую систему.

Задано множество, состоящее из m сравнимых объектов, которые описаны набором из n показателей.

Задана матрица описаний «объект-показатель» $A \in \mathbb{R}^{m\times n}$. Элемент матрицы a_{ij} — значение j-го показателя i-го объекта. Заданы начальные экспертные оценки. Каждому объекту поставлена в соответствие экспертная оценка качества объекта, а каждому показателю — экспертная оценка его важности. То есть заданы два упорядоченных набора $\mathbf{q}_0 \in \mathbb{R}^m$, $\mathbf{w}_0 \in \mathbb{R}^n$. Оценки $\mathbf{q}_0$, $\mathbf{w}_0$, допускают, по условию, произвольные монотонные преобразования. Без ограничения общности будем считать, что на

наборах экспертных оценок введено отношение порядка такое, что

$$\mathbf{q}_0 = \{q_i \colon q_1 \geqslant \cdots \geqslant q_m \geqslant 0\}, \quad \mathbf{w}_0 = \{w_j \colon w_1 \geqslant \cdots \geqslant w_n \geqslant 0\}. \qquad (1)$$

Назначена линейная модель вычисления индексов. *Интегральный индикатор* объекта — свертка вида $q_i = \sum\limits_{j=1}^{n} w_j g_j(a_{ij})$, где g_j — функция приведения показателей в единую шкалу. Далее предполагается $g_j = \mathrm{id}_j$.

Требуется построить индекс, основанный на измеряемых данных, и не противоречащий мнениям экспертов.

Согласованными значениями интегрального индикатора и весов показателей называются такие значения $\mathbf{q}$ и $\mathbf{w}$, при которых выполняется условие

$$\begin{cases} \mathbf{q} = A\mathbf{w}; \\ \mathbf{w} = A^{+}\mathbf{q}. \end{cases} \qquad (2)$$

В работе [3] описан метод согласования экспертных оценок в линейных шкалах. Он заключается в следующем. Каждый объект из множества заданных объектов можно оценить двумя путями: непосредственно через экспертную оценку $\mathbf{q}_0$ и через взвешенную сумму значений показателей объекта, $\mathbf{q}_1 = A\mathbf{w}_0$, где веса определяются экспертными оценками показателей $\mathbf{w}_0$. По исходным экспертным оценкам значения вектора интегрального индикатора $\mathbf{q}_0$ вычисляется вектор весов показателей $\mathbf{w}_1 = A^{+}\mathbf{q}_0$, где A^{+} — оператор, псевдообратный оператору A. В общем случае вектор экспертной оценки $\mathbf{q}_0$ объектов и вектор взвешенной суммы значений показателей объектов $\mathbf{q}_1$ различны. Также различны векторы $\mathbf{w}_0$ и $\mathbf{w}_1$. Согласованное решение отыскивается на отрезках, соединяющих соответствующие пары векторов.

Псевдообратный оператор A^{+}, такой что $A^{+}A = I_n$, $AA^{+} = I_m$, отыскивается с помощью сингулярного разложения $A = U\Lambda V^T$ и равен $A^{+} = V\Lambda^{-1}U^T$. При большой обусловленности A псевдообратный оператор регуляризуется. Для этого выполняется подстановка $\Lambda^{-1} \mapsto \Lambda^{-1} + r^{-1}I_n$, где r — коэффициент регуляризации.

Рассмотрим следующий алгоритм согласования экспертных оценок, выставленных в ранговых шкалах. Отношение порядка (1) задает конусы $\mathcal{Q}_0 \in \mathbb{R}^m_+$ и $\mathcal{W}_0 \in \mathbb{R}^n_+$ в пространстве оценок объектов и в пространстве показателей соответственно. Линейный оператор A отображает конус $\mathcal{W}_0$ в конус $A\mathcal{W}_0$. Оператор A^{+} отображает конус $\mathcal{Q}_0$ в конус $A^{+}\mathcal{Q}_0$. Обозначим $\mathcal{W}_\rho = \mathcal{W}_0 \cup A^{+}\mathcal{Q}_0$ и $\mathcal{Q}_\rho = \mathcal{Q}_0 \cup A\mathcal{W}_0$.

Справедливы утверждения. Если конус $\mathcal{Q}_\rho$, не пуст, то не пуст также и конус $\mathcal{W}_\rho$, в противном случае оба конуса пусты. Для каждого вектора $\mathbf{w}_\rho \in \mathcal{W}_\rho$ найдется согласованный с ним вектор $\mathbf{q}_\rho \in \mathcal{Q}_\rho$, такой что выполняется равенство (2).

Таким образом, в случае непустого пересечения конусов, экспертные оценки являются согласованными в ранговых шкалах и удовлетворяют условию (2). Согласованными экспертными оценками является любая согласованная пара $\mathbf{q}_\rho, \mathbf{w}_\rho$, принадлежащая пересечениям конусов в соответствующих пространствах.

Для отыскания пресечения конусов $\mathcal{Q}_\rho$ опишем соответствующие множества системами линейных неравенств. Представим конус $\mathcal{Q}_0$, элементы которого удовлетворяют условию (1), в виде двухдиагональной матрицы $\mathcal{Q}_0$, в которой элементы на главной диагонали равны 1, а элементы на диагонали $(1,2),\ldots,(n-1,n)$ равны -1. Представим произведение $A\mathcal{W}_0$ также в виде матрицы коэффициентов в пространстве $\mathbb{R}^{m\times m}$.

Множество векторов $\mathbf{q}_\rho \in \mathcal{Q}_\rho$ является решением объединенной системы линейных неравенств

$$\begin{cases} \mathcal{Q}_0\mathbf{q}_\rho \geqslant 0; \\ (A\mathcal{W}_0)\mathbf{q}_\rho \geqslant 0. \end{cases}$$

Полученное пересечение $\mathcal{Q}_\rho$ также является конусом, возможно, тривиальным, каждый элемент которого является интегральным индикатором, удовлетворяющим условию согласованности (2).

Следует отметить, что независимо от того, имеет система линейных неравенств решение или нет, представляется возможным найти согласованный линейный индикатор, минимально отличающийся от экспертных оценок. Для этого, если конус $\mathcal{Q}_\rho$ пуст, на единичной сфере S отыскивается точка $\mathbf{q}_\rho \in s = (A\mathbf{w}, \mathbf{q})$, на дуге $s = \arg\min \sigma(s)$, где $A\mathbf{w} \in A\mathcal{W} \cap S$, $\mathbf{q} \in \mathcal{Q} \cap S$ и $\sigma(s)$ — длина дуги. В противном случае, отыскивается точка $\mathbf{q}_\rho$, равноудаленная от вершин конуса $\mathcal{Q}_\rho \cap S$. Вектор весов показателей отыскивается посредством псевдообратного оператора A^+.

Полученные процедуры нахождения интегрального индикатора использованы для объективизации экспертных оценок в следующих прикладных задачах: для построения интегральных индикаторов развития человеческого потенциала регионов РФ и интегральных индикаторов качества управления Особо охраняемыми природными территориями РФ.

Работа выполнена при поддержке РФФИ, проект № 07-07-00181.

Литература

[1] *Айвазян С А., Мхитарян В. С.* Прикладная статистика и основы эконометрики. — М.: ЮНИТИ, 1998. — 363 с.

[2] *Strijov V., Shakin V.* Index construction: the expert-statistical method // Environmental research, engineering and management. — 2003. — V. 26, № 4. — Pp. 51–55.

[3] *Стрижов В. В.* Уточнение экспертных оценок с помощью измеряемых данных // Заводская лаборатория. — 2006. — Т. 92, № 6. — С. 59–64.

Построение инвариантов на множестве временных рядов путем динамической свертки свободной переменной

Стрижов В. В., Пташко Г. О.
strijov@ccas.ru
Москва, Вычислительный центр РАН

Дано множество временных рядов, на котором заданы классы эквивалентности. Требуется построить модель — параметрическое семейство функций, которая задает инвариант, отображающий временные ряды в соответствующие классы. Модель выбирается из индуктивно порожденного множества. Предложенный алгоритм основан на методе поиска оптимальной регрессионной модели как произвольной суперпозиции порождающих функций [1].

Ранее были предложены методы классификации временных рядов с использованием динамической свертки времени [2]. Классификация выполнялась с учетом порогового значения пути наименьшей стоимости. Также классификация может быть выполнена с использованием пространства параметров функций, аппроксимирующих временные ряды [3] в качестве входной информации.

Особенностью предложенного алгоритма является то, что аппроксимация выполняется не на множестве временных рядов, а на множестве их путей минимальной стоимости, что позволяет сократить множество моделей-претендентов и уменьшить их сложность.

Постановка задачи

Дано $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ — множество временных рядов, $\mathbf{x}_i = \{x_{it}\}_{t=1}^{T}$. На упорядоченных парах индексов (i, j) временных рядов $(\mathbf{x}_i, \mathbf{x}_j)$ задано отношение эквивалентности $[\mathbf{x}]$, которое интерпретируется как принадлежность временных рядов к некоторому классу.

Задано множество монотонных параметрических гладких функций $\big\{g(\mathbf{b}, \cdot, \dots, \cdot) \mid g\colon \mathbb{R} \times \dots \times \mathbb{R} \to \mathbb{R}\big\}$, которое порождает множество суперпозиций $\{f_k(\mathbf{w}_k)\colon t \to t\}$ путем обобщенного индуктивного определения. Вектор параметров $\mathbf{w}$ есть присоединенные векторы $\langle \mathbf{b}_1, \dots, \mathbf{b}_{r(k)} \rangle$ параметров функций, входящих в суперпозицию; размерность вектора $\mathbf{w}$ зависит от структуры суперпозиции: $\mathbf{w}_k \in \mathbb{R}^{W(k)}$.

Требуется выбрать такую модель $f_k(\mathbf{w})$ и найти такое множество допустимых значений ее параметров $[w_k] \subset \mathbb{R}^{W(k)}$, которые задают требуемый класс эквивалентности $[\mathbf{x}]$ при условии, что множество $[w_k]$ до-

ставляет максимум целевой функции S на всех парах временных рядов из $[\mathbf{x}]$. S есть функционал качества аппроксимации f_k пути наименьшей стоимости, определенного парой временных рядов.

Аппроксимация пути наименьшей стоимости

Построим матрицу расстояний $\Omega(\mathbf{x}_i, \mathbf{x}_j) = \{\omega_{pq} = \|x_{ip} - x_{jq}\|_2\}_{p,q=1}^{T}$ между всеми парами элементов (x_{ip}, x_{jq}) временных рядов $\mathbf{x}_i, \mathbf{x}_j$.

Для отыскания пути наименьшей стоимости рассмотрим в матрице Ω все пути $\mathbf{s} = \{s_1, \ldots, s_C\}$ такие, что $s_1 = \omega_{11}$, $s_C = \omega_{T,T}$ и для произвольного $s_c = \omega_{pq}$, где $c = 1, \ldots, C-1$ значение $s_{c+1} = \omega_{p+u,q+v}$, где $u + v \in \{1, 2\}$. Вычислим стоимость пути $\mathbf{s} = C^{-1}\big(\sum_{c=1}^{C} s_c^2\big)^{1/2}$. Обозначим $\bar{\mathbf{s}}_{ij}$ — путь наименьшей стоимости для пары временных рядов $(\mathbf{x}_i, \mathbf{x}_j)$. Подробнее об алгоритме динамической свертки временных рядов см. [4].

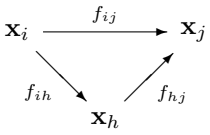
Найдем такие параметры $\mathbf{w}_k(i,j)$, при которых модель наилучшим образом приближает путь $\bar{\mathbf{s}}_{ij}$. Оптимальные значения параметров заданы как $\bar{\mathbf{w}}_k = \arg\min S\big(f_k(\mathbf{w}_k(i,j)), \bar{\mathbf{s}}_{ij}\big)$.

Выберем такую модель, которая не нарушает отношения эквивалентности $[\mathbf{x}]$, и для которой максимальное значение S по всем парам (i,j), таким, что $(\mathbf{x}_i, \mathbf{x}_j) \in [\mathbf{x}]$ минимально: $\max\limits_{i,j} S\big(f_k(\bar{\mathbf{w}}_k(i,j)), \bar{\mathbf{s}}_{ij}\big) \to \min$.

Классы эквивалентности в пространстве параметров

Обозначим $\bar{f}_{ij}$ фиксированную по k функцию $f_k(\bar{\mathbf{w}}_k(i,j))$. Множество последовательных отображений $\bar{f}_{12}, \bar{f}_{23}, \ldots, \bar{f}_{(\ell-1)\ell}$ задано множеством фиксированных параметров $\bar{\mathbf{w}}_{12}, \bar{\mathbf{w}}_{23}, \ldots, \bar{\mathbf{w}}_{(\ell-1)\ell}$, $\ell = N(N-1)/2$, полученных в результате идентификации.

Элементы множества $\{\bar{f}_{ij}\}$, сохраняющие класс эквивалентности $[\mathbf{x}]$, задают группу преобразований G. Для произвольной тройки i, j, h индексов временных рядов из $[\mathbf{x}]$ справедлива коммутативная диаграмма



и $\quad G = \begin{pmatrix} \mathrm{id} & f_{12} & \cdots & f_{1N} \\ f_{12}^{-1} & \mathrm{id} & \cdots & f_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ f_{1N}^{-1} & f_{2N}^{-1} & \cdots & \mathrm{id} \end{pmatrix}$.

Группа G задает классы эквивалентности в пространстве параметров, $[\mathbf{w}] \ni \{\bar{\mathbf{w}}_k(i,j)\}$. Рассмотрим такие модели f_k, которые позволяют доопределить $[\mathbf{w}]$ до выпуклой оболочки $[w_k]$ элементов этого класса без потери отношения эквивалентности.

Для каждого нового временного ряда $\mathbf{x}_{N+1} \notin X$ заданной ранее размерности находим $\{\bar{\mathbf{w}}_k(i,N)\}$ для $i \in \{1,\ldots,N\}$. Ряд $\mathbf{x}_{N+1}$ принадлежит $[\mathbf{x}]$, если вектор параметров $\bar{\mathbf{w}}_k(i,N+1)$ принадлежит $[w_k]$. Таким образом, по построенному классу эквивалентности $[w_k]$ можно определить принадлежность нового временного ряда к $[\mathbf{x}]$.

Заключение

Предложенный метод был использован для классификации временных рядов давления в камере внутреннего сгорания дизельного двигателя. Непосредственное вычисление значения пути оптимальной стоимости не позволило решить задачу классификации исследуемых временных рядов, так как стоимость двух несовпадающих путей временных рядов из разных классов часто оказывалась одинаковой. Создание моделей, аппроксимирующих эквивалентные временные ряды, также не позволило решить данную задачу, так как классификацию при этом приходилось выполнять в пространстве параметров, которое имело большую размерность. Предложенный метод позволяет разделить данные временные ряды на кластеры, так как классификация выполняется в пространстве параметров небольшой размерности.

Работа выполнена при поддержке РФФИ, проект № 07-07-00181.

Литература

[1] *Стрижов В. В., Пташко Г. О.* Алгоритмы поиска суперпозиций при выборе оптимальных регрессионных моделей. — М: ВЦ РАН, 2006. — 42 с.

[2] *Pavlidis I., Singh R., Papanikolopoulos N.* Recognition of On-Line Handwritten Patterns Through Shape Metamorphosis. // Proceedings of the 13th International Conference on Pattern Recognition. — 1996. — Vol. 3. — Pp. 18–22.

[3] *Стрижов В. В.* Поиск параметрической регрессионной модели в индуктивно заданном множестве // Вычислительные технологии. — 2007. — Т. 12, № 1. — С. 93–102.

[4] *Keogh E. J., Pazzani M. J.* Derivative Dynamic Time Warping // First SIAM International Conference on Data Mining (SDM'2001), Chicago, USA. 2001. — `http://www.ics.uci.edu/~pazzani/Publications/sdm01.pdf`.

Потенциальные функции на множестве векторных последовательностей разной длины

Сулимова В. В., Моттль В. В., Мучник И. Б.

`sulimova@tula.net, vmottl@yandex.ru, muchnik@dimacs.rutgers.edu`

Тула, ТулГУ; Москва, ВЦ РАН; США, Rutgers University

Последовательности разной длины $\omega = (\alpha_k,\, k = 1,\ldots,N_\omega)$ являются типовыми объектами задач анализа данных. Примитивы $\alpha_k \in A$, состав-

ляющие последовательности, могут быть действительными числами, векторами или символами некоторого конечного алфавита. В первых двух случаях последовательности являются скалярными или векторными сигналами, а в последнем случае принято говорить о символьных последовательностях.

Широко известны такие задачи анализа последовательностей разной длины, как задача идентификации личности по динамике подписи, распознавание речевых команд и слитного текста, прогнозирование биологических свойств белков на основе анализа составляющих их последовательностей аминокислотных остатков.

Удобным инструментом решения таких задач являются методы, основанные на понятии потенциальной функции [1, 2] — действительной функции двух аргументов, матрица значений которой для любой конечной совокупности объектов является неотрицательно определенной. Потенциальная функция $K(\omega', \omega'')$, определенная на множестве произвольных объектов $\omega \in \Omega$, погружает исходное множество в некоторое линейное пространство $\tilde{\Omega} \supseteq \Omega$, в котором она является скалярным произведением. Такое погружение позволяет применять для объектов произвольной природы практически любые классические методы анализа данных, разработанные для линейных пространств, минуя промежуточный этап выбора вектора числовых признаков $\mathbf{x}(\omega) \in R^n$, определяющего скалярное произведение $K(\omega', \omega'') = \mathbf{x}^{\mathrm{T}}(\omega')\,\mathbf{x}(\omega'')$. Для последовательностей разной длины поиск числовых признаков особенно проблематичен.

Можно предложить разные способы введения потенциальной функции на множестве последовательностей разной длины. Мы сначала изложим одну достаточно общую математическую структуру потенциальной функции, адекватную потребностям многих практических задач, а затем рассмотрим частный случай этой структуры, приводящий к простому алгоритму вычисления потенциальной функции для любых двух заданных последовательностей.

Потенциальная функция на множестве примитивов

Будем полагать, что множество примитивов A является линейным пространством со скалярным произведением (потенциальной функцией) $\mu(\alpha', \alpha'')$. Тогда величина $\sqrt{\mu(\alpha, \alpha)}$ играет роль нормы в этом линейном пространстве.

Интерпретация множества примитивов как линейного пространства представляется вполне естественной для сигналов, изначально являющихся последовательностями действительных чисел либо векторов. Можно показать, что такая интерпретация остается естественной и для аминокислотных последовательностей белков, представляющих собой символьные последовательности над алфавитом двадцати существую-

щих в природе аминокислот, если используется общепринятый в молекулярной биологии способ измерения сходства всяких двух аминокислот как вероятности их происхождения в процессе эволюции от одной и той же неизвестной аминокислоты [3].

Множество выравниваний двух последовательностей

Если бы все последовательности имели одинаковую длину $\Omega = \{\omega = (\alpha_k,\, k = 1, \ldots, N = \text{const})\}$, то, например, произведение $K(\omega', \omega'') = \prod_{k=1}^{N} \mu(\alpha'_k, \alpha''_k)$ обладало бы всеми свойствами потенциальной функции на множестве Ω, поскольку известно, что произведение любого числа потенциальных функций также является потенциальной функцией [4]. Однако мы имеем дело с множеством последовательностей разной длины $\Omega = \{\omega = (\alpha_k,\, k = 1, \ldots, N_\omega)\}$, и применение такого способа возможно лишь после выравнивания длин сравниваемых последовательностей.

Под выравниванием w двух последовательностей $\omega' = (\alpha'_k,\, k = 1, \ldots, N')$ и $\omega'' = (\alpha''_k,\, k = 1, \ldots, N'')$, $\alpha'_k, \alpha''_k \in A$, понимают приведение их к одинаковой длине за счет добавления «пустых» выравнивающих элементов в некоторые позиции каждой из последовательностей с последующей перенумерацией элементов: $\bar{\omega}'_w = (\bar{\alpha}'_{w,j},\, j = 1, \ldots, |w|)$ и $\bar{\omega}''_w = (\bar{\alpha}''_{w,j},\, j = 1, \ldots, |w|)$, где $|w| \geqslant \max\{N', N''\}$ — общая длина выровненных последовательностей. В качестве выравнивающего элемента мы произвольно выберем некоторый элемент исходного линейного пространства примитивов $\alpha^0 \in A$, имеющий единичную норму $\mu(\alpha^0, \alpha^0) = 1$.

Множество всех выравниваний упорядоченной пары последовательностей $\langle \omega', \omega'' \rangle$ длин N' и N'' будем обозначать $\mathcal{W}_{N'N''}$. Любое выравнивание $w \in \mathcal{W}_{N'N''}$ может быть представлено в виде пути на графе с горизонтальными, диагональными и вертикальными ребрами, ориентированными слева направо и сверху вниз, как показано на Рис. 1. Горизонтальное направление на таком графе будем связывать с первой последовательностью $\omega' = (\alpha'_k,\, k = 1, \ldots, N')$, а вертикальное направление — со второй последовательностью $\omega'' = (\alpha''_k,\, k = 1, \ldots, N'')$.

Будем рассматривать всякое выравнивание $w \in \mathcal{W}_{N'N''}$ как последовательность значений переменной из трехэлементного алфавита: $w = (h_k,\, k = 1, \ldots, |w|)$, $h_k \in \{h, h', h''\}$. Значение $h_k = h'$ означает продвижение на один шаг в горизонтальном направлении, т. е. вставку одного «пустого» выравнивающего элемента в первую последовательность, значение $h_k = h$ интерпретируется как диагональное продвижение, соответствующее отсутствию вставки, а $h_k = h''$ обозначает шаг в вертикальном направлении, вставляющий «пустой» элемент во вторую последовательность. Симметричный аналог всякого выравнивания w, получаемый

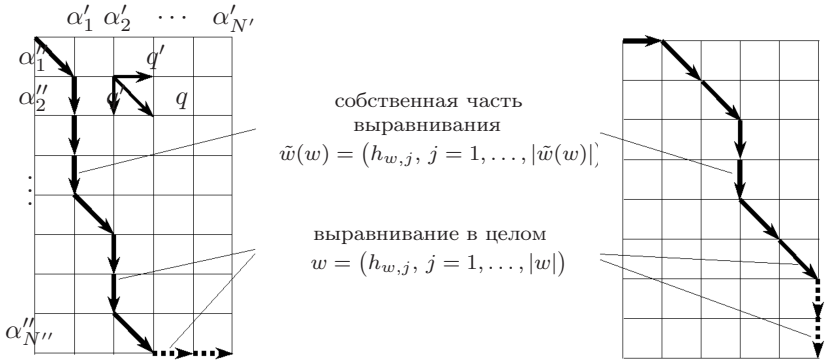


Рис. 1. Два разных выравнивания пары последовательностей.

заменой каждого шага $h_k = h'$ на $h_k = h''$ и наоборот, будем обозначать символом w^{T}, так что $\mathcal{W}_{N''N'} = \{w^{\mathrm{T}}\colon\ w \in \mathcal{W}_{N'N''}\}$.

Система весов на множестве выравниваний и структура потенциальной функции

С выравниванием w двух последовательностей ω' и ω'' свяжем действительную величину

$$K(\omega', \omega'' \,|\, w) = K(\omega'', \omega' \,|\, w^{\mathrm{T}}) = \prod_{j=1}^{|w|} \mu(\bar{\alpha}'_{w,j}, \bar{\alpha}''_{w,j}), \tag{1}$$

понимаемую как мера условного сходства двух последовательностей, зависящая от выбора выравнивания w.

Далее, выберем систему неотрицательных весов парных выравниваний $p(w) \geqslant 1$, общую для всех значений пар длин последовательностей N' и N'', и выражающую априорные предпочтения на множестве разных выравниваний одной и той же пары.

Традиционный способ измерения сходства двух последовательностей основан на поиске выравнивания, максимизирующего их условное сходство с учетом веса $K(\omega', \omega'') = \max\limits_{w \in \mathcal{W}_{N'N''}} p(w) K(\omega', \omega'' \,|\, w)$ [5], однако такая мера сходства не будет обладать свойствами потенциальной функции.

В данной работе вместо операции максимизации мы используем линейную комбинацию значений условного сходства двух последовательностей по всем выравниваниям с учетом их весов:

$$K(\omega', \omega'') = \sum_{w \in \mathcal{W}_{N'N''}} p(w) K(\omega', \omega'' \,|\, w). \tag{2}$$

Пусть w — некоторое выравнивание последовательностей ω' и ω'', имеющих длины N' и N''. Начальную часть выравнивания w до первого

касания правой или нижней границы области $\mathcal{W}_{N'N''}$ (Рис. 1), будем называть его собственной частью и обозначать символом $\bar{w}(w)$.

Дополним последовательности ω' и ω'' длин N' и N'' выравнивающими элементами $\alpha^0 \in A$ справа до некоторой длины N, и будем называть полученные последовательности $\bar{\omega}' = (\alpha'_{k'},\, k'{=}1,\ldots,N,\, \alpha'_{k'}{=}\alpha^0,\, k' > N')$ и $\bar{\omega}'' = (\alpha''_{k''},\, k''{=}1,\ldots,N,\, \alpha''_{k''}{=}\alpha^0,\, k'' > N'')$ расширенными. Все выравнивания расширенных последовательностей образуют множество $\mathcal{W}_{NN}$. Два выравнивания $w \in \mathcal{W}_{N'N''}$ и $\bar{w} \in \mathcal{W}_{NN}$ будем называть эквивалентными и обозначать как $w \sim \bar{w}$, если собственная часть выравнивания w является начальной частью выравнивания $\bar{w}$.

Система весов $p(w)$ называется согласованной, если, во-первых, веса симметричных выравниваний равны $p(w) = p(w^{\mathrm{T}})$, и, во-вторых, для любых N', N'' и N, таких, что $N \geqslant N'$ и $N \geqslant N''$, выполняется условие $p(w) = \sum_{\bar{w}\in\mathcal{W}_{NN},\, w\sim\bar{w}} p(\bar{w})$, т. е. вес выравнивания w исходных последовательностей равен сумме весов эквивалентных ему расширенных последовательностей.

Теорема 1. *Для того, чтобы линейная комбинация $K(\omega',\omega'')$ (2) условных мер сходства $K(\omega',\omega''\,|\,w)$ (1) обладала свойствами потенциальной функции на множестве последовательностей над линейным пространством примитивов $\Omega = \{\omega = (\alpha_k, k = 1,\ldots,N_\omega),\ \alpha_k{\in}A\}$, достаточно, чтобы выравнивающий элемент удовлетворял условию $\mu(\alpha^0,\alpha^0) = 1$ и система весов $p(w)$ была согласованной.*

Тот факт, что некоторая двухместная функция $K(\omega',\omega'')$ (2) формально обладает свойствами потенциальной функции на множестве последовательностей разной длины, еще не гарантирует ее практическую полезность. Важно удачно выбрать исходную потенциальную функцию $\mu(\alpha',\alpha'')$ на множестве примитивов $\alpha \in A$, выравнивающий элемент $\alpha^0 \in A$, а также систему весов выравниваний $p(w)$.

Радиальная потенциальная функция на множестве примитивов и мультипликативные веса выравниваний

Пусть в линейном пространстве примитивов с нулевым элементом $\varnothing \in A$ определена евклидова метрика, например, с помощью некоторой исходной потенциальной функцией $\rho(\alpha',\alpha'') = \big[\varkappa(\alpha',\varnothing) + \varkappa(\alpha'',\varnothing) - 2\varkappa(\alpha',\alpha'')\big]$. Известно [1], что в этом случае двухместная функция

$$\mu(\alpha',\alpha'') = \exp\big[-\beta\rho^2(\alpha',\alpha'')\big] \tag{3}$$

обладает свойствами потенциальной функции при любом значении параметра $\beta{>}0$, преобразуя линейное пространство A в некоторое другое линейное пространство со скалярным произведением $\mu(\alpha',\alpha'')$.

Потенциальную функцию (3), по своему смыслу являющуюся мерой сходства примитивов относительно исходной метрики $\rho(\alpha', \alpha'')$, принято называть радиальной.

Выбор нулевого элемента исходного линейного пространства в качестве выравнивающего элемента $\alpha^0 = \emptyset \in A$ удовлетворяет условию $\mu(\alpha^0, \alpha^0) = 1$ в Теореме 1.

С каждым из трех значений переменной h, h' и h'' свяжем неотрицательные числа $q(h) = q$ и $q(h') = q(h'') = q'$, $q + 2q' = 1$. Значение $q > 1/3$ задает предпочтительность отсутствия вставок и удалений элементов на каждом элементарном шаге сравнения последовательностей, Рис. 1.

Пусть $w = (h_{w,j}, j = 1, \ldots, |w|)$ — произвольное выравнивание, $\tilde{w}(w)$ — его собственная часть. Вес выравнивания определим как произведение

$$p(w) = \prod_{j=1}^{|\tilde{w}(w)|} q(h_{w,j}). \tag{4}$$

Теорема 2. *Система весов выравниваний (4) является согласованной.*

Таким образом, радиальная потенциальная функция на множестве примитивов и мультипликативная система весов выравниваний удовлетворяют всем требованиям Теоремы 1 и определяют потенциальную функцию на множестве последовательностей разной длины (2), явным образом выражающую степень их попарного сходства. Алгоритм вычисления такой потенциальной функции имеет сложность, пропорциональную произведению длин сравниваемых последовательностей.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00679, № 06-01-08042 и № 06-01-00412, а также INTAS, проекты № 04-77-7347 и № 06-1000014-6563;

Литература

[1] *Айзерман М. А., Браверманн Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — 384 с.

[2] *Vapnik V.* Statistical Learning Theory. — New York: John-Wiley & Sons, Inc., 1998. — 732 p.

[3] *Dayhoff M. O., Schwartz R. M., Orcutt B. C.* A model for evolutionary change in proteins. — Atlas for Protein Sequence and Structure (M. O. Dayhoff, ed.). — 1978. — Vol. 5. — Pp. 345–352.

[4] *Haussler D.* Convolution kernels on discrete structures. — Technical Report UCSC-CLR-99-10, University of California at Santa Cruz, 1999.

[5] *Dubin R., Eddy S., Krogh A., Mitchison G.* Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids. — Cambridge University Press, 1998. — 356 p.

Комбинирование классификаторов и потенциальных функций в многомодальном распознавании образов

Татарчук А. И., Елисеев А. П., Моттль В. В.
aitech@yandex.ru, andreyel@gmail.com, mottl@yandex.ru
Москва, МФТИ, ВЦ РАН

Под многомодальным распознаванием образов понимается комплекс задач и методов их решения, связанных с необходимостью принимать суждения о классах объектов, характеризующихся одновременно несколькими видами представлений (модальностей), часто имеющих разную физическую природу. В данной работе проведено теоретическое и экспериментальное исследование принципов многомодального обучения распознавания образов в рамках метода потенциальных функций при разных уровнях комбинирования модальностей.

Уровни комбинирования модальностей

Способ математического выражения информации об объектах в задачах анализа данных принято называть *модальностью* представления объектов. В терминах выбранной модальности каждый объект реального мира $\omega \in \Omega$ характеризуется некоторым значением в пространстве соответствующего обобщенного признака $x(\omega) \in \mathbb{X}$, например, в виде сигнала, изображения, а, в сравнительно простых ситуациях, в виде действительного числа или вектора.

В задачах обучения распознаванию образов по прецедентам такое преобразование приводит к обучающей совокупности, составленной из значений заданного обобщенного признака $x_j = x(\omega_j) \in \mathbb{X}$ и индексов классов $y_j = y(\omega_j) \in Y = \{1, \ldots, m\}$ для доступного подмножества объектов $\Omega^* = \{\omega_j, j = 1, \ldots, N\} \subset \Omega$. Анализируя предъявленный массив данных, требуется продолжить функцию $y(x(\omega_j))$, заданную в пределах обучающей совокупности $\omega_j \in \Omega^*$, на все множество объектов $\hat{y}(x(\omega))\colon \mathbb{X} \to Y$.

Обоснованность предположения, что в результате анализа обучающей совокупности $\{(x_j, y_j), j = 1, \ldots, N\}$ удастся построить решающее правило $\hat{y}(x(\omega))\colon \mathbb{X} \to Y$, хорошо аппроксимирующее фактическое разбиение генеральной совокупности объектов $\omega \in \Omega$ на множество классов $y(\omega)\colon \Omega \to Y$, зависит главным образом от того, насколько удачно выбрана модальность представления объектов $x(\omega) \in \mathbb{X}$.

Во многих практических случаях ни одна отдельно взятая модальность не обеспечивает достаточной надежности распознавания, а необходимость повышения обобщающей способности решающего правила приводит к концепции *многомодальных систем*, комбинирующих несколько разных способов представления объектов $\big(x_i(\omega) \in \mathbb{X}_i,\ i = 1, \ldots, n\big)$ в единой процедуре распознавания образов $\hat{y}\big(x_1(\omega), \ldots, x_n(\omega)\big)$.

Различают два уровня комбинирования модальностей [1] — *уровень сенсоров* [2], когда до обучения классификатора формируется единое представление объектов, комбинирующее все частные модальности, и *уровень классификаторов* [3], когда комбинируются классификаторы, обученные по каждой модальности в отдельности.

До последнего времени основное внимание в литературе уделялось принципам комбинирования классификаторов, поскольку считалось, что комбинировать модальности разной физической природы затруднительно, либо в принципе невозможно.

Однако развитие метода потенциальных функций, основанного на идее единообразного представления объектов любой природы в виде элементов линейного пространства, и появление методов комбинирования нескольких разных потенциальных функций [2] открывает путь к комбинированию модальностей фактически на уровне сигналов сенсоров.

Обучение по единственной модальности

В рамках метода потенциальных функций [4] для построения двухклассового классификатора по единственной модальности $x_i(\omega)$ необходимо выбрать потенциальную функцию $K(x_i(\omega'), x_i(\omega''))$, $\omega', \omega'' \in \Omega_i^*$, адекватную искомой классификации, определить направляющий элемент $\vartheta \in \mathbb{X}_i$ и сдвиг $b \in \mathbb{R}$ линейной дискриминантной функции:

$$f_i(x_i(\omega)) = K(\vartheta, x_i(\omega)) + b \quad \begin{cases} < 0 & \to \hat{y}(x_i(\omega)) = -1; \\ > 0 & \to \hat{y}(x_i(\omega)) = 1. \end{cases}$$

Если $f_i(x_i(\omega)) = 0$, то, строго говоря, решение о принадлежности объекта $\omega \in \Omega$ к одному из классов не может быть принято. Такие точки пространства $\mathbb{X}_{\varnothing,i} = \{x_i(\omega) \in \mathbb{X}_i \mid f_i(x_i(\omega)) = 0\}$ будем называть *нейтральными точками* [5] обобщенного признака $\mathbb{X}_i$.

Обучение по непересекающимся обучающим совокупностям: Метод нейтральной точки

При построении мультимодальных систем типична ситуация, когда разные модальности разрабатываются разными группами специалистов независимо друг от друга. В этом случае общая обучающая совокупность оказывается состоящей из непересекающихся подмножеств объектов $\Omega^* = \bigcup_{i=1}^n \Omega_i^*$, $\Omega_i^* \cap \Omega_j^* = \varnothing$, $i \neq j$, в пределах каждого из которых известны значения только одного обобщенного признака $x_i(\omega_j)$, $\omega_j \in \Omega_i^*$, следовательно, не представляется возможным напрямую строить решающее правило распознавания по методу комбинирования потенциальных функций [2].

В работе [5] предлагается рассматривать задачу обучения по непересекающимся совокупностям как задачу обучения по неполным данным.

Для объектов каждой совокупности $\omega_j \in \Omega_i^*$ неизвестные фактические значения их признаков по каждой из других модальностей $x_l(\omega_j),\ l \neq i$ заменяются одним общим значением $\hat{x}_{\emptyset,l} \in \mathbb{X}_{\emptyset,l}$, которое представляет собой нейтральную точку соответствующего обобщенного признака.

При таком способе восполнения недостающих значений обобщенных признаков метод комбинирования потенциальных функций [2] сводится к методу комбинирования классификаторов, известный в англоязычной литературе как Sum Rule of Classifier Fusion.

Экспериментальное сравнение комбинирования классификаторов и потенциальных функций

В работе [6] показано, что при независимых модальностях, имеющих примерно равную информативность, комбинирование классификаторов по методу Sum Rule предпочтительнее по ошибке на генеральной совокупности в сравнении с комбинированием потенциальных функций.

В случае независимых модальностей с существенно разной информативностью, а так же зависимых модальностей комбинирование потенциальных функций дает лучшие результаты на генеральной совокупности.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00679, № 06-01-08042, № 06-07-89249, а также INTAS, проект № 04-77-7347.

Литература

[1] *Ross A., Jain A.* Multimodal biometrics: An overview // 12th European Signal Processing Conference (EUSIPCO), Vienna, Austria, 2004. — С. 1221–1224.

[2] *Mottl V., Tatarchuk A., Seredin O., Krasotkina O., Sulimova V.* Combining pattern recognition modalities at the sensor level via kernel fusion // 7th International Workshop on Multiple Classifier Systems, Prague, Czech Republic, 2007. — С. 1–12.

[3] *Kittler J., Hatef M., Duin R., Matas J.* On combining classifiers. // IEEE Trans. on Patt. Anal. and Mach. Intelligence. — 1998. — Т. 20, No 3. — С. 226–239.

[4] *Vapnik V.* Statistical Learning Theory. John-Wiley & Sons, Inc. 1998.

[5] *Windridge D., Mottl V., Tatarchuk A., Eliseyev A.* The neutral point method for kernel-based combining disjoint training data in multi-modal pattern recognition // 7th International Workshop on Multiple Classifier Systems, Prague, Czech Republic, 2007. — С. 13–21.

[6] *Windridge D., Mottl V., Tatarchuk A., Eliseyev A.* The relationship between kernel and classifier fusion in kernel-based multi-modal pattern recognition: An experimental study // Int. Conf. on Machine Learning and Cybernetics, August 19-22, 2007, Hong Kong, China.

Об оптимальном выборе закономерностей, составляющих плавно меняющуюся закономерность

Филипенков Н. В.
filipenkov@mail.ru
Москва, Вычислительный Центр РАН

В работе [2] был предложен подход к поиску плавно меняющихся закономерностей в пучках временных рядов. Идея подхода состоит в разбиении исходного пучка временных рядов на отрезки, на каждом из которых применяется алгоритм поиска постоянных закономерностей. Наиболее близкие (в смысле определённой в работе [2] меры сходства) закономерности, полученные на различных отрезках, «склеиваются» в плавно меняющуюся закономерность. Однако в упомянутой работе достаточно слабо был освещён вопрос выбора постоянных закономерностей при построении плавно меняющейся закономерности. Настоящая работа ставит своей целью заполнить этот пробел.

Пучком временных рядов $\mathfrak{S}$ называется совокупность взаимосвязанных временных рядов $S_i,\ i = 1,\ldots,N$. Каждый ряд S_i представляет собой последовательность чисел конечнозначной логики E_{k_i}. Значение ряда S_i в момент времени $t \in \{1,\ldots,T\}$ обозначим $a(i,t)$. *Маской* ω на прямоугольнике $N \times \Delta$ называется булева матрица размерности $N \times \Delta$ (здесь параметр Δ определяет максимальный отступ по времени). Число единиц в маске ω называется *мощностью* маски и обозначается $\|\omega\|$. Элемент маски, находящийся в i-ой строке и j-ом столбце обозначается $\omega(i,j)$. *Закономерностью* R (постоянной) называется набор (p,ω,f), где число $p \in \{1,\ldots,N\}$ указывает на целевой ряд (то есть ряд, значения которого определяются закономерностью R); маска ω указывает на значения рядов, являющиеся аргументами функции f; частично-определённая функция f задаёт зависимость значений целевого ряда от переменных, на которые указывает маска ω.

$$f\colon E_{k_{i_1}} \times \ldots \times E_{k_{i_{\|\omega\|}}} \to E_{k_p} \cup \{\lambda\},$$

где $\omega(i_1,j_1),\ldots,\omega(i_{\|\omega\|},j_{\|\omega\|})$ — единичные элементы матрицы ω, p — номер целевого ряда, символ λ обозначает, что функция f не определена на соответствующем наборе значений переменных.

Отрезком $\mathfrak{S}_1$ с *началом* $t_b \in \{0,\ldots,T\}$ и *концом* $t_e \in \{0,\ldots,T\}$, $(t_b < t_e)$ на пучке временных рядов $\mathfrak{S}$ (обозначается $\mathfrak{S}_1 \subset \mathfrak{S}$) назовём матрицу $N \times \theta$, составленную из последовательных столбцов матрицы $\mathfrak{S}$, первым из которых является столбец с номером t_b, последним — столбец с номером t_e, где $\theta = t_e - t_b + 1$ называется *длиной* отрезка $\mathfrak{S}_1$.

Определим следующие показатели качества постоянных закономерностей. Их названия совпадают с принятыми в нечёткой логике [1] показателями качества правил, так как несут сходный смысл. *Репрезентативностью* $\mathrm{rep}(R)$ $(0 \leqslant \mathrm{rep}(R) \leqslant 1)$ закономерности $R = (p, \omega, f)$ назовём следующую величину:

$$\mathrm{rep}(R) = \frac{|D(\omega)|}{k_{i_1} \cdots k_{i_{\|\omega\|}}},$$

где $|D(\omega)|$ — число наборов из $E_{k_{i_1}} \times \ldots \times E_{k_{i_{\|\omega\|}}}$, на которых значение функции f отлично от λ, а $\omega(i_1, j_1), \ldots, \omega(i_{\|\omega\|}, j_{\|\omega\|})$ — единичные элементы матрицы ω.

Локальной эффективностью $\mathrm{eff}_\varepsilon(R, t)$ $(0 \leqslant \mathrm{eff}(R) \leqslant 1)$ закономерности R в точке t назовём следующую величину:

$$\mathrm{eff}_\varepsilon(R, t) = 1 - \frac{1}{2\varepsilon + 1} \sum_{i=-\varepsilon}^{\varepsilon} \Big(\frac{\hat{a}(t+i) - a(p, t+i)}{k_p} \Big)^2,$$

где $\hat{a}(t)$ — прогноз закономерности R для значения $a(p, t)$ пучка временных рядов.

Плавно меняющейся закономерностью $\tilde{R} \in \mathfrak{R}^{T-\Delta}$ на пучке временных рядов $\mathfrak{S}$ называется последовательность постоянных закономерностей $R_{\Delta+1}, \ldots, R_T$ такая, что элементы $a(p, \Delta+1), \ldots, a(p, T)$ пучка временных рядов $\mathfrak{S}$ прогнозируются соответственно постоянными закономерностями $R_{\Delta+1}, \ldots, R_T$. При этом $R_{\Delta+1}, \ldots, R_T$ могут представлять собой одни и те же закономерности.

Определим три основных показателя качества плавно меняющейся закономерности в точке пучка временных рядов.

Локальной репрезентативностью $\widetilde{\mathrm{rep}}_\varepsilon(\tilde{R}, t)$ и *локальной эффективностью* $\widetilde{\mathrm{eff}}_\varepsilon(\tilde{R}, t)$ плавно меняющейся закономерности $\tilde{R}$ в ε-окрестности точки t пучка временных рядов $\mathfrak{S}$ называются следующие величины:

$$\widetilde{\mathrm{rep}}(\tilde{R}, t) = \mathrm{rep}(R_t), \qquad \widetilde{\mathrm{eff}}_\varepsilon(\tilde{R}, t) = \mathrm{eff}_\varepsilon(R_t, t).$$

Третьим показателем качества плавно меняющейся закономерности в точке t пучка временных рядов $\mathfrak{S}$ является значение $\rho(R_t, R_{t+1})$ — меры сходства закономерностей R_t и R_{t+1}. Для точки T данный показатель принимается равным нулю. Определение меры сходства закономерностей подробно рассмотрено в работе [2]. Заметим, что значение меры сходства закономерностей несёт смысл штрафа за смену закономерностей, и этот штраф тем больше, чем больше различие между закономерностями.

На основании приведённых выше локальных показателей качества плавно меняющейся закономерности $\tilde{R}$ определяются показатели качества $\tilde{R}$ на всём пучке временных рядов. *Средней репрезентативностью* $\widetilde{\mathrm{rep}}(\tilde{R})$ и *средней эффективностью* $\widetilde{\mathrm{eff}}(\tilde{R})$ закономерности $\tilde{R}$ называются следующие показатели:

$$\widetilde{\mathrm{rep}}(\tilde{R}) = \frac{\sum_{i=\Delta+1}^{T} \widetilde{\mathrm{rep}}(\tilde{R},t)}{T-\Delta}, \qquad \widetilde{\mathrm{eff}}(\tilde{R}) = \frac{\sum_{i=\Delta+1}^{T} \widetilde{\mathrm{eff}}_0(\tilde{R},t)}{T-\Delta}.$$

Здесь $\widetilde{\mathrm{eff}}_0(\tilde{R},t)$ — локальная эффективность $\widetilde{eff}_\varepsilon(\tilde{R},t)$ при $\varepsilon = 0$.

Длиной $l(\tilde{R})$ изменяющейся закономерности $\tilde{R}$ называется следующий показатель качества плавно меняющейся закономерности $\tilde{R}$:

$$l(\tilde{R}) = \sum_{t=\Delta+1}^{T-1} \rho(R_t, R_{t+1}).$$

Поиск наилучшей плавно меняющейся закономерности $\tilde{R}$ сводится к задаче многокритериальной оптимизации:

$$\begin{cases} \widetilde{\mathrm{rep}}(\tilde{R}) \to \max\limits_{\tilde{R}\in\mathfrak{R}^{T-\Delta}}; \\ \widetilde{\mathrm{eff}}(\tilde{R}) \to \max\limits_{\tilde{R}\in\mathfrak{R}^{T-\Delta}}; \\ l(\tilde{R}) \to \min\limits_{\tilde{R}\in\mathfrak{R}^{T-\Delta}}. \end{cases}$$

Здесь оптимизация происходит по всем возможным закономерностям $\tilde{R}$, т. е. по всем последовательностям постоянных закономерностей $R_{\Delta+1}, \ldots, R_T$.

Литература

[1] *Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И.* Методы и модели анализа данных: OLAP и Data Mining. — СПб.: БХВ-Петербург, 2004.

[2] *Филипенков Н. В.* О задачах анализа пучков временных рядов с изменяющимися закономерностями // Искусственный интеллект — Донецк, 2006. — № 2. — С. 125–129.

Задача обучения распознаванию образов в нестационарной генеральной совокупности

Шавловский М. Б., Красоткина О. В., Моттль В. В.
shavlovsky@yandex.ru, ko180177@yandex.ru, vmottl@yandex.ru
Москва, МФТИ; Тула, ТулГУ; Москва, ВЦ РАН

Целью данного исследования является создание основного математического аппарата и простейших алгоритмов для решения типичных для практики задач обучения распознаванию образов в генеральных совокупностях, свойства которых изменяются во времени. Широко известная классическая постановка задачи распознавания основана на молчаливом предположении, что свойства генеральной совокупности в момент «экзамена» остаются теми же, что и при формировании обучающей выборки. Принятое в данной работе более реалистичное предположение о нестационарности генеральной совокупности неизбежно приводит к необходимости анализа последовательности выборок в некоторые моменты времени и поиска для них разных решающих правил распознавания.

Пусть каждый объект генеральной совокупности $\omega \in \Omega$ представлен точкой в линейном пространстве признаков $\mathbf{x}(\omega) = \big(x^1(\omega), \ldots, x^n(\omega)\big) \in \in \mathbb{R}^n$, а его скрытая фактическая принадлежность к одному из двух классов определяется значением индекса класса $y(\omega) \in \{1, -1\}$. Классический подход к обучению распознаванию двух классов объектов, развитый В. Н. Вапником [1], основан на понимании модели генеральной совокупности в виде дискриминантной функции $f\big(\mathbf{x}(\omega)\big) = \mathbf{a}^T\mathbf{x} + b$, определяемой гиперплоскостью с априори неизвестными наблюдателю направляющим вектором $\mathbf{a} \in \mathbb{R}^n$ и параметром положения $b \in \mathbb{R}$:

$$f\big(\mathbf{x}(\omega)\big) = \mathbf{a}^T\mathbf{x} + b \begin{cases} \text{преимущественно } > 0, & \text{если } y(\omega) = 1, \\ \text{преимущественно } < 0, & \text{если } y(\omega) = -1. \end{cases}$$

Неизвестные параметры разделяющей гиперплоскости подлежат оцениванию на основе анализа обучающей совокупности объектов $\{\omega_j,\ j = \\ = 1, \ldots, N\}$, представленных векторами их признаков и индексами принадлежности к классам, так что выборка в целом является конечным множеством пар $\big\{(\mathbf{x}_j \in \mathbb{R}^n,\ y_j = \pm 1),\ j = 1, \ldots, N\big\}$. Широко известен принцип оптимальной разделяющей гиперплоскости, выбираемой по критерию максимизации числа точек обучающей совокупности, правильно классифицируемых с гарантированным «запасом», условно равным единице:

$$\begin{cases} J(\mathbf{a}, b, \delta_j) = \mathbf{a}^T\mathbf{a} + C\sum\limits_{j=1}^{N}\delta_j \to \min, \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \geqslant 1 - \delta_j, \quad \delta_j \geqslant 0, \quad j = 1, \ldots, N. \end{cases} \tag{1}$$

Понятие времени здесь полностью отсутствует.

Принципиальное отличие предлагаемой в данной работе концепции нестационарной генеральной совокупности заключается во введении в рассмотрение фактора времени t. Предполагается, что основное свойство нестационарной генеральной совокупности выражается изменяющейся во времени разделяющей гиперплоскостью, характеризующей преимущественное различие векторов признаков объектов двух классов и, в свою очередь, полностью определяемой своим направляющим вектором и параметром положения, которые должны рассматриваться как функции времени $\mathbf{a}_t$ и b_t:

$$f_t\big(\mathbf{x}(\omega)\big) = \mathbf{a}_t^T\mathbf{x}+b_t \begin{cases} \text{преимущественно} > 0 \text{ в момент } t, & \text{если } y(\omega) = 1, \\ \text{преимущественно} < 0 \text{ в момент } t, & \text{если } y(\omega) = -1. \end{cases}$$

Здесь всякий объект $\omega \in \Omega$ рассматривается всегда только вместе с указанием момента времени, в который он предъявлен (ω, t). В результате обучающая совокупность приобретает структуру множества троек $\big\{(\mathbf{x}_j \in \mathbb{R}^n,\ y_j = \pm 1,\ t_j),\ j = 1, \ldots, N\big\}$, а не пар. Естественно нумеровать объекты обучающей совокупности в порядке поступления объектов, тогда уместно говорить скорее об обучающей последовательности, нежели об обучающей совокупности, рассматривая ее как временной ряд, вообще говоря, с переменным шагом по времени.

В разные момент времени t_j скрытая от наблюдателя разделяющая гиперплоскость характеризуется разными неизвестными значениями направляющего вектора и параметра положения. Таким образом, объективно существует двухкомпонентный временной ряд со скрытой и наблюдаемой компонентами, соответственно, $(\mathbf{a}_j, b_j)$ и $(\mathbf{x}_j, y_j)$.

В динамической постановке задача обучения превращается в задачу анализа двухкомпонентного временного ряда, в котором требуется, анализируя наблюдаемую компоненту, дать оценку скрытой компоненты. Это стандартная задача анализа сигналов (временных рядов), специфика которой заключается лишь в предполагаемой модели связи между скрытой и наблюдаемой компонентами. Согласно классификации задач оценивания скрытой компоненты сигнала, введенной Н. Винером [2], естественно различать, по крайнем мере, два вида задач обучения.

Задача фильтрации обучающей последовательности. Пусть t_j — момент поступления очередного объекта, к которому уже зарегистрированы векторы признаков и индексы классов объектов $\big\{\ldots, (\mathbf{x}_{j-2}, y_{j-2}), (\mathbf{x}_{j-1}, y_{j-1}), (\mathbf{x}_j, y_j)\big\}$, поступивших в предыдущие моменты времени до текущего момента включительно $(\ldots, t_{j-2}, t_{j-1}, t_j)$. Требуется непосредственно в процессе наблюдения давать оценку параметров разделяющей гиперплоскости $(\hat{\mathbf{a}}_j, \hat{b}_j)$ в каждый текущий момент времени t_j.

Задача интерполяции. Пусть к моменту обработки обучающая последовательность уже зарегистрирована в некотором интервале времени $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$. Требуется оценить изменяющиеся параметры разделяющей гиперплоскости во всем интервале наблюдения $\{(\hat{\mathbf{a}}_1, \hat{b}_1), \ldots, (\hat{\mathbf{a}}_N, \hat{b}_N)\}$.

Предполагается, что параметры разделяющей гиперплоскости $\mathbf{a}_t$ и b_t изменяются во времени достаточно медленно в том смысле, что величины

$$\tfrac{1}{t_j - t_{j-1}}(\mathbf{a}_j - \mathbf{a}_{j-1})^{\mathrm{T}}(\mathbf{a}_j - \mathbf{a}_{j-1}) \approxeq \varepsilon_{\mathbf{a}} \quad \text{и} \quad \tfrac{1}{t_j - t_{j-1}}(b_j - b_{j-1})^2 \approxeq \varepsilon_b$$

являются, как правило, достаточно малыми. Это предположение препятствует вырождению задач фильтрации и интерполяции в совокупность независимых некорректных задач обучения распознаванию двух классов объектов по единственному наблюдению.

С формальной точки зрения оценка параметров разделяющей гиперплоскости в последний момент интервала наблюдения $(\hat{\mathbf{a}}_N, \hat{b}_N)$, полученная при решении задачи интерполяции, является решением задачи фильтрации для этого момента. Однако смысл задачи фильтрации заключается в том, чтобы очередные оценки вычислялись непосредственно в процессе поступления новых наблюдений, без решения всякий раз задачи интерполяции для временного ряда возрастающей длины.

Предлагаемая постановка задачи обучения в режиме интерполяции отличается от совокупности классических задач обучения по методу опорных векторов (1) для каждого момента времени только наличием дополнительных членов, штрафующих различие между смежными значениями параметров гиперплоскости $(\mathbf{a}_{j-1}, b_{j-1})$ и $(\mathbf{a}_j, b_j)$:

$$\begin{cases} J(\mathbf{a}_j, b_j, \delta_j, j = 1, \ldots, N) = \sum\limits_{j=1}^{N} \left(\frac{1}{N}\mathbf{a}_j^T \mathbf{a}_j + \delta_j\right) + \\ + \sum\limits_{j=2}^{N} \frac{1}{t_j - t_{j-1}} \left[D^{\mathbf{a}}(\mathbf{a}_j - \mathbf{a}_{j-1})^{\mathrm{T}}(\mathbf{a}_j - \mathbf{a}_{j-1}) + D^b(b_j - b_{j-1})^2\right] \to \min, \\ y_j(\mathbf{a}_j^T \mathbf{x}_j + b_j) \geqslant 1 - \delta_j, \quad \delta_j \geqslant 0, \quad j = 1, \ldots, N. \end{cases} \tag{2}$$

Здесь коэффициенты $D^{\mathbf{a}} > 0$ и $D^b > 0$ являются параметрами критерия, задающими желаемую степень сглаживания оцениваемой последовательности мгновенных значений параметров разделяющей гиперплоскости.

Критерий (2) реализует концепцию оптимальной достаточно гладкой последовательности разделяющих гиперплоскостей в отличие от концепции единственной оптимальной гиперплоскости в (1). Искомые гиперплоскости должны обеспечивать правильную классификацию векторов признаков объектов для как можно большего числа моментов времени с гарантированным «запасом», принятым равным единице, как и в (1).

Как и классическая задача обучения, динамическая задача (2) является задачей квадратичного программирования, но содержит $N(n+1)+N$ переменных, в отличие от $(n+1)+N$ переменных в (1). Известно, что вычислительная сложность задачи квадратичного программирования общего вида пропорциональна кубу числа переменных, т. е. динамическая задача, на первый взгляд, существенно сложнее классической.

Однако целевая функция $J(\mathbf{a}_j, b_j, \delta_j,\, j = 1, \ldots, N)$ в динамической задаче является парно-сепарабельной, т.е. представляя собой сумму частных функций, каждая из которых зависит от переменных, связанных только с одним либо двумя моментами времени в порядке их возрастания. Это обстоятельство позволяет построить алгоритм численного решения задачи, вычислительная сложность которого линейна относительно длины обучающей последовательности N.

Применение теоремы Куна-Таккера к динамической задаче (2) переводит ее в двойственную форму относительно множителей Лагранжа $\lambda_j \geqslant 0$ при ограничениях-неравенствах $y_j(\mathbf{a}_j^T \mathbf{x}_j + b_j) \geqslant 1 - \delta_j$:

$$\begin{cases} W(\lambda_1, \ldots, \lambda_N) = \\ \quad = \sum\limits_{j=1}^{N} \lambda_j - \frac{1}{2} \sum\limits_{j=1}^{N} \sum\limits_{l=1}^{N} y_j y_l \big(\mathbf{a}_j^{\mathrm{T}} \mathbf{Q}_{jl} \mathbf{a}_l + f_{jl}\big) \lambda_j \lambda_l \to \max, \\ \sum\limits_{j=1}^{N} y_j \lambda_j = 0, \quad 0 \leqslant \lambda_j \leqslant C/2, \quad j = 1, \ldots, N. \end{cases} \tag{3}$$

Здесь матрицы $\mathbf{Q}_{jl} (n \times n)$ и $\mathbf{F} = (f_{jl})\,(N \times N)$ не зависят от обучающей последовательности и определяются только коэффициентами штрафа, соответственно, $D^{\mathbf{a}}$ на негладкость последовательности направляющих векторов искомых гиперплоскостей и D^b на негладкость последовательности их параметров положения в (2).

Теорема 1. *Решение задачи обучения (2) полностью определяется значениями множителей Лагранжа $(\lambda_1, \ldots, \lambda_N)$, полученными как решение двойственной задачи (3), и обучающей последовательностью:*

$$\hat{\mathbf{a}}_j = \sum_{l:\ \lambda_l > 0} y_l \lambda_l \mathbf{Q}_{jl} \mathbf{x}_l; \qquad \hat{b}_j = b + \sum_{l:\ \lambda_l > 0} y_l \lambda_l f_{jl}; \tag{4}$$

$$b = \frac{\sum\limits_{j:\ 0 < \lambda_j < C/2} \lambda_j \sum\limits_{l:\ \lambda_j > 0} y_l \lambda_l \big(\mathbf{x}_l^{\mathrm{T}} \mathbf{Q}_{jl} \mathbf{x}_j + f_{jl}\big) + (C/2) \sum\limits_{j:\ \lambda_j = C/2} y_j}{\sum\limits_{j: 0 < \lambda_j < C/2} \lambda_j}. \tag{5}$$

Из этих формул видно, что решение задачи динамического обучения определяется только теми элементами обучающей последовательности

$(\mathbf{x}_j, y_j)$, множители Лагранжа при которых получили положительные значения $\lambda_j > 0$. Уместно назвать векторы признаков соответствующих объектов опорными векторами, так что мы пришли к некоторому обобщению метода опорных векторов [1], вытекающего из концепции оптимальной разделяющей гиперплоскости (1).

Классическая задача обучения (1) является частным случаем задачи (2) при бесконечно больших значениях штрафов на изменение параметров гиперплоскости $D^{\mathbf{a}} \to \infty$ и $D^b \to \infty$. В этом случае $\mathbf{Q}_{jl} \to \mathbf{I}$, $f_{jl} \to 0$, и двойственная задача (3) превращается в классическую двойственную задачу [1], соответствующую исходной задаче (1), а формулы (4) и (5) определяют результат обучения по классическому методу опорных векторов $\hat{\mathbf{a}} = \hat{\mathbf{a}}_1 = \ldots = \hat{\mathbf{a}}_N$, $\hat{b} = \hat{b}_1 = \ldots = \hat{b}_N$.

Хотя двойственная задача (3) и не является парно-сепарабельной, в силу парной сепарабельности исходной задачи (2) вычисление градиента целевой функции $\nabla_{\boldsymbol{\lambda}} W(\lambda_1, \ldots, \lambda_N)$ в любой точке $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ и, далее, определение оптимального допустимого направления поиска с учетом ограничений обеспечиваются алгоритмом, имеющим линейную вычислительную сложность относительно длины обучающей последовательности. В результате оказывается, что использование любого градиентного метода для решения двойственной задачи приводит к алгоритму с линейной вычислительной сложностью относительно N. В частности, стандартный метод наискорейшего спуска для решения задач квадратичного программирования [3], примененный к функции $-W(\lambda_1, \ldots, \lambda_N)$, дает алгоритм, являющийся, по сути, обобщением известного алгоритма SMO (Sequential Minimum Optimization) [4], обычно используемого при решении двойственных задач.

Работа выполнена при поддержке РФФИ, проекты №05-01-00679, №06-01-00412.

Литература

[1] *Vapnik V.* Statistical Learning Theory. — New York: John-Wiley & Sons, Inc., 1998. — 732 p.

[2] *Wiener N.* Extrapolation, Interpolation, and Smoothing of Stationary Random Time Series with Engineering Applications. — Technology Press of MIT, John Wiley & Sons, 1949. — 163 p.

[3] *Базара М., Шетти К.* Нелинейное программирование. Теория и алгоритмы. — М.: Мир, 1982.

[4] *Platt J.C.* Fast training of support vector machines using sequential minimal optimization / Advances in Kernel Methods: Support Vector Learning. — MIT Press, Cambridge, MA, 1999.

Об одном алгебраическом подходе к обучению в теоретической нейроинформатике

Шибзухов З. М.
szport@gmail.com
НИИ прикладной математики и автоматизации КБНЦ РАН, Нальчик

В теоретической нейроинформатике известно, что любое непрерывное преобразование, определенное на компактном множестве, можно аппроксимировать при помощи сети из аддитивных сумматоров и функциональных преобразователей, реализующих произвольно заданную непрерывную нелинейную скалярную функцию [1, 2]. Из него следует, что любое непрерывное преобразование, определенное на компактном множестве, можно аппроксимировать при помощи сети из элементов, реализующих операции сложения, умножения, и произвольно выбранные нелинейные скалярные функции. Такие сети называют *полиномиальными сетями*. Оказывается, что можно построить *прямую алгебраическую процедуру* [3] для построения некоторых достаточно представительных классов полиномиальных сетей, обрабатывающих сигналы, кодируемые в произвольно заданном *коммутативном кольце*, не содержащем делителей нуля [4]. С ее помощью можно эффективно конструировать искусственные нейронные сети для решения задач распознавания и прогнозирования. Ниже рассматривается один класс полиномиальных нейронных сетей, содержащих специальный слой алгебраических *мультиплицирующих нейронов* и слой алгебраических *суммирующих нейронов*. Такая организация сети делает возможным построение прямой алгебраической процедуры обучения с учителем.

Алгебраические мультиплицирующие функции

Пусть $\mathbb{K}$ — это коммутативное кольцо, не содержащее делителей нуля, $\mathbb{X}$ — скалярная область значений входов, $\mathbb{Y}$ — скалярная область значений выходов. Пусть $\mathcal{F}_{0,1}(\mathbb{K})$ — некоторый класс функций $\eta\colon \mathbb{K} \to \mathbb{K}$, таких, что $\eta(p) = 0 \Leftrightarrow p = 0$ и $\eta(1) = 1$. Рассмотрим класс $\mathfrak{P}[x_1, \ldots, x_n]$ всех алгебраических мультиплицирующих функций, которые можно построить из тождественных функций $\mathrm{id}(x_1) = x_1$, …, $\mathrm{id}(x_n) = x_n$, применяя операции умножения и композиции функций из $\mathcal{F}_{0,1}(\mathbb{K})$. Например, функции вида $\eta\bigl(\prod_{i=1}^{n} \varphi_i(x_i)\bigr)$, где $\eta \in \mathcal{F}_{0,1}(\mathbb{K})$ и $\varphi_i \in \mathcal{F}_{0,1}(\mathbb{K})$, принадлежат $\mathfrak{P}^n[x_1, \ldots, x_n]$.

Пусть $\Pi(\boldsymbol{x}) \in \mathfrak{P}[x_1, \ldots, x_n]$ — функция n переменных $\boldsymbol{x} = (x_1, \ldots, x_n)$. Для любого мультииндекса $\mathbf{i} \subset \{1, \ldots, n\}$ определена функция $\Pi(\boldsymbol{x}, \mathbf{i})$ вида $\mathbb{K}^r \to \mathbb{K}$, зависящая от набора переменных $\{x_i : i \in \mathbf{i}\}$, $r = |\,\mathbf{i}\,|$, и получающаяся из $\Pi(\boldsymbol{x})$ в результате «исключения переменных» с индексами из $\{i\colon i \notin \mathbf{i}\}$.

Множество функций $\mathcal{P}[\boldsymbol{x}] = \big\{\Pi(\boldsymbol{x},\mathbf{i})\colon \mathbf{i} \subseteq \{1,\ldots,n\}\big\}$ удовлетворяет следующим условиям:

1) если $\Pi(\boldsymbol{x},\mathbf{i}) = 0$, то $\Pi(\boldsymbol{x},\mathbf{i}') = 0$ для любого $\mathbf{i}' \supset \mathbf{i}$;
2) если $\Pi(\boldsymbol{x},\mathbf{i}) \neq 0$, то $\Pi(\boldsymbol{x},\mathbf{i}') \neq 0$ для любого $\mathbf{i}' \subset \mathbf{i}$.

Алгебраические суммирующие функции

Пусть $\mathbb{A}$ — некоторый $\mathbb{K}$–модуль. Пусть $\mathcal{F}_0(\mathbb{A})$ — некоторый класс отображений $\sigma\colon \mathbb{A} \to \mathbb{A}$ таких, что $\sigma(s) = 0 \Leftrightarrow s = 0$. Рассмотрим класс $\mathfrak{S}[s_1,\ldots,s_N]$ всех алгебраических суммирующих функций, которые можно построить из тождественных функций $\mathrm{id}(s_1) = s_1$, …, $\mathrm{id}(s_n) = s_n$, применяя операции сложения, умножения на скаляры из $\mathbb{K}$ и композиции функций из $\mathcal{F}_0(\mathbb{A})$. Например, множества функций, определяемые индуктивно:

1) $\Sigma(s_1) = \sigma_1(s_1)$;
2) $\Sigma(s_1,\ldots,s_k) = \sigma_k\big(\Sigma(s_1,\ldots,s_{k-1}) + w_k s_k\big)$ при $k > 1$;

где $\sigma_k \in \mathcal{F}_0(\mathbb{A})$, $k = 1,\ldots,N$, принадлежат $\mathfrak{S}[s_1,\ldots,s_N]$.

Алгебраические ΣΠ–функции

Алгебраическая ΣΠ–функция реализует преобразование

$$\mathrm{spn}(\boldsymbol{x}) = \mathrm{out}\big(\mathrm{sp}(\boldsymbol{x})\big); \tag{1}$$

$$\mathrm{sp}(\boldsymbol{x}) = \Sigma\big(\theta(\boldsymbol{x}), \Pi_1(\boldsymbol{x},\mathbf{i}_1),\ldots,\Pi_N(\boldsymbol{x},\mathbf{i}_N)\big); \tag{2}$$

где $\theta\colon \mathbb{X}^n \to \mathbb{A}$, $w_k \in \mathbb{A}$, $\Sigma(s_0,\ldots,s_N) \in \mathfrak{S}[s_0,\ldots,s_N]$, $\Pi_k(\boldsymbol{x}) \in \in \mathfrak{P}[x_1,\ldots,x_n]$. Преобразование $\mathrm{out}\colon \mathbb{A} \to \mathbb{Y}$ является *допустимым*, т. е. для любых $s \in \mathbb{A}$, $0 \neq p \in \mathbb{K}$, $y \in \mathbb{Y}$ уравнение $\mathrm{out}(s + wp) = y$ имеет решение относительно w.

При $m = 1$ и $\mathbb{A} = \mathbb{K}$ соотношения (1)–(2) описывают алгебраический ΣΠ-нейрон $\mathrm{spn}(\boldsymbol{x}) = \mathrm{out}\big(\mathrm{sf}(\boldsymbol{x})\big)$. Если $m = 1$ и $\mathbb{A}$ — $\mathbb{K}$-модуль размерности, большей 1, то соотношения (1)–(2) описывают коллектив ΣΠ-нейронов, принимающих единственное решение, или алгебраический ΣΠ-корректор.

Треугольно упорядоченные последовательности алгебраических мультиплицирующих функций

Пусть $\mathbf{X} = \{\mathbf{x}_k\}$ — некоторая последовательность входов $\mathbf{x}_k \in \mathbb{X}^n$, $\mathbf{P} = \{\Pi_k(\boldsymbol{x},\mathbf{i}_k)\}$ — последовательность алгебраических мультиплицирующих функций из $\mathfrak{P}[x_1,\ldots,x_n]$, треугольно упорядоченная на $\mathbf{X}$, т. е. таких, что

1) $\Pi_k(\mathbf{x}_j,\mathbf{i}_k) = 0$ для любой пары $1 \leqslant j < k$;
2) $\Pi_k(\mathbf{x}_k,\mathbf{i}_k) \neq 0$ для всех k.

Она является минимальной, если каждая последовательность $\{\Pi_k(\boldsymbol{x}, \mathbf{i}'_k)\}$, где $\mathbf{i}'_k \subseteq \mathbf{i}_k$, не является треугольно упорядоченной на $\mathbf{X}$.

При определенных условиях для некоторых достаточно широких классов последовательностей $\{\mathbf{x}_k\}$ существуют конструктивные минимальные треугольно упорядоченные последовательности $\{\Pi_k(\boldsymbol{x}, \mathbf{i}_k)\}$ [4].

Лемма 1. *(о минимальных треугольно упорядоченных последовательностях алгебраических мультиплицирующих функций). Конструктивно перечисляются все минимальные последовательности* $\{\Pi_k(\boldsymbol{x}, \mathbf{i}^\star_k)\}$, *треугольно упорядоченные на* $\mathbf{X}$.

Рекуррентные последовательности алгебраических ΣΠ–функций

Пусть задана последовательность ожидаемых на выходе значений $\mathbf{Y} = \{y_k\}$, соответствующая $\mathbf{X}$. Определяется последовательность алгебраических ΣΠ-функций $\{\mathrm{spn}_k(\boldsymbol{x})\}$, где

$$\mathrm{sp}_k(\boldsymbol{x}) = \sigma_k\big(\mathrm{sp}_{k-1}(\boldsymbol{x}) + w_k\Pi_k(\boldsymbol{x}, \mathbf{i}_k)\big), \tag{3}$$

w_k — решение уравнения $\mathrm{out}\big(\sigma_k\big(\mathrm{sp}_{k-1}(\mathbf{x}_k) + w_k\Pi_k(\mathbf{x}_k, \mathbf{i}_k)\big)\big) = y_k$ относительно w_k, $\mathrm{sp}_0(\boldsymbol{x}) = \sigma_0\big(\theta(\boldsymbol{x})\big)$.

Лемма 2. *(о рекуррентной последовательности алгебраических* **ΣΠ***-функций). Для всех* $j = 1, \ldots, k$ *верно равенство* $\mathrm{spn}_k(\mathbf{x}_j) = y_j$.

Работа выполнена при поддержке ОМН РАН по проекту «Исследование конструктивных последовательностей алгебраических расширений распознающих алгоритмов» в 2007 году.

Литература

[1] *Gilev S. E., Gorban A. N.* On Completness of the Class of Functions Computable by Neural Networks // World Cong. on Neural Networks, 1996. — Pp. 984–991.

[2] *Горбань А. Н.* Возможности нейронных сетей. — Нейроинформатика, под ред. Новикова Е. А., Гл. 1. — Новосибирск: Наука, 1998. — С. 18–46.

[3] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания и классификации / Избр. науч. тр. — М.: Магистр, 1998. — С. 229–323.

[4] *Шибзухов З. М.* Конструктивные методы обучения ΣΠ-нейронных сетей. — М.: Наука, 2006 — 159 с.

Коллективные решения задачи кластерного анализа с помощью гиперграфов

Шмаков А. С.
ashmak@mail.ru

Исследуется выборка $S = x_1, \ldots, x_m$, заданная таблицей признаковых описаний

$$J_m(S) = \left\| \begin{matrix} a_{11} & \ldots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{m1} & \ldots & a_{md} \end{matrix} \right\| \tag{1}$$

на множестве вещественно-значимых признаков, где d — число признаков, а m — число объектов.

Пусть в результате работы некоторого алгоритма A исходная выборка разбита на l кластеров $K_1, \ldots, K_l$.

Определение 1. *Информационной матрицей для задачи кластеризации выборки (1) на l кластеров называется матрица $I = \|\alpha_{ij}\| \in R_{m\times l}$, такая что $\alpha_{ij} \in \{0, 1, \triangle\}$, причем $\alpha_{ij} = 1$, если $x_i \in K_j$; $\alpha_{ij} = 0$, если $x_i \notin K_j$; $\alpha_{ij} = \triangle$ соотвествует отказу от зачисления x_i в один из классов.*

В отличие от задач распознавания, где каждый кластер имеет свое априорное смысловое содержание, в задачах кластерного анализа кластеры могут нумероваться в произвольном порядке. Поэтому произвольная информационная матрица I, отличающаяся от I' лишь перестановкой столбцов, будет являться записью того же самого решения.

Пусть теперь рассматривается задача кластеризации исходной выбоки (1) на l кластеров коллективом из n алгоритмов $A^1, \ldots, A^n$, и, кроме того, пусть в результате анализа исследуемый выборки получены n матрицы оценок (информационных матриц) $I^1, \ldots, I^n$.

Определение 2. *Коллективной информационной матрицей $\hat{I} = \|\chi_{ij}\|$ для задачи кластеризации выборки (1) на l кластеров коллективом из n алгоритмов называется блочная матрица размера $m \times ln$, состовленная из информационных матриц $I^1, \ldots, I^n$, т.е. $\hat{I} = \|I_1 \cdots I_n\|$.*

На основании коллективной информационной матрицы можно ввести понятие гиперграфа сопряженности.

Определение 3. *Гиперграфом сопряженности G_H для задачи коллективной кластеризации называется гиперграф, для которого коллективная информационная матрица $\hat{I}$ является матрицей инцидентности.*

Для построения коллективного решения предлагается построить разбиения гиперграфа сопряженности на l узлов в следующем виде: пусть

дан гиперграф $G_H(V, E)$, где множество вершин $V = \{x_1, \ldots, x_n\}$ — объекты выборки, а множество ребер $E = \{e_j, j = 1, \ldots, ln\}$ — столбцы коллективной информационной матрицы. Вес вершин считается одинаковым и равным 1, а вес ребер задается множеством $\varphi_j = \sum_{i=1}^{m} \chi_{ij}$, $j = 1, \ldots, ln$, где χ_{ij} — элементы коллективной информационной матрицы.

Для решения поставленной задачи предлагается использовать известный алгоритм разделения гиперграфа — HMETIS. Упрощенно алгоритм работает следующим образом:

1. Coarsening phase — фаза стягивания или огрубления, на которой строится последовательность грубых приближений гиперграфа:
   - Edge coarsening — простейший способ группировки вершин состоит в выборе пары вершин, принадлежащих одному и тому же гиперребру.
   - Hyperedge coarsening — выбирается независимое множество гиперребер; вершины, принадлежащие одному гиперребру, объединяются.
2. Initial partitioning phase — фаза разделения, на которой наименьший граф подвергается декомпозиции: в качестве алгоритма деления может быть использовано спектральное деление, геометрическое деление или комбинаторное деление.
3. Refinement phase — фаза, на которой решение для наименьшего графа проецируется на следующий уровень и уточняется итерационным алгоритмом Kernighan-Lin.

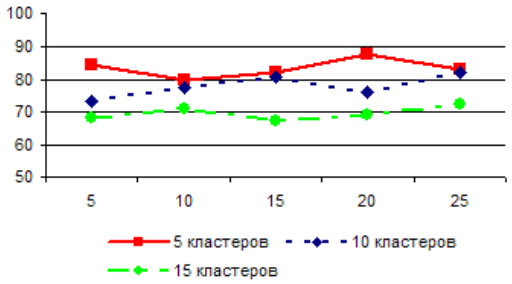
Таким образом, для решения задачи поиска локально оптимального решения по конечному набору кластеризаций предлагается следующий алгоритм:

1. Строим набор информационных матриц для каждого решения из коллектива $\{I_1, \ldots, I_n\}$;
2. Стром общую информационную матрицу для коллективного решения $\hat{I} = \|I_1 \cdots I_n\|$;
3. Строим гиперграф сопряженности H_G;
4. Используя алгоритм HMETIS, строим разделение гиперграфа H_G на l узлов;
5. Полученные в результате l подмножеств являются коллективным решением задачи.

Эксперименты с использованием гиперграфов проводились на выборках размера 500 объектов, состоящих из кластеров, которые имеют нормальные плотности распределения объектов по признакам. Были проведены испытания для выборок состоящих из $5, 10, 15$ кластеров. В качестве алгоритмов, участвующих в коллективах, использовались эври-

стические алгоритмы: k-внутригрупповых средних и алгоритм Форель. Количество решений для каждой задачи было 10. При исследованиях изучалась зависимость точности решения описываемым методом от возрастающей сложности решаемой задачи. Ниже приводится график зависимости точности кластеризации от числа признаков, описывающих выборку, при разном числе кластеров. На представленном ниже графике по оси Y отложено количество правильно классифицированных объектов в процентах, а по X — количество признаков, которые имела исследуемая выборка.

Работа выполнена при поддержке РФФИ, проект №05-07-90333, Целевой программы №14 Президиума РАН, Целевой программы №2 Отделения математических наук РАН.



Литература

[1] *Рязанов В. В.* О синтезе классифицирующих алгоритмов на конечных множествах алгоритмов классификации (таксономии).— ЖВМиМФ, 1982. — Т. 22, №2. — С. 429–440.

[2] *Рязанов В. В.* Комитетный синтез алгоритмов распознавания и классификации.— ЖВМиМФ, 1981. — Т. 21, №6. — С. 1533–1543.

[3] *Karypis G., Kumar V.* Multilevel k-way Hypergrap Partitioning. — VLSI Design. — Vol. 11, No. 3. — 2000, Pp. 285–300.

[4] *Karypis G., Selvakkumaran N.* Multi-Objective Hypergraph Partitioning Algorithms for Cut and Maximum Subdomain Degree Minimization. — IEEE Transactions on CAD, 2005.

Энтропийные методы исследования итеративных процедур коллективной оценки и выбора вариантов

Шоломов Л. А.
sholomov@isa.ru
Москва, Институт Системного анализа РАН

Содержательные задачи выбора лучших объектов, как правило, плохо формализованы (некорректны). Мы руководствуемся подходом к исследованию некорректных процедур, предложенным Ю. И. Журавлевым применительно к задачам распознавания и классификации, и состоящим в том, чтобы вместо конкретных процедур изучать строгими математическими методами свойства классов процедур [1]. Цель данной работы — рассмотрение с указанных позиций некоторого класса многотуровых процедур оценки вариантов. Подробное изложение имеется в [2].

Пусть для выбора лучшего объекта из k заданных используется процедура с n участниками, цель которой — приписывание каждому объекту j некоторого показателя $q_j \geqslant 0,\ q_1 + \ldots + q_k = 1$, интерпретируемого как «мера того, что объект является лучшим». Считаем, что на каждом шаге t участник i, $1 \leqslant i \leqslant n$, приписывает объекту j, $1 \leqslant j \leqslant k$, показатель $q_j^{(i)}(t) \geqslant 0,\ q_1^{(i)}(t) + \ldots + q_k^{(i)}(t) = 1$. По этой информации организаторы процедуры находят набор $Q(t) = (q_1(t), \ldots, q_k(t))$ агрегированных (средних) показателей и сообщают его участникам. Каждый участник i некоторым образом преобразует (субъективизирует) его, образуя новый нормированный набор $f_i(Q(t)) = S^{(i)}(t) = (s_1^{(i)}(t), \ldots, s_k^{(i)}(t))$, и, задавшись некоторым $\theta^{(i)}(t)$, $0 \leqslant \theta^{(i)}(t) \leqslant 1$, сдвигается от прежнего набора показателей $Q^{(i)}(t)$ в сторону $S^{(i)}(t)$, полагая $Q^{(i)}(t+1) = (1 - \theta^{(i)}(t))Q^{(i)}(t) + \theta^{(i)}(t)S^{(i)}(t)$. Если процедура сходится, результатом считается агрегированный набор $Q = (q_1, \ldots, q_k)$, соответствующий точке сходимости, иначе она безрезультатна. Варьируя f_i, получаем разные модели.

Будем полагать, что каждый участник i разбивает объекты на ряд классов $T_1^{(i)}, \ldots, T_{k_i}^{(i)}$, считая объекты из одного класса равно предпочтительными, а из каждого предыдущего класса более предпочтительными, чем из следующих. Субъективизирующие функции f_i, $1 \leqslant i \leqslant m$, зависят от разбиений $\mathbf{T}^{(i)} = (T_1^{(i)}, \ldots, T_{k_i}^{(i)})$.

Модель типа 0 относится к случаю $\mathbf{T}^{(i)} = (T_1^{(i)}, T_2^{(i)})$. В ней

$$s_j^{(i)}(t) = \frac{\lambda_j^{(i)} q_j(t))}{\sum_u \lambda_u^{(i)} q_u(t)}, \qquad \lambda_j^{(i)} = \begin{cases} 1, & j \in T_1^{(i)}; \\ 0, & j \in T_2^{(i)}. \end{cases}$$

Модель типа 1 обобщает предыдущую на случай нескольких классов $T_1^{(i)},\dots,T_{k_i}^{(i)}$. Поведение участника i задается цепочкой чисел $\gamma_1^{(i)} \geqslant \dots \geqslant \gamma_{k_i}^{(i)} \geqslant 0$. Компоненты $s_j^{(i)}(t)$ набора $S^{(i)}(t)$ вычисляются в соответствии с предыдущей формулой при $\lambda_j^{(i)} = \gamma_m^{(i)}$, где $j \in T_m^{(i)}$.

Модель типа 2. По исходным классам $T_1^{(i)},\dots,T_{k_i}^{(i)}$ строится цепочка вложенных классов $\hat{T}_1^{(i)} \subset \dots \subset \hat{T}_{k_i}^{(i)}$, $\hat{T}_l^{(i)} = T_1^{(i)} \cup \dots \cup \hat{T}_l^{(i)}$. Классу $\hat{T}_l^{(i)}$ сопоставляется набор $S^{(i,l)}(t) = (s_1^{(i,l)}(t),\dots,s_k^{(i,l)}(t))$, образуемый согласно модели типа 0. Поведение участника i описывается набором коэффициентов $\alpha_l^{(i)} \geqslant 0$, $1 \leqslant l \leqslant k_i$, $\alpha_1^{(i)} + \dots + \alpha_{k_i}^{(i)} = 1$. В качестве $S^{(i)}(t)$ берется набор $\alpha_l^{(1)} S^{(1,l)}(t) + \dots + \alpha_l^{(n)} S^{(n,l)}(t)$.

Общая модель. Поведение участника i задается некоторым количеством m_i цепочек $\gamma_1^{(i,l)} \geqslant \dots \geqslant \gamma_{k_i}^{(i,l)} \geqslant 0 \quad (1 \leqslant l \leqslant m_i)$, где k_i — число классов разбиения в $\mathbf{T}^{(i)}$, и набором $(\alpha_1^{(i)},\dots,\alpha_{m_i}^{(i)})$, $\alpha_1^{(i)} + \dots + \alpha_{m_i}^{(i)} = 1$, положительных чисел. Вначале, в соответствии с моделью типа 1, при $\lambda_j^{(i,l)} = \gamma_m^{(i,l)}$, $j \in T_m^{(i)}$, находятся m_i наборов $S^{(i,l)}(t)$. Затем, подобно модели 2, вычисляется их линейная комбинация $S^{(i)}(t)$.

С общей моделью M свяжем функцию энтропийного типа от набора переменных $Q = (q_1,\dots,q_k)$, $q_1 \geqslant 0,\dots,q_k \geqslant 0$, $q_1 + \dots + q_k = 1$,

$$H_M(Q) = -\sum_{i,l} \alpha_l^{(i)} \ln\biggl(\sum_j \lambda_j^{(i,l)} q_j\biggr),$$

где $1 \leqslant i \leqslant n$, $1 \leqslant l \leqslant m_i$, $1 \leqslant j \leqslant k$.

Доказательство сходимости и устойчивости процедур основывается на следующем утверждении.

Теорема 1. *Если $\theta^{(1)}(t) = \dots = \theta^{(n)}(t) = \theta(t) > 0$, то справедливо неравенство $H_M(Q(t)) \geqslant H_M(Q(t+1))$, которое при $Q(t+1) \neq Q(t)$ является строгим.*

Обозначим через D_M множество точек минимума функции H_M. С помощью теоремы 1 доказывается

Теорема 2. *Если выполнено условие $\theta^{(1)}(t) = \dots = \theta^{(n)}(t) = \theta(t) > 0$, то последовательность $\{Q(t)\}$ агрегированных показателей сходится к множеству D_M.*

Можно показать, что в типичном случае множество D_M состоит из единственной точки, которая и является точкой сходимости процедуры.

Процедуру назовем *устойчивой*, если найдется точка Q такая, что процедура может сойтись лишь к точке Q, и существует вариант осуществления процедуры, при котором она сходится (к точке Q).

Следующее утверждение, в сочетании с теоремой 2, показывает, что в типичном случае (когда множество D_M одноэлементно) процедуры рассматриваемого вида при весьма общих предположениях устойчивы.

Теорема 3. *Если* $\theta(t) = \min\{\theta^{(1)}(t), \ldots, \theta^{(n)}(t)\}$, *ряд* $\sum_t \theta(t)$ *расходится, и набор агрегированных показателей, получаемых в результате процедуры, сходится к точке* Q, *то* $Q \in D_M$.

Будем говорить, что решение $Q = (q_1, \ldots, q_k)$ *противоречит мнению участника* i, если $q_j = 0$ для всех $j \in T_1^{(i)}$. Решение Q назовем *корректным*, если оно не противоречит мнению ни одного из участников. Множество моделей назовем *корректно полным*, если в нем могут быть получены все корректные решения, реализуемые моделями общего вида.

Будем считать, что множеством решений, связанных с моделью M, является D_M. Модель M назовем *максимально точной*, если $D_M \subseteq D_{M'}$ для любой модели M' такой, что $D_M \cap D_{M'} \neq \varnothing$.

Теорема 4. *Все модели типа 2 являются корректными и максимально точными, а множество моделей типа 2 корректно полно.*

Таким образом, модели типа 2 обладают в рассматриваемом классе моделей наилучшими свойствами.

Работа выполнена при поддержке РФФИ, проект № 06-01-00577 и ОИТВС РАН (программа «Фундаментальные основы информационных технологий и систем»).

Литература

[1] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. Вып. 33. — М.: Наука, 1978. — С. 5–68.

[2] *Шоломов Л. А.* Исследование одного класса динамических процедур коллективного выбора // Нелинейная динамика и управление. Вып. 5. — М.: Физматлит, 2006.

Статистический кластер-алгоритм

Шурыгин А. М.
a.shurygin@bk.ru
Москва, МГУ им. М. В. Ломоносова, факультет ВМиК

Большинство алгоритмов кластер-анализа требуют указания числа k классов, на которое надо поделить совокупность точек наблюдения $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$. Но в приложениях величину k можно указать лишь тогда, когда кластеризация уже проведена на интуитивном уровне. Во всех алгоритмах остаётся неопределённым само понятие кластера, а обычно к кластер-анализу обращаются, когда надо разделить сгустки точек «объективно», например, в биологии разделить виды, не путая их с подвидами.

Задача явно статистическая. Против предположения о нормальности распределений в кластерах никто не возражает. К. Пирсон [1] вывел критерий проверки гипотезы о совпадении центров двух многомерных нормальных совокупностей с известными ковариациями, различающихся сдвигом. Условия очень жёсткие и для кластер-анализа непригодные.

Предположив, что кластеры являются выборками из нормальных распределений, построить «объективный» кластер-алгоритм можно следующим образом. Кластер-критерием [2] на уровне значимости α проверить гипотезу о принадлежности выборки одному нормальному распределению. Если гипотеза принимается, то выборка с вероятностью $1-\alpha$ содержит один кластер. Критерий использует свойство оценок Мешалкина [3] оценивать параметры распределения кластера, наибольшего по количеству точек. Если гипотеза отвергается, то самый большой кластер вырезается эллипсоидом на уровне значимости α и к оставшимся точкам применяется описанная процедура в цикле. Процедура выделения кластеров заканчивается, когда все точки распределены по кластерам, либо их количество невелико, так что они могут считаться засоряющими или $\alpha-$остатками от проверки гипотез.

Рассмотрим элементы этого решения.

Оценки Мешалкина [3] $\mathbf{m}_\lambda = \left(m_\lambda^{(1)}, \ldots, m_\lambda^{(p)}\right)^{\mathrm{T}}$ и $\mathbf{C}_\lambda = \{c_\lambda^{ij}\}$ параметров нормального распределения $\mathcal{N}_p(\mathbf{m}, \mathbf{C})$ удовлетворяют системе уравнений, которую удобно решать итерациями:

$$\begin{cases} \mathbf{m}_\lambda = \dfrac{\Sigma_i\, \mathbf{x}_i \exp(-\lambda q_i/2)}{\Sigma_i\, \exp(-\lambda q_i/2)}; \\ \mathbf{C}_\lambda = (1+\lambda)\dfrac{\Sigma_i\, (\mathbf{x}_i - \mathbf{m}_\lambda)(\mathbf{x}_i - \mathbf{m}_\lambda)^{\mathrm{T}} \exp(-\lambda q_i/2)}{\Sigma_i\, \exp(-\lambda q_i/2)}; \end{cases}$$

где $q_i^2 = (\mathbf{x}_i - \mathbf{m}_\lambda)^{\mathrm{T}} C_\lambda^{-1} (\mathbf{x}_i - \mathbf{m}_\lambda)$ — квадрат расстояния от точки $\mathbf{x}_i$ до оценки $\mathbf{m}_\lambda$ центра распределения $\mathbf{m}$, измеренный оценкой $\mathbf{C}_\lambda$ матрицы ковариаций $\mathbf{C}$.

Кластер-критерий. Предлагается по величине

$$K_\lambda = \frac{1}{np} \sum\nolimits_i (\mathbf{x}_i - \mathbf{m}_\lambda)^{\mathrm{T}} \mathbf{C}_\lambda^{-1} (\mathbf{x}_i - \mathbf{m}_\lambda)$$

проверять «однородность» выборки при альтернативе наличия излишнего количества точек на периферии. Доказана асимптотическая сходимость к нормальному распределению величины

$$\sqrt{np} \ln K_\lambda \xrightarrow{d} \mathcal{N}(0,\ \xi^2);$$
$$\text{где } \xi^2 = \frac{\left[2 + 4\lambda + (p+2)\lambda^2\right](1+\lambda)^{p+2}}{(1+2\lambda)^{p/2+2}} - 2.$$

Если эта величина попадает в доверительный интервал на уровне значимости $1\% \leqslant \alpha \leqslant 5\%$, то выборка считается однородной и процесс заканчивается.

Если гипотеза отвергается, производится следующая процедура: **выделение точек кластера**. Они локализуются внутри p-мерного эллипсоида, $p \geqslant 2$, с центром в точке $\mathbf{m}$, с плотностью, постоянной на поверхности эллипсоида. Для проверки гипотезы на уровне значимости α эллипсоид должен содержать $1 - \alpha$ часть распределения. Пусть соответствующая часть эллипсоида удовлетворяет по вероятности равенству

$$\mathsf{P}\left\{\sqrt{(\mathbf{x} - \mathbf{m}_\lambda)^{\mathrm{T}} \mathbf{C}_\lambda^{-1} (\mathbf{x} - \mathbf{m}_\lambda)} \leqslant a_p\right\} = 1 - \alpha.$$

Положим $\mathbf{m}_\lambda = 0$, и соответственно центрируем плотность распределения. Сожмём эллипсоид вместе с плотностью распределения по главным осям так, чтобы распределение полученного вектора $\mathbf{y}$ стало стандартным нормальным $\mathcal{N}_p(\mathbf{0}, \mathbf{I})$. Найдём радиус b_p сферы $R_p(b)$, содержащей $1 - \alpha$ часть распределения. Решим уравнение

$$\mathsf{P}\left\{\sqrt{\mathbf{y}^{\mathrm{T}}\mathbf{y}} \in R_p(b_p)\right\} = 1 - \alpha$$

относительно b_p. Тогда получим равенство

$$1 - \alpha = \frac{1}{2^{p/2-1}} \int_0^{b_p} r^{p-1} \mathbf{e}^{-r^2/2}\, dr \Big/ \Gamma(p/2),$$

которое нетрудно решить численно, увеличивая b_p от $b_p = 2$. Пренебрегая небольшими отличиями истинных параметров распределения от их оценок, можно написать приближённое равенство

$$a_p \approx b_p,$$

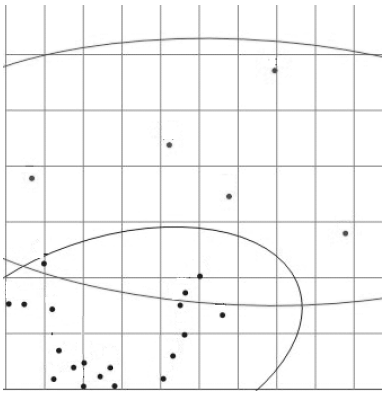


Рис. 1. Задача с пересечением двух кластеров.

решающее поставленную задачу.

О. Медведева составила программу для двумерного случая и решила ряд задач. Наиболее интересными из них были задачи со сложными наложениями и пересечениями нескольких кластеров.

Работа выполнена при поддержке РФФИ, проект № 04-01-00064.

Литература

[1] *Pearson K.* On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. // Phil. Mag. — 1900. — Vol. 50. — Pp. 157–175.

[2] *Шурыгин А. М.* Статистический кластер-критерий. // Алгоритмическое и программное обеспечение прикладного статистического анализа. — М.: Наука, 1980. — С. 360–366.

[3] *Meshalkin L. D.* Some mathematical methods for the study of non-communicable diseases. // Proc. 6-th Intern. Meeting of Uses of Epidemiol. in Planning Health Services. — Yugoslavia, Primosten. — Vol. 1. — Pp. 250–256.

Генетический алгоритм формирования оптимального подмножества диагностических тестов

Янковская А. Е., Цой Ю. Р.
yank@tsuab.ru, qai@mail.ru
Томск, Томский архитектурно-строительный университет,
Томский политехнический университет

Рассматривается проблема выбора оптимального подмножества безызбыточных безусловных диагностических тестов (ББДТ) с использованием эволюционного подхода. Для повышения качества результатов и обеспечения параллельного поиска предлагается использование нишинга (niching).

Введение

Выбор «хороших» безызбыточных диагностических тестов является одним из наиболее важных при принятии решений в интеллектуальных системах, поскольку от свойств используемых тестов существенно зависит качество получаемых решений. Однако этот выбор не всегда приводит к оптимальному решению, поскольку общее количество признаков в выбранном множестве тестов может быть слишком большим, также как временные и стоимостные затраты или ущерб (риск) [1], наносимый в результате выявления значений признаков исследуемого объекта, например, при решении геоэкологических или биомедицинских задач.

Впервые критерии оптимальности и безусловно актуальная задача поиска оптимального подмножества ББДТ поставлена в статье [2]. В статье [3] сформулированы критерии оптимальности, а в статье [4] предложено три алгоритма, обеспечивающие выполнение этих критериев: логико-комбинаторный, алгоритм на основе метода анализа иерархии и генетический алгоритм (ГА).

Постановка задачи

Кратко сформулируем рассматриваемую задачу. Пусть $\mathbf{T} = \{t_{ij}\}$, $i = 1, \ldots, n,\ \ j = 1, \ldots, m$ — матрица ББДТ, n — количество ББДТ, m — количество характеристических признаков, $\mathbf{T}_i$ соответствует i-му ББДТ (i-я строка матрицы $\mathbf{T}$). Обозначим через $\mathbf{z} = \{z_j : j = 1, \ldots, m\}$ множество характеристических признаков, причем $t_{ij} = 1 \leftrightarrow z_j \in \mathbf{T}_i$. Для каждого признака z_j зададим весовой коэффициент w_j и коэффициенты стоимости w'_j и ущерба (риска) w''_j [1].

Далее для краткости будем использовать термины «вес», «стоимость» и «ущерб» признака вместо соответственно «весовой коэффициент», «коэффициент стоимости» и «коэффициент ущерба».

Для данной матрицы $\mathbf{T}$ с заданными весами, стоимостью и ущербами признаков, необходимо выделить такую подматрицу $\mathbf{T}_0$, содержащую

n_0 строк, чтобы соответствующее ей множество тестов $\mathbf{N}^0$ обеспечивало выполнение следующих критериев, в порядке их следования [4]:

1. В выбранном множестве тестов $\mathbf{N}^0$ мощности n_0 должно содержаться максимальное число псевдообязательных признаков.
2. Множество $\mathbf{N}^0$ должно содержать минимальное общее число признаков.
3. Множество $\mathbf{N}^0$ должно иметь максимальный суммарный вес.
4. Множество $\mathbf{N}^0$ должно иметь наименьшую суммарную стоимость.
5. Множество $\mathbf{N}^0$ должно иметь наименьший суммарный ущерб.

Генетический алгоритм

Для решения поставленной задачи предлагается использовать ГА [5], представляющий итерационный вероятностный эвристический алгоритм поиска. Отличительной особенностью ГА является одновременная работа со множеством точек (популяцией) из пространства потенциальных решений.

Для решения рассматриваемой задачи каждое возможное решение представлено бинарной хромосомой (строкой) длины n, каждый i-й символ которой кодирует включение i-го диагностического теста в итоговое подмножество. В результате процесса искусственной эволюции, включающего отбор, рекомбинацию и вариацию (мутацию) «хороших», с точки зрения сформулированного критерия оценки [1], решений, качество решений в популяции постепенно улучшается. Окончанием генетического поиска, как правило, является нахождение субоптимального решения.

Для повышения качества результатов работы ГА предлагается использование нишинга [6, 7], для которого характерно сохранение в одной популяции как можно большего количества недоминируемых (несравнимых, nondominated) решений, соответствующих различным точкам на границе Парето, что дает возможность организовать параллельный поиск нескольких решений, в отличии от ГА без нишинга, в котором вся популяция сходится к одному решению.

Наиболее известны следующие разновидности нишинга:

1. Разделение приспособленности (fitness sharing) [6], при котором приспособленность хромосомы пересчитывается в зависимости от количества схожих с ней хромосом (в соответствии с выбранной метрикой) в данной популяции и их приспособленности.
2. «Перенаселение» (crowding) [7]. При использовании данной разновидности нишинга хромосомы-потомки замещают в популяции только наиболее близкие к ним хромосомы, что препятствует сходимости популяции к одному единственному решению.

Дальнейшие исследования будут посвящены исследованию влияния различных видов нишинга на качество результатов при решении поставленной задачи. Программный модуль, реализующий ГА будет включен в состав интеллектуального инструментального средства ИМСЛОГ [8].

Работа выполнена при поддержке РФФИ, проект № 07-01-00452, и РГНФ, проект № 06-06-12603В.

Литература

[1] *Yankovskaya A. E., Tsoy Y. R.* Optimization of a set of tests selection satisfying the criteria prescribed using compensatory genetic algorithm // Proc. of IEEE East-West Design & Test Workshop. Odessa, Ukraine, September 15–19, 2005. — Pp. 123–126.

[2] *Янковская А. Е.* Построение логических тестов с заданными свойствами и логико-комбинаторное распознавание на них // Интеллектуализация обработки информации (ИОИ-2002). — Симферополь, 2002. — С. 100–102.

[3] *Yankovskaya A. E., Mozheiko V. I.* Optimization of a set of tests selection satisfying the criteria prescribed // 7th Int. Conf. on Pattern Recognition and Image Analysis. — St. Petersburg: SPbETU, 2004. — Vol. I. — Pp. 145–148.

[4] *Колесникова С. И., Можейко В. И., Цой Ю. Р., Янковская А. Е.* Алгоритмы выбора оптимального множества безызбыточных диагностических тестов в интеллектуальных системах поддержки принятия решений // Межд. конф. «Системный анализ и информационные технологии» САИТ-2005, Переславль-Залесский. — М.: КомКнига, 2005. — Т. 1. — С. 256–262.

[5] *Емельянов В. В., Курейчик В. М., Курейчик В. В.* Теория и практика эволюционного моделирования. — М.: ФИЗМАТЛИТ, 2003. — 432 с.

[6] *Goldberg D. E., Richardson J.* Genetic Algorithms with sharing for multimodal optimization // Proc. of the 2nd Int. Conf. on Genetic Alg., 1987. — P. 41–49.

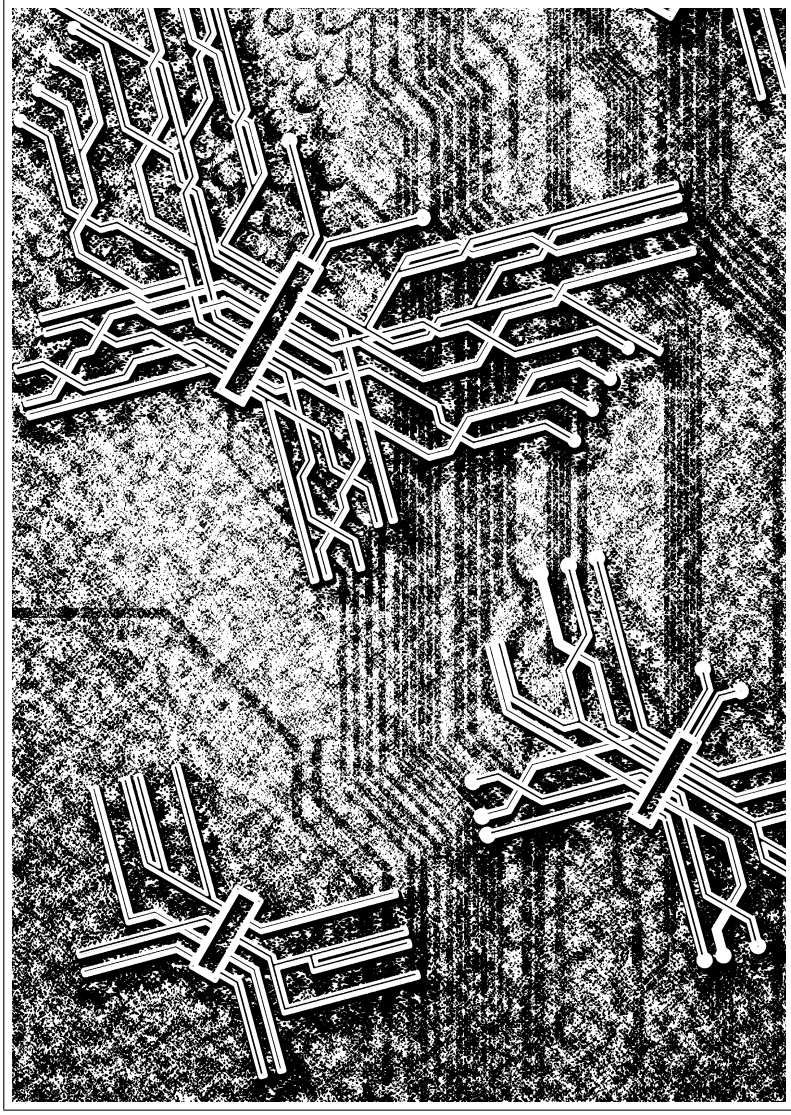
[7] *De Jong K.* An analysis of the behavior of a class of genetic adaptive systems. PhD thesis. — University of Michigan, Ann Arbor, 1975.

[8] *Yankovskaya A.E., Gedike A.I., Ametov R.V., Bleikher A.M.* IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition // Pattern Recognition and Image Analysis. — 2003. — Vol. 13, № 4. — P. 650–657.

Проблемы эффективности вычислений и оптимизации

Код раздела: CO (Computation and Optimization)

- Проблемы алгоритмической сложности.
- Построение вычислительно эффективных алгоритмов распознавания и прогнозирования.
- Численные методы оптимизации в задачах интеллектуального анализа данных.
- Параллельные вычисления.



Сведение задач криптоанализа асимметричных шифров к решению ассоциированных задач «ВЫПОЛНИМОСТЬ»

Дулькейт В. И., Файзуллин Р. Т., Хныкин И. Г.
r.t.faizullin@mail.ru, hig82@rambler.ru
Омск, Омский государственный университет им. Ф. М. Достоевского

Предложен способ решения задач криптоанализа асимметричных шифров (факторизации, дискретного логарифмирования) путем сведения к задаче ВЫПОЛНИМОСТЬ (SAT) и минимизации, специальным образом построенного, функционала.

Задача SAT заключается в поиске набора булевых значений, на котором заданная булева формула, представленная в КНФ, принимает значение ИСТИНА. Перспективным направлением в решении задачи SAT представляется её сведение к непрерывному аналогу, к задаче поиска точек глобального минимума ассоциированного с КНФ функционала [1]. В работе обосновывается выбор функционала специального вида. Предлагается применить к решению системы нелинейных алгебраических уравнений, определяющих стационарные точки функционала, модифицированный метод последовательных приближений.

Показано, что метод, в отличие от большинства существующих методов, позволяет достаточно эффективно находить решение для КНФ, эквивалентных указанным задачам криптоанализа. Рассматривается применимость метода к другим реальным задачам.

Сведение задач криптоанализа к задачам SAT осуществляется путем кодирования элементарных составляющих операции умножения и дискретного логарифмирования в терминах булевой алгебры. Метод кодирования представлен в работе [1].

Построение модифицированного метода последовательных приближений

Пусть $K(x) = \bigcap_{i=1}^{M} C_i(x)$ — КНФ. Переход от задачи SAT к задаче поиска глобального минимума функционала осуществляется по формуле

$$\min_{x \in E^n} F(x) = \min_{x \in E^n} \sum_{i=1}^{M} \prod_{j=1}^{N} Q_{i,j}(x) = 0; \tag{1}$$

$$Q_{i,j}(x) = \begin{cases} (1 - x_j)^2, & \text{если } x_j \in C_i(x); \\ x_j^2, & \text{если } \bar{x}_j \in C_i(x); \\ 1, & \text{иначе.} \end{cases}$$

Легко заметить, что $\min\limits_{x\in E^n} F(x) = 0$ соответствует достижению значения ИСТИНА на исходной КНФ.

Дифференцируя функционал по всем x_i, получим систему уравнений:

$$\sum_{\xi\in\Xi}\prod_{j\neq i}^{N} Q_{i,j}(x)\cdot x_i = \sum_{\xi\in\Lambda}\prod_{j\neq i}^{N} Q_{i,j}(x) \text{ где } i=1,\ldots,N; \tag{2}$$

$$\Xi = \{\xi, k\in\xi : x_i \text{ или } \bar{x}_i \in C_k(x)\},\quad \Lambda = \{\xi, k\in\xi : x_i \in C_k(x)\}.$$

Для ее решения предлагается применить метод последовательных приближений с «инерцией»:

$$\left[\sum_{p=0}^{K}\alpha_p\sum_{\xi\in\Xi}\rho_\xi\prod_{j\neq i}^{N} Q_{i,j}(x(t-p))\right]\cdot x_i(t+1) = \sum_{\xi\in\Lambda}\prod_{j\neq i}^{N} Q_{i,j}(x(t-p)) \tag{3}$$

$$\sim A^i\cdot x_i(t+1) = B^i, \text{ где } \sum_{p=0}^{K}\alpha_p = 1,\quad \alpha_p\geqslant 0,\quad \rho_\xi\geqslant 0.$$

Положив в (3) $K=0$, $\rho_\xi = 1$, получим простой метод последовательных приближений. В отличие от него, модифицированный метод формирует приближения не так быстро, что позволяет избегать областей притяжения аттракторов.

Преобразование исходной КНФ методом резолюции

Преобразование позволяет получить КНФ с меньшим количеством дизъюнктов и литералов, эквивалентную исходной.

Резольвента — дизъюнкция конъюнктов, отличающихся знаком по единственной переменной. Все возможные резольвенты добавляются к КНФ и используются для вычисления других резольвент. Дублирующие конъюнкты и тавтологии удаляются. Здесь используется сокращенная процедура с глубиной рекурсии 1. Вычислительная сложность процедуры $O(n\log n)$.

Метод резолюции в применении к КНФ, ассоциированных с задачами факторизации и дискретного логарифмирования позволяет уменьшить исходное число конъюнктов до 50% и разрешить до 20% переменных.

Гибридизация и распараллеливание алгоритма.

Гибридизация алгоритма состоит в добавлении дополнительных методов, позволяющих ускорить сходимость первоначального метода.

Основная процедура состоит из последовательных итераций, которые совмещают метод последовательных приближений (3) и сдвиг по градиенту: $x_i(t+1) = 2x_i(t) - B^i/A^i$, т.к. правая часть (2) суть градиент исходного функционала. Используется схема Зейделя.

1	Размерность,бит	40	44	52	56	60
	Число литералов	990	1199	1677	1946	2235
	Число дизъюнктов	22333	27291	38695	45141	52079
	Время решения, м.	7	36	360	36	612
2	Размерность,бит	18	20	22	24	26
	Число литералов	28224	38840	51832	67440	85904
	Число конъюнктов	448018	623239	839032	1099630	1409250
	Время решения, с.	63.57	108.20	182.73	277.46	417.71

Таблица 1. Результаты численных экспериментов: 1 — для задачи факторизации; 2 — для задачи дискретного логарифмирования.

Если скорость сходимости падает, применяется т. н. метод смены траектории. Текущее приближение проектируется на $B^n\{0,1\}$, и с некоторой вероятностью значения компонент вектора из множества $E = \{x_k \mid \exists$ конъюнкт $C_i\colon x_k$ или $\bar{x}_k \in C_i$ и $C_i(x) = 0\}$ меняются на противоположные. Работа алгоритма возобновляется с использованием нового полученного приближения.

Подробно о методах распараллеливания и ускорении сходимости: [1].

Результаты численных экспериментов.

При тестировании использовались КНФ библиотеки SATLib (`satlib.org`) и КНФ, сформированные для задач факторизации и дискретного логарифмирования. Результаты для тестов SATLib сравнимы с результатами ведущих алгоритмов [1]. Тесты, сформированные для задач факторизации и дискретного логарифмирования, оказываются наиболее трудными. Ведущие алгоритмы (RANOV, SATz) не смогли за обозримое время найти решение уже для задачи факторизации размерности 40 бит.

Предложенный алгоритм показал приемлемый (предположительно субэкспоненциальный) рост времени решения и возможное наличие «слабых» примеров больших размерностей (табл. 1).

При модификациях простого метода последовательных приближений было достигнуто равномерное улучшение сходимости по всем тестам.

Был разработан способ генерации трудных для решения КНФ, основанных на задачах криптоанализа.

Учитывая, что многие реальные задачи могут быть представлены в булевой форме, представляется перспективным применение параллельной версии метода для решения практически любых реальных задач.

Литература

[1] *Дулькейт В. И., Файзуллин Р. Т., Хныкин И. Г.* Алгоритм минимизации функционала, ассоциированного с задачей 3-SAT и его практические применения // ПаВТ, Челябинск, 2006.

О построении тупиковых покрытий булевых и целочисленных матриц

Дюкова Е. В., Инякин А. С.
djukova@ccas.ru, inyakin@ccas.ru
Москва, Вычислительный центр РАН

Задача построения тупиковых покрытий целочисленной, в частности, булевой, матрицы возникает при конструировании логических процедур распознавания и классификации и относится к числу трудных в вычислительном плане задач [1, 2, 3, 4]. Данная задача может быть также сформулирована как задача построения максимальных конъюнкций двузначной логической функции, заданной конъюнктивной нормальной формой (КНФ). Поиски эффективных алгоритмов ее решения ведутся с середины 1950-х годов.

Первостепенное значение имеет задача построения неприводимых покрытий (или тупиковых $(0,\dots,0)$-покрытий) булевой матрицы (или задача построения максимальных конъюнкций монотонной булевой функции, заданной КНФ).

Пусть $L = (a_{ij})$, $i = 1,\dots,m$, $j = 1,\dots,n$, — булева матрица. Будем предполагать, что L не содержит строк вида $(0,\dots,0)$. *Неприводимым* покрытием матрицы L называется набор H из r столбцов этой матрицы такой, что подматрица матрицы L, образованная столбцами набора H, не содержит строку $(0,\dots,0)$ и содержит каждую из строк $Q_1 = (1,0,0,\dots,0,0)$, $Q_2 = (0,1,0,\dots,0,0)$, $\dots$, $Q_r = (0,0,0,\dots,0,1)$. Подматрица матрицы L, составленная из строк $Q_1,\dots,Q_r$, называется единичной подматрицей.

Пусть $P(L)$ — множество всех неприводимых покрытий матрицы L.

Как правило, число неприводимых покрытий растет экспоненциально с ростом размера матрицы, поэтому эффективность алгоритмов построения неприводимых покрытий имеет смысл оценивать в терминах полиномиальной задержки.

Будем говорить, что алгоритм, порождающий множество $P(L)$, имеет задержку (delay) d, если выполнены следующие условия: 1) алгоритм выдаст первое неприводимое покрытие, выполнив не более d элементарных операций; 2) после выдачи очередного неприводимого покрытия он либо выполнит не более d элементарных операций прежде, чем выдаст следующее неприводимое покрытие, либо остановится. Под элементарной операцией понимается просмотр одного элемента матрицы L.

Алгоритм с задержкой, ограниченной сверху полиномом от размера матрицы, назовем алгоритмом с полиномиальной задержкой.

В [1, 2, 3] рассмотрен случай, когда число строк m матрицы имеет более низкий порядок роста, чем число столбцов n, при условии, что

$n \to \infty$. Для этого случая построен асимптотически оптимальный алгоритм поиска неприводимых покрытий (алгоритм АО1). Данный алгоритм строит с задержкой, не превосходящей $O(mn)$, приближенное решение, в качестве которого рассматривается совокупность наборов столбцов, содержащих единичные подматрицы. Каждый такой набор алгоритм АО1 строит столько раз, сколько единичных подматриц он содержит. Показано, что если $m^\alpha \leqslant n \leqslant 2^{m^\beta}$, $\alpha > 1$, $\beta < 1$, то при $n \to \infty$ число шагов алгоритма, равное числу единичных подматриц, почти всегда (для почти всех матриц размера $m \times n$) асимптотически равно мощности $P(L)$. При конструировании алгоритма АО1 в качестве приближенного решения может быть рассмотрена совокупность наборов столбцов, содержащих «максимальные» единичные подматрицы. Каждый такой набор столбцов строится столько раз, сколько максимальных единичных подматриц он содержит. Тогда алгоритм будет делать меньшее число шагов и работать с задержкой, не превосходящей $O(qmn)$, здесь и далее $q = \min(m, n)$. Данная модификация алгоритма АО1 обычно используется при его реализации на ЭВМ.

В [4] построен алгоритм, основанный на переборе с задержкой, не превосходящей $O(qm^2n)$ неприводимых покрытий матрицы L (алгоритм АО2). Однако данный алгоритм строит каждый набор из $P(L)$ столько раз, сколько единичных подматриц он содержит. Из сказанного выше следует, что указанный недостаток не существенен при $m^\alpha \leqslant n \leqslant 2^{m^\beta}$, $\alpha > 1$, $\beta < 1$ (в этом случае число шагов алгоритма, равное числу единичных подматриц, порождающих неприводимые покрытия, почти всегда при $n \to \infty$ асимптотически равно мощности $P(L)$).

Отметим, что проверка на повторяемость построенного на очередном шаге набора столбцов в алгоритмах АО1 и АО2 требует просмотра не более qm элементов матрицы L.

В докладе предложен «точный» алгоритм поиска неприводимых покрытий булевой матрицы с задержкой, не превосходящей $O\left(qmn^2(m+q)\right)$. Проведено экспериментальное обоснование алгоритма на случайных матрицах. Аналогичный результат получен для задачи поиска тупиковых покрытий целочисленной матрицы. При этом использовалось сведение задачи поиска тупиковых покрытий целочисленной матрицы к задаче поиска неприводимых покрытий булевой матрицы, приведенное в [5].

Работа выполнена при поддержке проектов РФФИ № 07-01-00516, № 05-01-00495 и гранта Президента РФ по поддержке ведущих научных школ НШ № 5833.2006.1 «Алгебраические и логические методы в задачах распознавания и прогнозирования».

Литература

[1] *Дюкова Е. В.* Об асимптотически оптимальном алгоритме построения тупиковых тестов // ДАН СССР. 1977. Т. 233. № 4. С. 527–530.

[2] *Дюкова Е. В.* Алгоритмы распознавания типа Кора: сложность реализации и метрические свойства // Распознавание, классификация, прогноз (математические методы и их применение). — М.: Наука, 1989. Вып. 2. С. 99–125.

[3] *Дюкова Е. В., Журавлёв Ю. И.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // ЖВМиМФ. — 2000. — Т. 40, № 8. — С. 1264–1278.

[4] *Дюкова Е. В.* О сложности реализации дискретных (логических) процедур распознавания // ЖВМиМФ. — 2004. — Т. 44, № 3. — С. 550–572.

[5] *Дюкова Е. В., Инякин А. С.* О процедурах классификации, основанных на построении покрытий классов // ЖВМиМФ. — 2003. — Т. 43, № 12. — С. 1910–1921.

Поиск минимальных покрытий булевой матрицы с использованием параллельных вычислений

Дюкова Е. В., Инякин А. С., Нефёдов В. Ю.

djukova@ccas.ru

Москва, Вычислительный центр РАН

В ряде случаев при построении процедур распознавания и кластеризации используется аппарат дискретной математики, в частности методы поиска покрытий булевой или целочисленной матрицы (например, при конструировании алгоритмов вычисления оценок, при синтезе элементарных классификаторов в логических процедурах распознавания, при построении логического корректора на базе элементарных классификаторов). Среди разного вида задач поиска покрытий одной из центральных является задача поиска минимальных покрытий булевой матрицы.

Пусть L — булева матрица. Набор столбцов H матрицы L называется *покрытием*, если каждая строка матрицы L в пересечении хотя бы с одним из столбцов, входящих в H, даёт единицу. Покрытие называется *неприводимым*, если никакое его собственное подмножество покрытием не является. Покрытие минимальной длины называется *минимальным* покрытием. Требуется найти все минимальные покрытия матрицы L.

Данная задача является трудной в вычислительном плане [1], поэтому важна разработка эффективных (в практическом плане) алгоритмов её решения.

Поскольку все минимальные покрытия являются неприводимыми, то для их поиска можно адаптировать алгоритмы нахождения неприводимых покрытий. В настоящее время разработан целый ряд таких алго-

ритмов [2, 3, 4, 5], среди которых следует выделить асимптотически оптимальные алгоритмы. Асимптотически оптимальный алгоритм работает с полиномиальной задержкой на каждом шаге, и число его шагов асимптотически равно числу всех неприводимых покрытий для почти всех булевых матриц размера $m \times n$ при условии, что $n \to \infty$ и $m^\alpha \leqslant n \leqslant 2^{m^\beta}$, $\alpha > 1$, $\beta > 1$.

В работе рассмотрен асимптотически оптимальный алгоритм поиска неприводимых покрытий из [4] (алгоритм A). Данный алгоритм основан на нахождении на каждом шаге набора столбцов H, содержащего максимальную единичную (перестановочную) подматрицу. Если при этом набор столбцов H не содержит нулевой строки, то он является неприводимым покрытием. Алгоритм работает с полиномиальной задержкой $O\left(q^2mn\right)$, где $q = \min(m, n)$.

Реализованы и экспериментально исследованы три модификации алгоритма A для поиска минимальных покрытий булевой матрицы. Первые два алгоритма, а именно, алгоритм A_1 и алгоритм A_2, осуществляют односторонний обход дерева решений. Третий алгоритм (алгоритм A_3) осуществляет обход дерева решений широким фронтом. Алгоритм A_1 использует в качестве первоначальной верхней оценки длины минимального покрытия оценку, выдаваемую жадным алгоритмом (алгоритмом наискорейшего спуска), а алгоритм A_2 — оценку, выдаваемую «улучшенным» жадным алгоритмом.

В отличие от классического жадного алгоритма, «улучшенный» жадный алгоритм взвешивает единичные элементы.

Для оценки эффективности алгоритмов A_1, A_2 и A_3 проведён ряд экспериментов на однопроцессорной ЭВМ, которые показали, что «улучшенный» жадный алгоритм позволяет ускорить решение задачи в среднем на 5–10% и сократить число шагов алгоритма. Под шагом алгоритма понимается построение очередной висячей вершины в дереве решений. Эксперименты также показали, что при больших размерах матриц обход широким фронтом невыгоден.

Алгоритм A_2, показавший наилучшие результаты, адаптирован для использования на многопроцессорных комплексах. При этом реализованы две схемы распараллеливания. В обеих схемах присутствует выделенный управляющий процесс, который раздаёт задания считающим процессам. В первой схеме в случае нахождения покрытия меньшей длины, чем текущая верхняя оценка длины минимального покрытия, считающий процесс сообщает об этом управляющему, который рассылает информацию всем остальным. Во второй схеме считающий процесс сам рассылает новую оценку всем остальным считающим процессам.

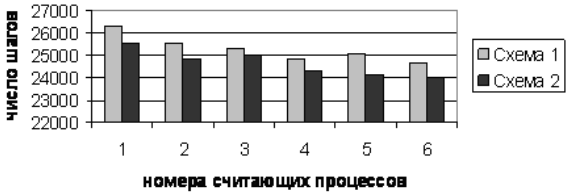


Рис. 1. Загруженность процессов.

Рис. 2. Масштабируемость.

Важными показателями качества распараллеливания являются равномерность загрузки считающих процессов (Рис. 1) и масштабируемоть, т. е. зависимость времени счёта от числа считающих процессов. Счёт на случайных матрицах показал, что при использовании как первой, так и второй схемы все считающие процессы загружены достаточно равномерно. Обе реализованные схемы показали типичную для задач дискретной оптимизации масштабируемость (Рис. 2). Вторая схема имеет небольшое преимущество по времени счёта и числу шагов.

Работа выполнена при поддержке РФФИ, проекты № 07-01-00516 и № 05-01-00495.

Литература

[1] *Гэри М., Джонсон Д.* Вычислительные машины и труднорешаемые задачи. — М.: Мир, 1982. — 416 с.

[2] *Дюкова Е. В.* О сложности реализации дискретных (логических) процедур распознавания // ЖВМиМФ. — 2004. — Т. 44, № 3. — С. 550–572.

[3] *Дюкова Е. В.* Об асимптотически оптимальном алгоритме построения тупиковых тестов // ДАН СССР. — 1977. — Т. 233, № 4. — С. 527–530.

[4] *Дюкова Е. В., Инякин А. С.* О сложности решения задачи построения поиска неприводимых покрытий булевой матрицы. — М.: ВЦ РАН, 2006. — 24 с.

[5] *Инякин А. С.* Алгоритмы поиска неприводимых покрытий булевой матрицы. — М.: ВЦ РАН, 2004. — 25 с.

Параллельная реализация алгоритма выделения оптимальной совместной подсистемы системы линейных неравенств

Катериночкина Н. Н., Шмаков А. С.

nnkater@@ccas.ru, ashmak@mail.ru

В процессе оптимизации некоторых моделей алгоритмов распознавания требуется как можно точнее удовлетворить определенной системе условий, которая описывается достаточно большим числом линейных неравенств и в целом может быть противоречивой. В общем случае решение такой проблемы сводится к выделению максимальной по мощности совместной подсистемы из заданной системы линейных неравенств.

Постановка задачи

Пусть задана система линейных неравенств:

$$\sum_{j=1}^{n} a_{ij}x_j \leqslant b_i, \quad i = 1, \ldots, m, \tag{1}$$

разбитая на блоки одинаковой длины l. Система (1) несовместна. Требуется найти совместную подсистему, содержащую максимальное число полных блоков заданной длины l. В частности, при $l = 1$ осуществляется поиск максимальной совместной подсистемы системы (1).

Разработан алгоритм, который осуществляет поиск совместной подсистемы, содержащей максимальное число полных блоков. В работах [1] и [2] доказывается, что для выделения всех нерасширяемых совместных подсистем системы (1) достаточно перебрать все ее подсистемы мощности r и ранга r. Именно на этом утверждении основана работа алгоритма, состоящая в переборе всех подсистем мощности r и ранга r системы (1). Для каждой из таких подсистем требуется решить несложную задачу: найти одно узловое решение. Для этого надо заменить все знаки неравенств в подсистеме на равенства и найти одно решение полученной системы линейных уравнений. Найденное узловое решение подсистемы подставляется во все неравенства системы (1) и выделяется из нее те

неравенства, которым это решение удовлетворяет. Выделенные неравенства образуют совместную подсистему. Перебрав таким образом все подсистемы, получаем некоторое множество совместных подсистем, среди которых будут все нерасширяемые совместные подсистемы. Среди них можно выбирать оптимальные подсистемы с интересующими свойствами.

Однако, трудоемкость поставленной задачи равна перебору $C_m^r = \frac{m!}{(m-r)!r!}$ r-подсистем. При больших значениях m эта задача требует больших вычислительных ресурсов и времени. Для ее решения предлагается построить параллельный алгоритм.

Параллельный алгоритм

Данную задачу удается достаточно легко и эффективно распараллелить на уровне данных, поскольку её решение состоит в переборе всех возможных r-подсистем системы (1). При этом задача рассмотрения очередной подсистемы является по сути самостоятельной задачей, включающей в себя следующие этапы:

1) поиск ранга подсистемы;
2) сравнение полученного ранга с рангом всей системы;
3) если ранг меньше, то происходит переход к следующей подсистеме;
4) если ранг равен, то подсистема является r-подсистемой, и далее ищется узловое решение и соответствующая ему подсистема неравенств.

Предлагается разбить множество всех подсистем на непересекающиеся подмножества и параллельно решать поставленную задачу для каждого из подмножеств.

Для эффективного распараллеливания необходимо добиться максимальной загрузки всех доступных процессоров. Для этого разделим множество всех возможных r-подсистем следующим образом. Введем булевский вектор $\gamma = (\gamma_1, \ldots, \gamma_m) \in E^m$, который будет описывать конкретную r-подсистему системы (1), где $\gamma_i = 1$, если неравенство с номером i входит в подсистему, и $\gamma_i = 0$ иначе. Будем перебирать все возможные векторы γ, такие что $|\gamma| = r$. Пусть $\Lambda = \{\gamma \in E^m \colon |\gamma| = r\}$, тогда $|\Lambda| = C_m^r$. Используя известные формулу для биномиальных коэффициентов:

$$C_m^r = C_{m-k}^r + C_{m-k}^{r-1} + \cdots + C_{m-k}^{r-k}, \tag{2}$$

будем делить множество всех векторов Λ на подмножества. Так, из (2) при $k = 1$ следует $\Lambda = \Lambda_1 \cup \Lambda_2$, где

$$\Lambda_1 = \Big\{\gamma \in \Lambda \;\Big|\; \gamma_1 = 1,\ \textstyle\sum_{i=2}^m \gamma_i = r - 1\Big\};$$
$$\Lambda_2 = \Big\{\gamma \in \Lambda \;\Big|\; \gamma_1 = 0,\ \textstyle\sum_{i=2}^m \gamma_i = r\Big\}.$$

Кол-во проц.	Время (1 проц.)	Время (p проц.)	Ускорение	Эффективность
2	5646, 12	3029, 15	1, 86	0, 932
4	5646, 12	1718, 26	3, 29	0, 821
8	5646, 12	930, 88	6, 07	0, 758
9	5646, 12	794, 40	7, 11	0, 790
10	5646, 12	739, 18	7, 64	0, 764
16	5646, 12	504, 07	11, 20	0, 700
20	5646, 12	405, 92	13, 91	0, 695
25	5646, 12	496, 85	11, 36	0, 455
30	5646, 12	305, 04	18, 51	0, 617
31	5646, 12	275, 56	20, 49	0, 661
32	5646, 12	284, 50	19, 85	0, 620

Таблица 1. Показатели эффективности распараллеливания.

Аналогично образом можно разделить множество Λ на $l = 2^q$ подмножеств. В результате получится упорядоченный набор булевских векторов, который представляет собой булевский куб размерности q.

Пусть у нас имеется p доступных процессоров. Выделяем один из них для основного или родительского процесса. Остальные $p - 1$ будут использоваться для вычислений. Назовем их *пулом свободных процессоров*. Основной процесс начинает работу с того, что определяет $q = \left[log_2(p-1)\right] + 1$ и делит множество Λ на 2^q подмножеств. Не ограничивая общности, будем считать, что задача задана корректно, т. е. выполнены следующие ограничения: $r \geqslant q - 1$, $m \gg r$, иначе задача имеет достаточное простое решение и не требует больших вычислительных мощностей.

Итак, главный процесс формирует нетривиальные задания (т. е. непустые подмножества) на решения задачи поиска максимальной совместной подсистемы и узлового решения для задачи меньшей размерности, затем отдает эти задания процессорам, причем они переходят в *пул занятых процессоров*. В начальный момент все процессы получают, задание, поскольку задача разбивается на большое число подзадач, чем количество процессов. Каждый из них решает самостоятельную задачу меньшей размерности либо по количеству неравенств, либо по рангу системы, либо и по тому, и по другому. Завершив свое выполнение, каждый из процессов возвращает результаты в основной процесс и переходит в пул свободных процессов. Если остались еще задания, то главный процесс выдает последовательно произвольным процессам, попадающим в пул свободных процессов, новое задание на обработку. Если заданий больше нет, главный процесс посылает сигнал на уничтожение процессу.

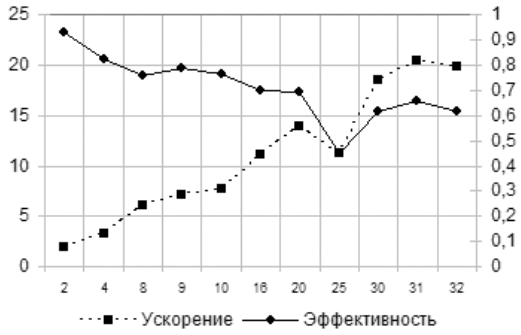


Рис. 1. Зависимость ускорения и эффективности работы алгоритма от количества процессоров.

Главный процесс собирает результаты от всех процессов по каждому из заданий и последовательно выбирает максимальное размер совместной подсистемы, а также номера соответствующих неравенств и узловое решение. После обработки всех заданий главный процесс выдает результат задачи.

Результаты

Данная задача достаточно просто и эффективно распараллеливается на уровне данных, т. е. для решения всей задача надо перебрать некоторое множество, решив для каждой выборки некоторую самостоятельную задачу, тогда как при параллельном подходе мы можем разделить исходное множество на подмножества и параллельно решать ту же задачу для каждого из них. В идеальном случае при делении множества на абсолютно одинаковые подмножества (с точки зрения их сложности, как вычислительных задач, а не с точки зрения их мощности, как множеств), можно получить идеальной случай распараллеливания. Однако, заранее невозможно определить вычислительную сложность каждого из подмножеств. В предложенном подходе исходное множество заранее делится на количество подмножеств, превышающих количество свободных процессоров. В начальный момент каждый из процессоров получает задачу на обработку подмножества. Закончив свою работу, он либо завершается, либо получает новое задание. Процессоры, получившие «большие» задания, будут работать только над ними; процессоры, получившие «небольшие» задания, смогут обработать несколько из них.

Для оценки эффективности распараллеливания будем применять следующие показатели (см. Таблицу 1, Рис. 1):

— Ускорение $A_p = \frac{T_1}{T_p}$, где T_p — время исполнения распараллеленной программы на p процессорах, T_1 — время исполнения исходной программы.
— Эффективность $S_p = \frac{T_1}{pT_p}$, показывающая долю использования процессоров.

Вычислительные эксперименты проводились на многопроцессорном вычислительном комплексе MVS-15000BM, установленном в Межведомственном суперкомпьютерном центре РАН (МСЦ РАН).

Работа выполнена при поддержке проектов РФФИ № 05-07-90333, Целевой программы № 14 Президиума РАН, Целевой программы № 2 Отделения математических наук РАН.

Литература

[1] *Катериночкина Н. Н.* Методы выделения максимальной совместной подсистемы системы линейных неравенств. Сообщение по прикладной математике. — Москва: Вычислительный центр РАН, 1997.

[2] *Черников С. Н.* Линейные неравенства. — Москва: Наука, 1968.

[3] *Воеводин В. В., Воеводин Вл. В.* Параллельные вычисления.— СПб.: БХВ-Петербург, 2002.

О некоторых полиномиально разрешимых и NP-трудных задачах анализа и распознавания последовательностей с квазипериодической структурой

Кельманов А. В.

kelm@math.nsc.ru

Новосибирск, Институт математики им. С. Л. Соболева СО РАН

Рассматривается нетрадиционный подход к помехоустойчивому компьютерному анализу и распознаванию числовых последовательностей, сущность которого состоит в апостериорном (off-line) способе обработки последовательности в сочетании с формализацией содержательной задачи как задачи принятия решения (проверки гипотез). Изложены результаты по исследованию сложности, решению и систематизации дискретных экстремальных задач, к которым сводится реализация этого подхода в случае, когда последовательности включают квазипериодически чередующиеся информационные фрагменты (подпоследовательности), имеющие одну и ту же размерность (число членов).

Задачи, входящие в анализируемую совокупность, возникают в приложениях, связанных с обработкой массивов зашумленных структурированных данных — результатов измерения характеристик изучаемых объектов различной природы. Эти задачи типичны, в частности, для

электронной разведки, дистанционного зондирования, телекоммуникации, геофизики, обработки речевых сигналов, биометрики, медицинской и технической диагностики, радиолокации, гидроакустики, криминалистики, поиска по мультимедийным базам данных и др.

В основе трех хорошо изученных традиционных подходов лежат последовательный (on-line) и апостериорный способы обработки последовательности в сочетании с формализацией содержательной задачи как задачи оценивания (оптимальной фильтрации), а также последовательный способ обработки в комбинации с формализацией задачи как задачи проверки гипотез. Эти подходы имеют глубокую историю и связаны с фундаментальными работами Колмогорова, Котельникова, Пугачева, Ширяева, Харкевича, Андерсона, Вальда, Винера, Калмана, Пэйджа, Хинкли и множества других отечественных и зарубежных исследователей. При реализации традиционных подходов проблемы комбинаторной оптимизации как правило не возникают. В противоположность этому, реализация рассматриваемого в работе подхода сопряжена с решением специфических задач комбинаторной оптимизации с целью выбора наилучшего из множества допустимых решений, мощность которого растет экспоненциально при увеличении длины последовательности.

Числовая последовательность, включающая квазипериодически чередующиеся ненулевые информационные фрагменты размерности q, определяется следующей формулой общего члена:

$$x_n = \sum_{m \in \mathbb{M}} u_{n-n_m}(m), \quad n = 0, \ldots, N-1,$$

где $u_{n-n_m}(m) = 0$, если $n - n_m \neq 0, \ldots, q-1$; $\big(u_0(m), \ldots, u_{q-1}(m)\big) \in \mathbb{R}^q$, $0 < \big\|(u_0(m), \ldots, u_{q-1}(m))\big\| < \infty$ при каждом $m \in \mathbb{M} = \{1, \ldots, M\}$, а $(n_1, \ldots, n_M) \in \Omega$, причем

$$\Omega = \bigcup_{M=M_{\min}}^{M_{\max}} \Omega_M,$$

где

$$\Omega_M = \left\{ (n_1, \ldots, n_M) \;\middle|\; \begin{array}{l} 0 \leqslant n_1 \leqslant N^+ \leqslant N - q, \;\; 0 \leqslant N^- \leqslant n_M \leqslant N - q \\ 0 < q \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N - q, \\ m = 2, \ldots, M \end{array} \right\},$$

а $M_{\min}$ и $M_{\max}$ находятся из решения системы неравенств, входящих в определение множества Ω_M, в которой N^-, N^+, $T_{\min}$ и $T_{\max}$ — целые числа. Термин «квазипериодическая последовательность» (т. е. последовательность, квазипериодически изменяющая свои свойства) обусловлен спецификой ограничений, входящих в определение множества Ω_M.

Положим $U_m = \big(u_0(m), \ldots, u_{q-1}(m)\big)$ и назовем $U_m,\ m \in \mathbb{M}$, *информационным вектором*, последовательность его компонент — *информационной последовательностью*. Фрагмент $(x_{n_m}, \ldots, x_{n_m+q-1}),\ m \in \mathbb{M}$, последовательности $x_n,\ n = 0, \ldots, N-1$, совпадающий с вектором U_m, будем называть *информационным фрагментом*.

Предполагается, что вектор $X = (x_0, \ldots, x_{N-1})$, последовательность компонент которого содержит чередующиеся информационные фрагменты, недоступен для непосредственой обработки из-за вектора помехи $E = (\varepsilon_0, \ldots, \varepsilon_{N-1}) \in \Phi_{0,\sigma^2 I}$, где $\Phi_{0,\sigma^2 I}$ — нормальное распределение. Доступным для обработки считается вектор $Y = X + E$. При этом вектор X рассматривается как функция $X(n_1, \ldots, n_M, U_1, \ldots, U_M)$, совокупность аргументов которой уточняется при формулировке различных вариантов задач анализа и распознавания.

Все рассмотренные в докладе дискретные экстремальные задачи выявлены (возникают) в результате формализации содержательных задач обработки данных как задач принятия решения (о среднем $X(\cdot)$ случайного гауссовского вектора $Y \in \Phi_{X(\cdot),\sigma^2 I}$), доставляющего максимум функционалу правдоподобия. К идентичным формулировкам экстремальных задач приводит минимизация функционала $\|Y - X(\cdot)\|^2$ суммы квадратов уклонений.

Совокупность выявленных экстремальных задач включает следующие классы, объединяющие содержательно похожие задачи:

1) обнаружение в числовой последовательности повторяющегося фрагмента;
2) распознавние последовательности, включающей повторяющийся фрагмент;
3) обнаружение и идентификация фрагментов;
4) распознавние последовательности, включающей фрагменты из алфавита;
5) обнаружение фрагментов в последовательности и разбиение этой последовательности на серии идентичных фрагментов;
6) распознавние последовательности, включающей серии идентичных фрагментов;
7) обнаружение в числовой последовательности повторяющегося набора фрагментов;
8) распознавание последовательности, включающей повторяющийся набор фрагментов;
9) обнаружение и идентификация наборов фрагментов;
10) кластеризация последовательностей.

Для части экстремальных задач из анализируемой совокупности установлена полиномиальная разрешимость либо NP-трудность, обоснованы

точные и приближенные алгоритмы их решения (см. [1-3] и цитированные там работы). Однако статус вычислительной сложности многих задач из этой совокупности пока не выяснен. Установление статуса комбинаторной сложности этой части задач представляется важным делом ближайшей перспективы, поскольку они являются специальными случаями задач обработки последовательностей с более сложной структурой.

Работа поддержана РФФИ, проекты № 06-01-00058 и № 07-07-00022.

Литература

[1] *Кельманов А. В.* Апостериорный подход к решению типовых задач анализа и распознавания числовых квазипериодических последовательностей: обзор результатов // ММРО-12 — Москва: МаксПресс, 2005. — С. 125–128.

[2] *Кельманов А. В.* Проблемы оптимизации в типовых задачах помехоустойчивой апостериорной обработки числовых последовательностей с квазипериодической структурой // Докл. 3-й Всеросс. конф. «Проблемы оптимизации и экономические приложения». — Омск: ОмГТУ, 2006. — С. 37–41.

[3] *Кельманов А. В.* Полиномиально разрешимые и NP-трудные варианты задачи оптимального обнаружения в числовой последовательности повторяющегося фрагмента // Докл. Всеросс. конф. «Дискретная оптимизация и исследование операций». — Владивосток-Новосибирск: ИМ СО РАН, 2007. — http://math.nsc.ru/conference/door07/.

Распознавание числовой квазипериодической последовательности, включающей повторяющийся набор эталонных фрагментов

Кельманов А. В., Михайлова Л. В., Хамидуллин С. А.

kelm@math.nsc.ru, okolnish@math.nsc.ru, kham@math.nsc.ru

Новосибирск, Институт математики им. С. Л. Соболева СО РАН

Рассматриваемая задача дополняет список изученных полиномиально разрешимых и NP-трудных задач комбинаторной оптимизации, возникающих в рамках нетрадиционного слабо изученного подхода к помехоустойчивому анализу и распознаванию числовых последовательностей с квазипериодической структурой [1]. Сущность этого подхода состоит в апостериорном (off-line) способе обработки последовательности при формализации содержательной задачи как задачи принятия решения.

Одна из возможных содержательных трактовок задачи состоит в следующем. Источник сообщений через канал связи с помехами передает информацию о некотором физическом объекте в виде повторяющегося упорядоченного эталонного набора импульсов различной формы, но одинаковой длительности. Имеется конечная совокупность отличающихся объектов. Каждому объекту этой совокупности соответствует единствен-

ный эталонный набор импульсов, а совокупности объектов — множество (словарь) эталонных наборов, размерности которых в общем случае различны. Словарь известен. На приёмную сторону через канал передачи поступает квазипериодическая последовательность импульсов, искаженная аддитивным шумом. Моменты времени появления импульсов в принятой (наблюдаемой) зашумленной последовательности и общее число переданных импульсов неизвестны. Требуется установить, какому объекту из совокупности соответствует принятая импульсная последовательность. Иными словами, требуется идентифицировать (распознать) принятую последовательность как последовательность, порожденную неизвестным эталонным набором из словаря.

Числовая квазипериодическая последовательность, в составе которой имеется повторяющийся упорядоченный набор из L ненулевых фрагментов размерности q, определяется следующей формулой общего члена:

$$x_n = \sum_{m\in\mathbb{M}} u_{n-n_m}\big(l(m,L)\big), \quad n = 0,\dots,N-1,$$

где

$$\begin{aligned}
&l(m,L) = (m-1) \bmod L + 1,\\
&u_{n-n_m}\big(l(m,L)\big) = 0, \text{ если } n - n_m \neq 0,\dots,q-1,\\
&\big(u_0(l(m,L)),\dots,u_{q-1}(l(m,L))\big)\in\mathbb{R}^q,\\
&0 < \big\|\big(u_0(l(m,L)),\dots,u_{q-1}(l(m,L))\big)\big\| < \infty
\end{aligned}$$

при каждом $m \in \mathbb{M} = \{1,\dots,M\}$, причем $M \geqslant L$,

$$(n_1,\dots,n_M) \in \Omega = \bigcup_{M=M_{\min}}^{M_{\max}} \Omega_M,$$

$$\Omega_M = \left\{ (n_1,\dots,n_M) \;\middle|\; \begin{array}{l} 0 \leqslant n_1 \leqslant N^+ \leqslant N-q \\ 0 \leqslant N^- \leqslant n_M \leqslant N-q \\ 0 < q \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N-q \\ m = 2,\dots,M \end{array} \right\},$$

а $M_{\min}$ и $M_{\max}$ находятся из решения системы неравенств, входящих в определение множества Ω_M, в которой N^-, N^+, $T_{\min}$ и $T_{\max}$ — целые числа. Термин «квазипериодическая последовательность» (т. е. последовательность, квазипериодически изменяющая свои свойства) обусловлен спецификой ограничений, входящих в определение множества Ω_M.

Заметим, что $l(m,L) \in \{1,\dots,L\}$ для любого $m \in \mathbb{M}$. Положим $U_j = \big(u_0(j),\dots,u_{q-1}(j)\big)$, и назовем U_j, $j = 1,\dots,L$, *эталонным вектором*, а $(U_1,\dots,U_L)$ — *эталонным набором*. Фрагмент $(x_{n_m},\dots,x_{n_m+q-1})$,

$m \in \mathbb{M}$, последовательности x_n, $n = 0, \ldots, N-1$, совпадающий с вектором U_j, $j = 1, \ldots, L$, будем называть *эталонным фрагментом*.

Допустим, что $(U_1, \ldots, U_L) \in \mathbb{W}$, где $\mathbb{W}$ — конечное подмножество (словарь, $|\mathbb{W}| = K$) множества всевозможных наборов, составленных из векторов размерности q:

$$\mathbb{W} \subset \left\{ (U^{(1)}, \ldots, U^{(i)}) \;\middle|\; \begin{array}{l} U^{(k)} \in \mathbb{R}^q,\ 0 < \|U^{(k)}\| < \infty,\ k = 1, \ldots, i, \\ 1 \leqslant i \leqslant M_{\max} \end{array} \right\}.$$

Предполагается, что вектор $X = (x_0, \ldots, x_{N-1})$, последовательность компонент которого включает повторяющиеся упорядоченые совокупности фрагментов, совпадающие с набором $(U_1, \ldots, U_L)$ эталонных векторов, недоступен для непосредственой обработки из-за вектора помехи $E = (\varepsilon_0, \ldots, \varepsilon_{N-1}) \in \Phi_{0,\sigma^2 I}$, где $\Phi_{0,\sigma^2 I}$ — нормальное распределение. Доступным для обработки считается вектор $Y = X + E$. При этом вектор X рассматривается как функция $X(n_1, \ldots, n_M, U_1, \ldots, U_L)$.

Задача распознавания состоит в следующем.

Дано: вектор Y и множество $\mathbb{W}$, включающее K эталонных наборов; если значения N^-, N^+, $T_{\min}$ и $T_{\max}$ неизвестны, то полагаем $N^- = 0$, $T_{\min} = q$, $T_{\max} = N^+ = N - q$.

Найти: набор $(U_1, \ldots, U_L) \in \mathbb{W}$, доставляющий максимум функционалу правдоподобия $\mathcal{L}\big(X(n_1, \ldots, n_M, U_1, \ldots, U_L) \mid Y\big)$.

В работе показано, что максимально правдоподобное распознавание числовой квазипериодической последовательности (искаженной аддитивной гауссовской некоррелированной помехой), включающей повторяющийся набор эталонных фрагментов, в случае, когда суммарное число M фрагментов в последовательности неизвестно, сводится к задаче отыскания наборов $(n_1, \ldots, n_M) \in \Omega$ и $(U_1, \ldots, U_L) \in \mathbb{W}$ таких, что

$$\sum_{m \in \mathbb{M}} \Big\{ 2\big(Y_{n_m}, U_{l(m,L)}\big) - \big\|U_{l(m,L)}\big\|^2 \Big\} \to \max,$$

где $(\cdot, \cdot)$ — скалярное произведение векторов, а $Y_n = (y_n, \ldots, y_{n+q-1})$, $n = 0, \ldots, N - q + 1$. Обоснован точный эффективный алгоритм решения этой задачи, имеющий временную сложность

$$O\big(L_{\max} K (T_{\max} - T_{\min} + 1)(N - q + 1)\big) = O(KN^3),$$

где $L_{\max}$ — максимальная размерность эталонного набора в словаре $\mathbb{W}$. Этот алгоритм является ядром алгоритма распознавания, устойчивого к помехам.

Полиномиальная разрешимость близких по своей сути задач установлена в [2–4]. В [2–3] обоснованы точные алгоритмы, обеспечивающие решение задач помехоустойчивого обнаружения повторяющегося набора эталонных фрагментов размерности L в случаях, когда суммарное число M фрагментов в последовательности известно и неизвестно. Алгоритмы имеют временную сложность, соответственно,

$$O\big(M(T_{\max} - T_{\min} + 1)(N - q + 1)\big) = O(MN^2);$$
$$O\big(\min\{L, M_{\max}\}(T_{\max} - T_{\min} + 1)(N - q + 1)\big) = O(N^3).$$

В [4] обоснован алгоритм, гарантирующий максимально правдоподобное распознавание числовой последовательности, включающей повторяющийся набор эталонных фрагментов в случае, когда суммарное число фрагментов в последовательности известно. Этот алгоритм имеет временную сложность

$$O\big(MK(T_{\max} - T_{\min} + 1)(N - q + 1)\big) = O(MKN^2).$$

Работа поддержана РФФИ, проекты № 06-01-00058 и № 07-07-00022.

Литература

[1] *Кельманов А. В.* О некоторых полиномиально разрешимых и NP-трудных задачах анализа и распознавания последовательностей с квазипериодической структурой // ММРО-13 (в настоящем сборнике). — 2007. — С. 261–264.

[2] *Кельманов А. В., Михайлова Л. В., Хамидуллин С. А.* Апостериорное обнаружение в квазипериодической последовательности повторяющегося набора эталонных фрагментов // ЖВМиМФ. (в печати).

[3] *Кельманов А. В., Михайлова Л. В., Хамидуллин С. А.* Оптимальное обнаружение в квазипериодической последовательности повторяющегося набора эталонных фрагментов // Докл. Всеросс. конф. «Дискретная оптимизация и исследование операций». — Владивосток-Новосибирск: Ин-т математики СО РАН, 2007. — `http://math.nsc.ru/conference/door07/`.

[4] *Кельманов А. В., Михайлова Л. В., Хамидуллин С. А.* Задача распознавания квазипериодической последовательности, включающей повторяющийся набор эталонных фрагментов // Докл. Всеросс. конф. «Дискретная оптимизация и исследование операций». — Владивосток-Новосибирск: Ин-т математики СО РАН, 2007. — `http://math.nsc.ru/conference/door07/`.

Задача монотонизации выборки

Таханов Р. С.
takhanov@mail.ru
Москва, МФТИ, ЗАО «Форексис»

Требования к классифицирующим правилам в задачах обучения по прецедентам состоят из двух частей — требования согласованности с прецедентными данными и удовлетворения некоторым заранее установленным дополнительным ограничениям. Одним из популярных типов подобных дополнительных ограничений являются ограничения монотонности. В некоторых случаях, однако, эти два типа ограничений могут быть взаимно противоречивыми, тогда возникает задача минимальной коррекции прецедентных данных.

Итак, рассмотрим следующее обобщение этой задачи, которую обозначим как MaxCMS (Maximal Consistent with Monotonicity Set).

MaxCMS. Заданы конечные множества B_n, B_m, где $B_r = \{1, \ldots, r\}$, на них отношения частичного порядка $\succcurlyeq^1, \succcurlyeq^2$ соответственно, и функция $\varphi\colon B_n \to B_m$. Для каждого элемента $i \in B_n$ задан положительный целочисленный вес w_i. Требуется найти максимальное по весу подмножество $B \subseteq B_n$, такое, что функция φ, ограниченная на B, является монотонной, то есть для любых $i, j \in B$ из $i \succcurlyeq^1 j$ следует $\varphi(i) \succcurlyeq^2 \varphi(j)$. Его вес обозначим MaxCMS.

Введем на множестве B_n частичный предпорядок (транзитивный и рефлексивный бинарный предикат): $i \succ j \Leftrightarrow \varphi(i) \succcurlyeq^2 \varphi(j)$. Рассмотрим орграф $G = (V, E)$, где $V = B_n$, а $E = \left\{(i,j) \;\middle|\; i \succcurlyeq^1 j,\ \varphi(i) \not\succcurlyeq^2 \varphi(j)\right\}$. Орграф G также может быть задан равенствами $V = B_n$ и $E = (\succcurlyeq^1 \cap \overline{\succ})$, где $\overline{\succ}$ — дополнение бинарного предиката.

Определение 1. *Всякий орграф, множество дуг которого может быть представлено как пересечение некоторых частичного порядка и дополнения частичного предпорядка на вершинах орграфа, называется специальным.*

Теорема 1. *Решение* MaxCMS *равно максимальному независимому множеству специального орграфа G.*

Теорема 2. MaxCMS — *NP-трудная задача.*

Определение 2. *Задача* MaxCMS *с входом $\left(\succcurlyeq^1, \succcurlyeq^2, \varphi, w\right)$ для случая, когда размерность частичного порядка $\succcurlyeq^2$ равна d, называется* d-MaxCMS.

Теорема 3. *d*-MaxCMS *сводится к нахождению максимального независимого множества в орграфе* $G = (V, E)$, *где* $E = (\succ^1 \cup \cdots \cup \succ^d)$, *и предикаты* $\succ^s$ *транзитивны, причем в* G *нет циклов.*

Теорема 4. 1-MaxCMS *полиномиально разрешима [4].*

1-MaxCMS сводится к следующей задаче линейного программирования, решаемую посредством метода эллипсоидов Хачияна [1]:

$$\begin{aligned}
&\sum\nolimits_{v\in V} w_v y(v) \to \max;\\
&\sum\nolimits_{v\in\Gamma} y(v) \leqslant 1, \quad \Gamma \in \mathbb{G}(s,t);\\
&y(v) \geqslant 0, \quad v \in V;
\end{aligned}$$

где $\mathbb{G}(s,t)$ — множество всех путей в орграфе $G = (V, E)$ из некоторого минимального элемента в некоторый максимальный элемент частичного порядка, определяемого орграфом G. Данный политоп обозначим через $\Pi(G)$.

Рассмотрим теперь задачу 2-MaxCMS. Рассмотрим два орграфа: $G_1 = \left(V, \succ^1\right)$ и $G_2 = \left(V, \succ^2\right)$. Заметим, что максимальное независимое множество орграфа $G = (V, E)$ является независимым множеством в обоих G_1 и G_2.

Рассмотрим следующую оптимизационную задачу:

$$\begin{aligned}
&\varphi(\overline{x}, \overline{y}) \to \max;\\
&\overline{x} \in \Pi(G_1); \quad \overline{y} \in \Pi(G_2);
\end{aligned}$$

где $\varphi(\overline{x}, \overline{y}) = -\frac{1}{2}\sum\limits_{v\in V} w_v (x_v - y_v)^2 - w_v(x_v + y_v)$. Будем называть ее выпуклой.

Рассмотрим следующий приближенный алгоритм для 2-MaxCMS.

1. Найти методом эллипсоидов для выпуклой оптимизации [2, 1] пару $\left(\overline{x'}, \overline{y'}\right)$ такую, что

$$\max_{(\overline{x},\overline{y})\in\Pi(G_1)\times\Pi(G_2)} \varphi(\overline{x}, \overline{y}) \leqslant \varphi(\overline{x}', \overline{y}') + \varepsilon, \quad |x'_i - y'_i| \leqslant \tfrac{1}{2}, \quad \varepsilon = \tfrac{1}{16}.$$

2. Найти $\overline{x}^* = \arg\max\limits_{\overline{x}\in\Pi(G_1)} \psi(\overline{x}, \overline{y}')$ и $\overline{y}^* = \arg\max\limits_{\overline{y}\in\Pi(G_2)} \psi(\overline{x}^*, \overline{y})$, где $\psi(\overline{x}, \overline{y}) = \sum\limits_{v\in V} w_v x_v y_v$. Здесь $\overline{x}^*, \overline{y}^*$ целочисленны.

Ответ алгоритма — множество вершин $\{v | x^*_v y^*_v = 1\}$.

Рассмотрим MaxCMS как задачу минимального вершинного покрытия. Фактически это означает, что задача заключается в удалении «шума» в обучающей выборке для построения корректного классификатора.

Очевидно, что этим самым «шумом» и является соответствующее минимальное вершинное покрытие.

Введем обозначения $W = \sum_{v \in V} w_v$ и $\mathrm{MaxCMS} = \alpha W$.

Ясно, что $0 \leqslant \alpha \leqslant 1$.

Теорема 5. *Алгоритм является полиномиальным.*

Теорема 6. *Алгоритм аппроксимирует минимальное вершинное покрытие G с константой $1 + \alpha \leqslant 2$, при $\alpha \geqslant \frac{1}{2}$.*

С учетом того, что стандартный алгоритм для вершинного покрытия [3] имеет константу аппроксимации 2, предложенный алгоритм является более точным.

Литература

[1] *Хачиян Л. Г.* Полиномиальный алгоритм в линейном программировании // Доклады АН СССР. — 1979. — Т. 244. — С. 1093–1096.

[2] *Grotshel M., Lovasz L., Schrijver A.* Geometric algorithms and combinatorial optimization. — Springer-Verlag, 1988.

[3] *Hochbaum D. S.* Approximation algorithms for the set covering and vertex cover problems // SIAM Journal on Computing. — 1982. — No 11. — Pp. 555–556.

[4] *Mohring R. H.* Algorithmic aspects of comparability graphs and interval graphs // Graphs and Order. — Dordrecht: Reidel, 1985. — Pp. 41–101.

Вычислительная и аппроксимационная сложность задачи о комитетной отделимости конечных множеств

Хачай М. Ю.

mkhachay@imm.uran.ru

Екатеринбург, Институт математики и механики УрО РАН

В работах [1, 2] исследована вычислительная сложность задачи MASC о минимальном аффинном разделяющем комитете для конечных множеств $A, B \subset \mathbb{Q}^n$. В частности, показано, что эта задача NP-трудна и не принадлежит классу Apx (если $P \neq NP$). В сообщении приведена оценка порога эффективной аппроксимируемости задачи MASC. Отдельно исследуется вопрос о вычислительной сложности задачи при дополнительных ограничениях, например, фиксированной размерности пространства. Показывается, что задача о комитетной отделимости остается труднорешаемой, даже будучи заданной на плоскости (т. е. в наиболее простом нетривиальном случае).

Формулировка задачи

Определение 1. *Конечная последовательность функций*

$$Q = (f_1, \ldots, f_q), \quad f_i(x) = \alpha_i^T x - \beta_i,$$

называется аффинным комитетом, разделяющим множества $A, B \subset \mathbb{R}^n$, *если выполнено условие*

$$\big|\{i \in \mathbb{N}_q \,|\, f_i(a) > 0\}\big| > \tfrac{q}{2}, \quad a \in A;$$
$$\big|\{i \in \mathbb{N}_q \,|\, f_i(b) < 0\}\big| > \tfrac{q}{2}, \quad b \in B.$$

Число q *называется числом элементов (членов) комитета* Q.

Задача о минимальном аффинном разделяющем комитете (MASC). Заданы множества $A = \{a_1, \ldots, a_{m_1}\}$ и $B = \{b_1, \ldots, b_{m_2}\}$, $A, B \subset \mathbb{Q}^n$. Требуется указать аффинный комитет Q с наименьшим числом элементов, разделяющий множества A и B.

Теорема 1 ([2]). *Задача MASC NP-трудна. При условии* $P \neq NP$ *задача MASC не принадлежит классу Apx.*

Задача MASC остается NP-трудной при дополнительном ограничении

$$A \cup B \subset \big\{z \in \{0,1,2\}^n \colon\ |z| \leqslant 2\big\}.$$

Ниже приведен новый результат, усиливающий утверждение Теоремы 1.

Теорема 2. *Справедливость условия* $NP \not\subseteq TIME(2^{\mathrm{poly}(\log n)})$ *влечет отсутствие приближенных алгоритмов для задачи MASC с точностью аппроксимации* $O(\log\log\log m)$.

Труднорешаемость задачи о комитетной отделимости на плоскости

Остановимся на важном частном случае задачи о комитетной отделимости при дополнительном условии фиксированности размерности n пространства признаков. Известно [3], что задача о минимальном аффинном разделяющем комитете, заданная в одномерном пространстве, полиномиально разрешима. Ниже показано, что при $n = 2$ (следовательно, и при произвольном фиксированном $n > 1$) она NP-трудна.

Определение 2. *Множество прямых*

$$\mathcal{L} = \{l_1, \ldots, l_s\}, \quad l_j = \big\{x \in \mathbb{R}^2 \;\big|\; c_j^T x = d_j\big\},$$

называется покрытием множества $P = \{p_1, \ldots, p_k\} \subset \mathbb{R}^2$, *если для каждой точки* $p \in P$ *найдется прямая* $l = l(p) \in \mathcal{L}$ *такая, что* $p \in l$.

Как обычно, перейдем к рассмотрению комбинаторных задач, сформулированных в виде задач распознавания свойства.

Задача о покрытии прямыми конечного множества на плоскости (PC). Заданы множество $P = \{p_1, \ldots, p_k\} \subset \mathbb{Z}^2$ и число $s \in \mathbb{N}$. Существует ли покрытие $\mathcal{L}$ множества P по мощности не превосходящее s?

Задача о комитетной отделимости конечных множеств на плоскости (PASC). Заданы множества $A = \{a_1, \ldots, a_{m_1}\}$ и $B = \{b_1, \ldots, b_{m_2}\}$, $A, B \subset \mathbb{Q}^2$, и число $t \in \mathbb{N}$. Существует ли аффинный комитет Q, разделяющий множества A и B и состоящий из не более чем t элементов?

Известно [4], что задача PC NP-полна. Задача PASC является частным случаем задачи ASC [2], полученным фиксацией размерности пространства. Легко убедиться в том, что обе задачи принадлежат классу NP. Ниже описывается полиномиальная сводимость задачи PC к задаче PASC, влекущая принадлежность последней классу NP-полных задач.

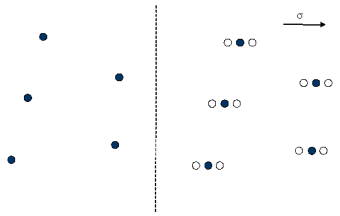


Рис. 1. Сведение задачи PC к задаче PASC.

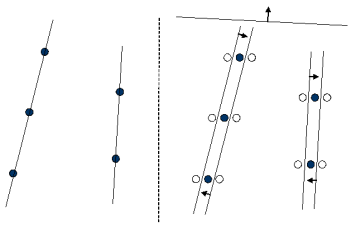


Рис. 2. Идея доказательства теоремы 3.

Пусть условие частной задачи PC задается множеством $P = \{p_1, \ldots, p_k\} \subset \mathbb{Z}^2$ и числом $s \in \mathbb{N}$. Вычислим $\rho = \max\{|p_i| : i \in \mathbb{N}_k\}$ и положим $\varepsilon = \frac{1}{6(2\rho+1)+1}$. Зафиксируем вектор σ, $|\sigma| = 1$ так, чтобы для любого $\{i, j\} \subset \mathbb{N}_k$ отрезки $[p_i - \varepsilon\sigma, p_i + \varepsilon\sigma]$ и $[p_j - \varepsilon\sigma, p_j + \varepsilon\sigma]$ не лежали на одной прямой. Сопоставим исходной задаче PC частную задачу PASC с условием: $A = P$, $B = (P - \varepsilon\sigma) \cup (P + \varepsilon\sigma)$ и $t = 2s + 1$ (Рис. 1). Нетрудно видеть, что описанные выше действия могут быть произведены за время, ограниченное сверху полиномом от длины записи условия задачи PC.

Теорема 3. *Множество $P = \{p_1, \ldots, p_k\} \subset \mathbb{Z}^2$ обладает покрытием из s прямых тогда и только тогда, когда множества $A = P$ и $B = (P - \varepsilon\sigma) \cup (P + \varepsilon\sigma)$ отделимы аффинным комитетом из $2s + 1$ элемента.*

Доказательство теоремы в большей степени конструктивно и представляет собой, по сути, обоснование корректности алгоритма сопоставления известному покрытию множества P аффинного комитета, разделяющего соответствующие множества A и B, иллюстрация которого приведена на Рис. 2.

Следствие 1. *Задачи PASC и ASC (в пространстве фиксированной размерности $n > 1$) NP-полны в сильном смысле.*

Следствие 2. *Задача MASC при фиксированном $n > 1$ NP-трудна в сильном смысле.*

Работа выполнена при поддержке РФФИ, проект № 07-07-00168 и грантов Президента РФ, проекты МД-6768.2006.1 и НШ-5955.2006.1.

Литература

[1] *Хачай М. Ю.* О вычислительной сложности задачи о минимальном комитете и смежных задач // ДАН. — 2006. — Т. 406, № 6. — С. 742–745.

[2] *Хачай М. Ю.* О вычислительной и аппроксимационной сложности задачи о минимальном аффинном разделяющем комитете // ТВИМ. — 2006. — № 1. — С. 34–43.

[3] *Мазуров Вл. Д.* Комитеты систем неравенств и задача распознавания // Кибернетика. — 1971. — № 3. — С. 140–146.

[4] *Megiddo N., Tamir A.* On the complexity of locating linear facilities in the plane // Operations Research Letters. — 1982. — Vol. 1 № 5. — С. 194–197.

Обработка сигналов и анализ изображений

Код раздела: SI (Signal Processing and Image Analysis)

- Теория, методы и прикладные задачи обработки, анализа и распознавания сигналов.
- Фурье-анализ и вейвлет-анализ.
- Обработка и распознавание речи.
- Теория, методы и прикладные задачи обработки, анализа, распознавания, понимания и синтеза изображений.
- Обработка видеоизображений.

Определение моментов начала нот (онсетов) при анализе музыкальных произведений

Андреенко С. А.
andreenk@widisoft.com
Москва, WIDISOFT

В настоящее время в мире проводятся интенсивные исследования, посвященные созданию автоматических систем анализа музыкальных записей. Такие системы начинают применяться для идентификации музыкальных композиций, контекстного музыкального поиска и контроля над соблюдением авторских прав. Чтобы создать эффективную систему анализа музыки, необходимо решить достаточно большое количество различных задач, таких как анализ ритма, тональности, гармонии, и т. д.

Данная работа посвящена поиску онсетов — моментов начала нот. Мы ограничимся инструментами с хорошо выраженной атакой, но достаточно сложным звуком, например, фортепиано. В записи одновременно с онсетом могут присутствовать частотные компоненты (основной тон и гармоники) уже звучащих нот, кроме того, частотные компоненты не являются непрерывными из-за биений, поэтому поиск онсетов является непростой задачей (Рис. 1).

В работах [1, 2] предлагается разделить спектр звукового сигнала на несколько полос и следить за уровнем и фазой сигнала в каждой полосе. Решение о наличии онсета предлагается принимать на основании вычисления расстояния между векторами амплитуд и фаз в последовательные моменты времени. Наши попытки повторить эту методику не дали ожидаемых результатов, и мы предлагаем другой способ поиска онсетов.

Нейросетевой подход

Сигнал разбивается банком фильтров на достаточно большое число полос, шириной в полутон ($2^{1/12}$ от центральной частоты), при этом центральные частоты соответствуют высотам нот в традиционной равномерно темперированной шкале. Вычисляются отсчёты интенсивности сигнала в каждой полосе через промежуток времени порядка 10 мс, таким способом строится сонограмма, аналогичная изображенной на Рис. 1. Поскольку нота представляет из себя сумму основного тона и гармоник, то в момент начала ноты появляется сразу несколько частотных компонент. Поэтому онсеты определяются в два этапа — сначала определяются те частотные полосы, в которых появились компоненты ноты, а потом определяется нота или ноты, способные породить такой набор частотных компонент.

Для решения этой задачи строится двухуровневая нейронная сеть. Первый уровень определяет появление частотной компоненты на каждой

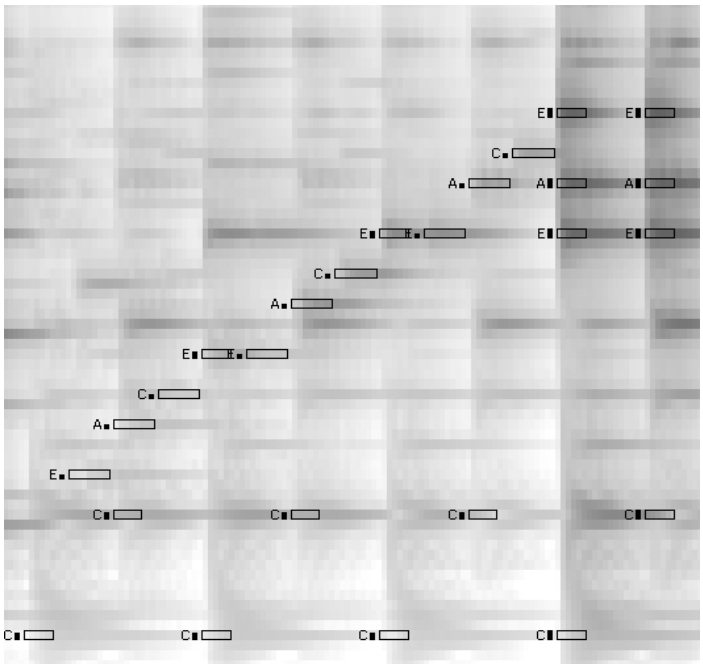


Рис. 1. Фрагмент сонограммы фортепианного произведения.

из полос частот. На вход этого уровня подаются близкие (и по времени, и по частоте) отсчёты сонограммы, т. е. значения из квадрата с центром в текущем моменте времени и полосе частот. Нейросеть обучается на эталонных записях, где положения онсетов известны, методом обратного распространения ошибки. При этом для каждой полосы частот (высоты ноты) строится собственная нейросеть.

Нейросеть второго уровня получает на вход результаты работы всех нейросетей первого уровня в текущий момент времени, и предназначена для определения собственно начал нот. Эта сеть анализирует появление частотных компонент в совокупности и принимает решение о появлении ноты. Данная сеть также позволяет скорректировать ошибки сетей первого уровня, поскольку детектирование появления только одной частотной компоненты с большой вероятностью не означает появления ноты. Это позволяет бороться с эффектом биений, когда возможно периодическое пропадание основного тона ноты. Нейросеть второго уровня также обучается методом обратного распространения ошибки по эталонным записям.

Использование такого механизма позволяет получить достаточно надежное и эффективное детектирование онсетов для фортепианных записей. Полученные таким образом онсеты могут быть использованы для системы определения ритма и темпа, а также для восстановления нотного текста произведения.

Литература

[1] *Duxbury C, Sandler M, Davies M* A Hybrid Approach to Musical Note Onset Detection // Proc. of the 5th Int Conference on Digital Audio Effects (DAFx-02), Hamburg, 2002.

[2] *Duxbury C, Bello J, Sandler M, Davies M* A Combined Phase and Amplitude Based Approach to Onset Detection for Audio Segmentation // Proc. of WIAMIS 2003, Queen Mary: University of London, 2003.

Идентификация векторных полей при анализе изображений

Бакина И. Г., Голов Н. И.

irina_msu@mail.ru, golov@forecsys.ru

Москва, МГУ им. М. В. Ломоносова

В данной работе исследуется возможность описания линейными векторными полями двумерных изображений с ярко выраженной морфологической структурой: узоров папиллярных линий, радужных оболочек глаз или рельефов местности. Изображения представляются в виде набора векторов, для описания которых строится аппроксимирующее векторное поле. В данной работе векторное поле рассматривается как набор линейных векторных функций. Каждая векторная функция ищется явно, методом наименьших квадратов. В работе описывается разработанный алгоритм идентификации линейной векторной функции и исследуется устойчивость отдельных линейных векторных функций к поворотам и искажениям исходных изображений.

Представление изображения в виде набора векторов

Получение множества векторов из исходного изображения может осуществляться несколькими методами. В данной работе рассматривается построение множества векторов для дактилоскопического узора как набора направляющих векторов ветвей дискретно-непрерывного скелета.

Линейное векторное поле

Линейное векторное поле задается системой линейных дифференциальных уравнений

$$\begin{cases} \partial x = ax + by - m; \\ \partial y = cx + dy - n; \end{cases} \tag{1}$$

и описывает следующие семейства кривых: узел, седло, фокус, центр, вырожденный узел и диакритический узел. Классификация осуществляется на основе анализа собственных значений основной матрицы системы (1).

Математическая модель линейного векторного поля задается небольшим набором параметров. Ее выбор обусловлен также: наличием в системе (1) только линейной зависимости, что позволяет применить методы линейной алгебры для идентификации и исследования модели; если $\partial x, \partial y$ рассматривать как градиент некоторой функции потенциала, то систему (1) можно рассматривать как линейную составляющую разложения функции потенциала в ряд Тейлора.

Идентификация модели линейного векторного поля

Пусть $\mathbf{P} = \{\mathbf{p_i}\}_{i=1}^{N}$ — множество исходных векторов, где $\mathbf{p_i} = \big((x_i, y_i); (\partial x_i, \partial y_i)\big)$ — вектор с координатами $(\partial x_i, \partial y_i)$ в точке (x_i, y_i), а N — общее число векторов. Введем функционал ошибки, определяющий качество аппроксимации множества векторов $\{\mathbf{p_i}\}_{i=1}^{N}$ векторным полем (1):

$$\mathrm{Err}\big(\mathbf{P}, \{a, b, c, d, m, n\}\big) = \sum_{i=1}^{N} (\partial x_i \cdot \partial y_i^{\nu} - \partial y_i \cdot \partial x_i^{\nu})^2, \tag{2}$$

где $(\partial x_i^{\nu}, \partial y_i^{\nu})$ — координаты аппроксимирующего векторного поля (1) в точке (x_i, y_i).

Введение функционала ошибки позволяет сформулировать задачу идентификации модели векторного поля как оптимизационную задачу:

$$\mathrm{Err}\big(\mathbf{P}, \{a, b, c, d, m, n\}\big) \to \min_{\{a,b,c,d,m,n\}}.$$

Результат одновременной оптимизации всех шести параметров оказался неустойчивым. Поэтому шестимерная задача была сведена к четырехмерной путем введения дополнительной информации.

Рассмотрим *опорный* вектор поля $\mathbf{p_0} = \big((x_0, y_0); (\partial x_0, \partial y_0)\big)$ и выразим параметры m и n поля через его координаты:

$$\begin{cases} m = ax_0 + by_0 - \partial x_0; \\ n = cx_0 + dy_0 - \partial y_0. \end{cases} \tag{3}$$

Опорный вектор задает смещение центра поля относительно начала координат. В работе рассмотрен выбор опорного вектора как центра тяжести системы, а также перебор по множеству исходных векторов. Подставляя (1), (3) в (2), получим квадратичный полином от четырех параметров a, b, c, d, причем все квадраты входят в него с положительными

Алгоритм 1. Поиск оптимальных собственных значений.

Вход: $\{\mathbf{p_i}\}_{i=1}^{N}$;
Выход: $\lambda_{1,2}^{*}$ — оптимальные собственные значения;
1: **для** $i = 1, \ldots, N$
2: Построить аппроксимирующее векторное поле, взяв вектор $\mathbf{p_i}$ в качестве опорного;
3: Подсчитать собственные значения $\lambda_{1,2}^{i}$ для построенного поля;
4: Вычислить взвешенное среднее собственных значений $\lambda_{1,2}^{*}$:

$$\lambda_{1,2}^{*} = \frac{\sum_{i=1}^{N} \alpha_i \lambda_{1,2}^{i}}{\sum_{i=1}^{N} \alpha_i}, \text{ где } \alpha_i = \frac{1}{\mathrm{Err}(\mathbf{p_i}, \{a_i, b_i, c_i, d_i, m_i, n_i\})};$$

знаками. Если не рассматривать вырожденные случаи, должен присутствовать единственный глобальный экстремум — минимум, в котором все производные по переменным равны нулю. В результате имеем систему линейных алгебраических уравнений с квадратной матрицей, решая которую, получим параметры поля a, b, c, d.

Для поиска устойчивых собственных значений основной матрицы системы (1) предлагается Алгоритм 92.

Анализ отпечатков пальцев человека

Общность и устойчивость описанного выше подхода исследовались на примере задачи построения описания отпечатков пальцев человека. Под общностью понимается возможность применять описанное выше представление для описания всех возможных типов отпечатков пальцев. Под устойчивостью понимается устойчивость описания к условиям и погрешностям съемки отпечатков пальцев. Так как очевидно, что одной линейной векторной функции вида (1) недостаточно, чтобы описать все разнообразие реальных изображений отпечатков пальцев, был разработан многошаговый алгоритм построения описания, состоящего из нескольких линейный векторных функций. Алгоритм включает в себя следующие шаги:

1. Построение скелета изображения.
2. Выделение векторов на основе скелета.
3. Разбиение множества векторов на кластеры.
4. Аппроксимация каждого кластера векторным полем.

Шаг 4 выполняется на основе описанного выше алгоритма. При работе с реальными изображениями отпечатков пальцев на шаге 3 образовывался один центральный кластер, и некоторое количество вспомога-

Отпечаток	$\lambda^*_{1,2}$	Отпечаток	$\lambda^*_{1,2}$
`Finger1`	$0,00038 \pm 0,00373i$	`Finger2`	$0,00154 \pm 0,00289i$
`Finger1`(90^0)	$0,00037 \pm 0,00376i$	`Finger2`(90^0)	$0,00155 \pm 0,00287i$
`Finger1`(180^0)	$0,00038 \pm 0,00372i$	`Finger2`(180^0)	$0,00154 \pm 0,00288i$
`Finger1`(270^0)	$0,00037 \pm 0,00371i$	`Finger2`(270^0)	$0,00155 \pm 0,00284i$

Таблица 1. Результат работы метода на двух отпечатках пальцев.

тельных. Линейное векторное поле центрального кластера в дальнейшем и использовалось для оценки общности и устойчивости метода.

Результаты работы метода

Результаты работы метода представлены в Таблице 1. В ней рассмотрено изменение собственных значений при трансформации отпечатков — повороты на 90°, 180° и 270° градусов. Каждый отпечаток описывался одним полем вида (1).

Работа поддержана РФФИ, проекты № 05-01-00542, № 07-07-00181.

Литература

[1] *Mestetskii L. M.* Fat curves and representation of planar figures. — Computers & Graphics, 2000 — No. 24. — Pp. 9–21.

[2] *Шильников Л. П., Шильников А. Л., Тураев Д. В., Чуа Л.* Методы качественной теории в нелинейной динамике, часть 1. — Институт компьютерных исследований, Москва-Ижевск, 2004.

[3] *Mumford D., Shah J.* Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems // Comm. Pure Appl. Math. — 1989. — Vol. XLII.

Реконструкция и визуализация городской обстановки по изображениям

Бобков В. А., Борисов Ю. С., Кудряшов А. П.

`kudryashova@iacp.dvo.ru`

Владивосток, Институт автоматики и процессов управления ДВО РАН

В работе рассматриваются методы трехмерной реконструкции и анимационной визуализации объектов городской среды по ограниченной, некалиброванной последовательности изображений. Предлагаемый подход базируется на интеграции в рамках единой методологии известных и активно развиваемых сегодня методов и алгоритмов компьютерного зрения (кросс-корреляционная идентификация особенностей, эпиполярные ограничения с вычислением фундаментальной матрицы и трифокального тензора, и др.) и методов, учитывающих специфику прикладно-

го контекста задачи (полигональность, параллелизм и ортогональность, свойственные городским сооружениям).

Калибровка камер

В качестве исходных данных рассматривается последовательность снимков городской сцены, представленной множеством зданий. Снимки сделаны с перекрытием, таким образом, чтобы для каждой пары соседних изображений нашлась одна точка-особеность (угол здания, окна), видимая на обоих изображениях. Требуется найти матрицы преобразований из некоторой единой евклидовой мировой системы координат (МСК) в системы координат камер (СКК). Решение ищем для тройки соседних видов. Привяжем единую МСК к некоторому углу (здания, окна) на снимке 1. Плоскость стены здания, содержащей этот угол, является плоскостью XY этой МСК. Одно из рёбер этой стены соответствует оси X. Если для некоторой тройки выполняется оговоренное выше условие: угол (МСК1) виден на снимках 1 и 2, но не виден на снимке 3, то для снимка 2 определяется дополнительно матрица преобразования из другого угла (МСК2), видного также на снимке 3. Задача полной калибровки решается с применением нелинейной оптимизации, отталкиваясь от вычисления точек схода (vanishing points) применительно к ортогональным семействам параллельных линий плоскостей (грани зданий) в МСК.

Определение точечных соответствий

Для выполнения калибровки камер необходимо определение точечных соответствий на изображениях. Для решения задач реконструкции трехмерных сцен городской обстановки по фотоизображениям необходимо высокое разрешение, порядка 5–10 мегапикселей, а известные методы эффективно работают лишь при разрешении порядка 0.3–1 мп [1]. Поэтому был разработан алгоритм для работы с фотоснимками высого разрешения, высокой скоростью выделения точечных особенностей и с приемлемыми для указанного приложения достоверностью и количеством особенностей. Схематично работа алгоритма может быть представлена следующим образом:

1. Разрешение изображения уменьшается до оптимального(примерно 400×300 точек).
2. На уменьшенном изображении ищутся точечные особенности с помощью метода, получившего название Difference-of-Gaussian (DoG).
3. На паре уменьшенных изображений ищутся соответствия кросс-корреляционным и методом нормированного евклидового расстояния.
4. Полученные соответствия перерасчитываются с уменьшенных изображений на исходные, учитывая возможное неточное сопоставление.

5. Применяется фильтр, позволяющий избавиться от большинства ложных сопоставлений.

Определение соответствия линий

Метод сопоставления линий [2] на изображениях полигональной сцены по трем видам основан на построении интегральной оценки сходства отрезков, учитывающей геометрическое и текстурное подобие отрезков. Геометрическое подобие оценивается с помощью трифокального тензора, а текстурное — применением кросскорреляции с учетом гомографии. Растровые изображения предварительно векторизуются, в результате чего формируется векторное представление каждого изображения в виде множества полилиний (ломаных). Для каждого из отрезков первого изображения ищется его образ на втором и третьем изображении. Для оценки вероятности правильного решения и отбраковки ложных решений выполняется последовательная многоступенчатая фильтрация с предварительно заданными порогами. Основными типами применяемой фильтрации являются: проверка геометрического правдоподобия отрезков-кандидатов, выполняемая на основе эпиполярного соответствия; текстурное сходство, выполняемое с учетом вычисленной гомографии; проверка геометрической связности (примыкание двух пространственных отрезков, наблюдаемое на любом виде). Эпиполярное соответствие отрезков всех трех изображений проверяется с помощью фундаментальной матрицы и трифокального тензора.

Трехмерная реконструкция

На первом этапе, используя фундаментальную матрицу и эпиполярные ограничения, необходимо каждую из трех сопоставленных линий достроить до максимально видимой длины на всех трех видах. Это даст нам соответствие концевых точек линий. После этого можно восстанавливать трехмерное положение концевых точек сопоставленных троек линий, используя полученные на этапе калибровки матрицы положения камер. Пространственное положение точки определяем решением триангуляционной задачи. С использованием полученного множества пространственных линий решается задача построения полигонов при ограничениях на трехмерные объекты («крыши» зданий параллельны «земле» (плоскости XZ), «стены» перпендикулярны той же плоскости и являются прямоугольниками). На полученные полигоны наносится наибольшая по площади текстура с одного из изображений сцены.

Построение новых видов

Наряду с алгоритмом полной реконструкции сцены, рассматривается построение новых видов плеоптическим методом без явного восстанов-

ления геометрии. Указанный метод является развитием подхода, описанного в [3].

Для выбранной точки нового вида по известной калибровке камер рассчитываются эпиполярные линии на базовых видах. Используя фундаментальную матрицу между базовыми видами, рассчитывается соответствие точек на построенных эпиполярных линиях. Полученные области сравниваются для нахождения максимально схожих, которые определяют образы пространственной точки. Полученные координаты образов применяются как текстурные, которые используются для нанесения на соответствующие полигональные участки нового вида частей базовых изображений.

Работа выполнена при финансовой поддержке Президиума РАН (Программа фундаментальных исследований №14, раздел 2).

Литература

[1] *Lowe D. G.* Distinctive Image Features from Scale-Invariant Keypoints — International Journal of Computer Vision, 60, 2, 2004 — Pp. 91–110

[2] *Бобков В. А., Роншин Ю. И., Кудряшов А. П.* Сопоставление линий по трем видам пространственной сцены — Информационные технологии и вычислительные системы. — № 2. — 2006. — С. 71–78.

[3] *Борисов Ю. С.* Визуализация городской обстановки пленоптическим методом — Сибирский журнал вычислительной математики. — № 2, Т. 9. — 2006 — С. 215–224

Распознавание составных объектов изображения на базе структурного и корреляционно-экстремальных методов

Васин Ю. Г., Лебедев Л. И.

lebedev@pmk.unn.ru

Нижний Новгород, НИИ прикладной математики и кибернетики
Нижегородского государственного университета им. Н. И. Лобачевского

В работе предлагается решение задачи распознавания составных объектов изображения, получаемых в результате «слипания» дискретных объектов (ДО) и, поэтому, не имеющих эталонных прототипов. Распознавание такого рода объектов осуществляется последовательно в два этапа на базе структурного и корреляционно-экстремальных контурных методов. Структурное распознавание применяется для идентификации частей составного объекта в целях разбиения его на совокупность ДО и последующего распознавания последних корреляционно-экстремальным контурным методом определения сходства разномасштабных форм. Приводятся результаты применения этого подхода для распознавания

составных объектов изображений в задачах автоматизации ввода документов.

Существующие методы распознавания составных объектов изображения с приемлемой сложностью вычислений накладывают, как правило, жесткие ограничения на тип входной информации и, поэтому, предназначены в основном для решения узкоспециализированных задач. Следует также отметить, что невозможно напрямую и применение контурного корреляционно-экстремального метода из-за возможного разнообразия эталонов, которые необходимо задать для решения задачи распознавания составных объектов. В то же время решение задачи автоматизации ввода документов с описанными дефектами является актуальной.

Постановка задачи и методы решения

В данной работе решение задачи распознавания составных объектов предлагается осуществлять на базе корреляционно-экстремальных контурных методов и структурного анализа информации о составном объекте.

Для этого решение задачи разбивается на два этапа. На первом этапе проводится структурный анализ составного объекта. Структурное распознавание осуществляется на основе инвариантных признаков, которые вычисляются на основе заданных примитивов корреляционно-экстремальным контурным методом. В качестве непроизводных элементов для вычисления инвариантных признаков выбираются фрагменты контуров эталонов. Фрагменты следует выбирать так, чтобы они, желательно, характеризовали только те ДО, с которых были взяты примитивы, то есть по возможности были уникальными с одной стороны, а с другой, не содержали бы участков контуров, наиболее вероятных к слипанию и образованию составного объекта. Очевидно, что получение правил вывода для решения классификационной задачи распознавания составных объектов в окончательном варианте является сложной проблемой обучения. К тому же структурное распознавание не обеспечивает в полном объеме характеристиками о распознанных ДО, необходимых для идентификации сформированных надписей. Поэтому, структурное распознавание используется здесь как предварительное в целях вычленения и формирования контуров ДО. Окончательное решение о распознанном дискретном объекте и его характеристиках осуществляется на базе корреляционно-экстремального контурного метода определения сходства разномасштабных форм. Это второй этап распознавания составных объектов. При таком подходе к распознаванию составных объектов значительно снижаются требования к качеству структурного распознавания, а следовательно, к упрощению как самой схемы классификатора, так и получения его решающих правил (например, в этом случае теряется необходимость в по-

строении решающих правил распознавания символов «Н» и «П» и/или формировании для этой цели отличительных фрагментов). Для получения описания ДО на основании контура составного объекта каждый фрагмент эталона, выделенный для структурного распознавания, дополнительно сопровождается заданием габаритных точек самого эталона. Габаритное описание эталона, в зависимости от его целевого использования, может иметь несколько ступеней, начиная с грубого, задаваемого вершинами выпуклого четырехугольника, более точного, — вершинами выпуклого многоугольника минимальной площади и, кончая, собственно, метрикой самого эталона. Грубое габаритное задание эталона предназначено в основном для целей «разрезания» контура составного объекта, тогда как его более точные габаритные задания используются для построения более приближенной к эталонному описанию контурной модели ДО в очерченной области. Для повышения эффективности и быстродействия алгоритма распознавания составных объектов к эталонному описанию фрагмента и габаритному описанию самого эталона могут быть добавлены так называемые контрольные точки, роль которых сводится к функции отсева бесперспективных областей местоположения ДО, полученных после структурного распознавания. Например, в качестве таких точек для распознавания символов могут быть взяты точки задающие направление надписи.

Полученные результаты и эксперимент

На основе предложенного подхода был реализован алгоритм распознавания составных объектов со следующей пошаговой структурой:

1. масштабирование эталонных фрагментов;
2. обнаружение на составном объекте участка контура сходного с текущим эталонным фрагментом и вычисление коэффициента сходства между ними ε_i^R корреляционно-экстремальным контурным методом;
3. определение множества эталонных фрагментов M_Ω, удовлетворяющих критерию отбора по величине коэффициента сходства;
4. отбор множества эталонных фрагментов M_ω из списка M_Ω на основе контрольных точек;
5. формирование описания ДО на основе описания составного объекта в выделенной габаритной области для каждого примитива из M_ω;
6. вычисление коэффициента сходства ДО с эталоном корреляционно-экстремальным контурным методом сравнения разномасштабных форм;
7. идентификация ДО и определение области его местоположения;
8. формирование нового описания составного объекта на основе предыдущего описания за вычетом описания распознанного ДО.

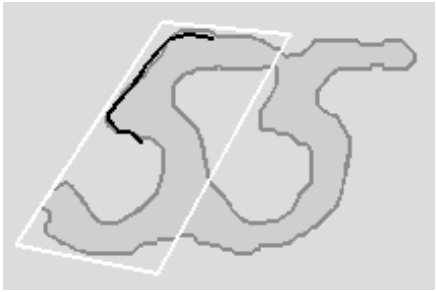


Рис. 1. Иллюстрация к распознаванию составных объектов.

Графическая демонстрация распознавания составного объекта «55» приведена на Рис. 1. Показано положение примитива на контуре составного объекта, дающее максимальное сходство и его габаритное описание, обеспечивающее формирование метрики двух дискретных объектов.

Работа выполнена при поддержке РФФИ, проект № 05-01-00590.

Критерии формирования примитивов и контрольных точек в структурном распознавании составных объектов

Васин Ю. Г., Лебедев Л. И.

lebedev@pmk.unn.ru

Нижний Новгород, НИИ прикладной математики и кибернетики
Нижегородского государственного университета им. Н. И. Лобачевского

В работе формулируются критерии формирования примитивов и контрольных точек при решении задачи распознавания составных объектов изображения с использованием предложенного в [1] подхода. Даются рекомендации по вопросам грамотного формирования примитивов и контрольных точек в целях оптимизации вычислительной сложности алгоритма распознавания составных объектов.

Введение

В предложенном подходе [1] структурное распознавание является предварительным и применяется для идентификации фрагментов контуров составного объекта в целях разбиения его на совокупность дискретных объектов по габаритным описаниям эталонов, примитивы которых обеспечивают заданный уровень сходства. Фрагментов контуров исследуемого составного объекта, сходных по форме с заданными примитивами и отмеченных контурным корреляционно-экстремальным ме-

тодом, может быть достаточно большое количество. В соответствии с общей концепцией распознавания составных объектов для каждого такого фрагмента должен быть по габаритному описанию эталона сформирован дискретный объект. Далее проводится его распознавание контурным корреляционно-экстремальным методом определения сходства разномасштабных форм. Таким образом, в зависимости от используемого набора примитивов, состав и численность формируемых и, следовательно, распознаваемых дискретных объектов может существенно меняться. Поэтому от выбора примитивов напрямую зависит быстродействие алгоритма распознавания составных объектов.

Постановка задачи и методы решения

Из вышесказанного следует, что решение задачи распознавания составных объектов на основе предложенного подхода должно предусматривать оптимизацию вычислительной сложности алгоритма. Вычисление оценок сходства с эталонами и примитивами контурными корреляционно-экстремальными методами уже является оптимизированной процедурой, поэтому, сложность алгоритма распознавания составных объектов зависит только от количества получаемых дискретных объектов. А для того, чтобы на контуре составного объекта было как меньше участков, сходных с примитивами, необходимо при их задании выбирать на эталонах фрагменты с уникальной формой. Это первый критерий формирования примитивов из группы Π_1 на основе эталонных дискретных объектов. Второе правило, которому необходимо следовать при формировании примитивов, состоит в том, чтобы по возможности исключить из их описаний проблемные участки контуров эталонов. Проблемными участками будем считать те фрагменты контуров дискретных объектов, в зоне которых вероятность слипания и образования составного объекта объективно выше. Например, у цифры «*5*», выполненной курсивом, это нижняя левая и верхняя правая области контурного описания эталона. Как правило, использование примитивов, содержащих проблемные участки, дает малый эффект при структурном распознавании, в то же время, значительно повышает вычислительную сложность алгоритма в целом. Это следует из того, что в зонах слипания контуров дискретных объектов конфигурация фрагментов коренным образом меняется, поэтому сходство их с примитивами будет отсутствовать, а следовательно, вхолостую будет осуществлена работа контурного корреляционно-экстремального метода с этим непроизводным элементом.

Очевидным фактом является то, что получить примитивы с уникальной формой, не содержащие проблемных участков, удается не для всех эталонов. Поэтому возможно формирование описаний непроизводных элементов на базе фрагментов контуров, идентичных у многих эталонов,

которые будут образовывать некоторые классы этих примитивов. Более того, такие фрагменты могут встречаться несколько раз и на одном и том же эталоне. Это примитивы группы Π_2. Примером примитива этой группы может служить фрагмент контура буквы «Н» шрифта Times New Roman, имеющий форму скобы «⊓» соответствующих размеров. Этому примитиву будет соответствовать класс эталонов, содержащий символы «Н», «П», «Ц», «Е», то есть $\mathbf{C}(\text{«⊓»}) = \{\text{«Н»}, \text{«П»}, \text{«Ц»}, \text{«Е»}\}$. При этом, при заданном пороговом уровне сходства, на символах «П», «Ц», «Е» будет определено контурным корреляционно-экстремальным методом по одному фрагменту, сходному с примитивом «⊓», а на букве «Н» — два участка (минимальный уровень сходства установливался по результатам масштабирования примитива на полпункта и вычисления соответствующего порогового коэффициента сходства). Дальнейшую идентификацию символов в классе $\mathbf{C}(\text{«⊓»})$ можно провести путем задания дополнительного набора примитивов, отражающих форму символов данного класса. Однако, можно решить эту задачу менее затратными процедурами, не привлекая к решению контурный корреляционно-экстремальный метод. Для этого определим котрольные точки эталонов. Первая группа контрольных точек определяет ориентацию эталона, а вторая — относительное местоположение. Для символов шрифта их естественной ориентацией является линия надписи, а точкой относительного местоположения — геометрический центр. Теперь, если габаритное описание эталонов и контрольных точек связано с местоположением фрагментов на эталонах, обеспечивающих заданный уровень сходства с примитивом, то идентифицировать символы в классах при известном направлении надписи достаточно просто. При неизвестном направлении надписи по соответствующим габаритным описаниям необходимо сформировать либо один дискретный объект, либо три с их дальнейшим распознаванием.

Следующий критерий формирования дискретных объектов — это принцип поглощения. Это означает, что если при заданных примитивах возможно по их габаритным описаниям формирование различных дискретных объектов с их последующим распознаванием, то предпочтение при всех прочих условиях необходимо отдавать тем эталонам, которые имеют бóльшие габариты. Это позволяет избежать рассыпания дискретных объектов. В качестве примера можно привести распознавание букв «Ш», «н», «п» на базе примитива в форме скобы, полученного из описания соответствующего фрагмента символа «н».

Полученные результаты и эксперимент

Предлагаемые принципы формирования примитивов, соответствующих габаритных описаний и контрольных точек в целях автоматизации получения решающих правил структурного распознавания и оптими-

зации вычислительной сложности алгоритма распознавания составных объектов были использованы для решения задачи распознавания отметок глубин на морских навигационных картах. Полученные результаты свидетельствуют об эффективности предлагаемых методов и методик распознавания составных объектов.

Работа выполнена при поддержке РФФИ, проект № 05-01-00590.

Литература

[1] *Васин Ю. Г., Лебедев Л. И.* Распознавание составных объектов изображения на базе структурного и корреляционно-экстремальных методов // ММРО-13 (в настоящем сборнике). — 2007. — С. 285–288.

Волоконно-оптический безлинзовый микроскоп

Власов Н. Г., Каленков Г. С., Каленков С. Г., Штанько А. Е.
vlasovng@rol.ru, kalenkov@mami.ru
Москва, МГТУ «МАМИ»

Описаны принципы построения безлинзовых цифровых микроскопов оптического диапазона.

Метод фазовых шагов первоначально был предложен для автоматизации интерференционных измерений на элементной базе современных цифровых средств обработки изображений, и позволил не только повысить информационную емкость измерений, но и в ряде случаев повысить их точность [1, 2].

Очень скоро стало понятно, что область применения метода значительно шире, чем предполагалось ранее. Так, еще в 1994 году в наших лекциях [3, 4] на Школе по когерентной оптике и голографии мы отметили, что в данном методе осуществляется цифровая запись волнового поля, и он является, таким образом, своеобразным компьютерным аналогом голографической записи, позволяющим осуществить все варианты голографической интерферометрии с компенсацией аберраций оптических элементов, с возможностью сравнения изделий по размерам и форме, причем информация об одном из них может задаваться математически и храниться в памяти компьютера. Было отмечено также, что на основе фазовых шагов можно обобщить известный метод фазового контраста Цернике, применявшийся в основном в микроскопии фазовых объектов, и использовать его для визуализации фазовых объектов с произвольным значением фазы. Метод фазовых шагов является также хорошей основой для решения фазовой проблемы [4, 5].

Перспективность применения голографии в микроскопии была отмечена еще Д. Габором, показавшим, что если записывать волновой фронт в электронных волнах, а восстанавливать в оптическом диапа-

зоне, то, с учетом геометрии схемы, можно получить увеличение $\cong 10^6$. Отсутствие когерентных источников надолго задержало реализацию идей Д. Габора, однако разработка высокоразрешающих ПЗС-матриц, содержащих несколько миллионов элементов (пикселей) сделало вполне реальным развитие голографии в видимом диапазоне [6], основанное на записи голограмм на ПЗС-матрицы и на цифровом восстановлении изображений.

Перспективность названного подхода обусловлена самой природой явления дифракции: чем меньше предмет, тем больше угловые размеры его дифракционной картины. Таким образом, в дальней зоне можно получать распакованную, то есть растянутую по пространственным координатам информацию об исследуемом объекте. При согласовании, на основе теоремы отсчетов, информационной емкости картины дифракции и разрешения ПЗС-матрицы можно, после ввода полученной голограммы в компьютер, восстановить в нем изображение и вывести на монитор с соответствующим увеличением.

Недостатком такой цифровой голографической регистрации является необходимость разрешения несущей пространственной частоты, образованной интерференцией объектного и опорного волновых полей. Она должна быть сравнительно высокой для того, чтобы при восстановлении пространственно отделить информативный плюс первый от других порядков дифракции [7]. От названного недостатка свободна запись пространственных распределений амплитуды и фазы волнового поля, выполненная на основе метода фазовых шагов. Перспективность такого подхода для реализации микроскопии видимого диапазона показана в [8]. В настоящей работе будет продемонстрировано, что безлинзовые микроскопы среднего увеличения могут быть созданы на базе недорогих цифровых фотоаппаратов, в которых, как ни парадоксально, излишним и даже мешающим элементом является их объектив.

Метод регистрации амплитуды и фазы волнового поля, использованный в настоящей работе, состоит в следующем. Пусть в плоскости приемной светочувствительной матрицы накладываются два поля. Первое — объектное, получено в результате дифракции волны на структуре микрообъекта, второе представляет собой опорную сферическую волну, центр кривизны которой расположен в пределах самого объекта. Результат интерференции двух волн регистрируется матрицей. Регистрацию производят несколько раз, меняя каждый раз фазу опорной волны. В данной работе использован алгоритм записи фазы волнового поля тремя последовательными экспозициями с изменением в промежутке между ними фазы опорной волны на 120°. Фазу $\varphi(x, y)$ объектного волнового поля

вычисляют с помощью простых соотношений [1]

$$\varphi(x,y) = \arctan[B(x,y)/(x,y)],$$

где

$$A(x,y) = 2I_1(x,y) - I_2(x,y) - I_3(x,y),$$
$$B(x,y) = [I_2(x,y) - I_3(x,y)],$$

I_1, I_2, I_3 — интенсивности, зарегистрированные каждым элементом приемной матрицы в ходе трех экспозиций. Дополнительно необходимо произвести проверку знака величины A.

$$\varphi = \begin{cases} \operatorname{arctg}[B/A], & \text{если } A > 0; \\ \operatorname{arctg}[B/A] + \pi, & \text{если } A < 0. \end{cases}$$

Для определения амплитуды объектного поля производят еще одну, четвертую экспозицию, в ходе которой регистрируют интенсивность I_4 одной лишь объектной волны (опорную волну перекрывают непрозрачным экраном). Распределение амплитуды объектного поля находят как $\sqrt{I_4(x,y)}$.

Описанный метод регистрации волнового поля был реализован с помощью устройства, основанного на обычных оптических элементах, входящих в комплект оптической скамьи [9, 10]. В настоящем докладе описан безлинзовый цифровой микроскоп на световодах. Устройство состоит из источника когерентного излучения, оптически сопряженной с ним системы световодов, формирующих две когерентные сферические волны, предметного столика и регистрирующей приемной матрицы. Источником излучения служит одномодовый полупроводниковый лазер ближнего ИК диапазона.

Система формирования двух когерентных сферических волн выполнена из одномодового световода 2 с разветвителем Y-типа. Излучающие торцы световодов располагаются у предметного стекла 3. Торец 4, пропущенный через отверстие в предметном стекле, излучает опорную волну. Излучение, исходящее из другого торца 5, расположенного под предметным стеклом, проходит через объект 6, лежащий на стекле 3, и образует объектную волну. Обе волны попадают на приемную матрицу 7, расположенную над предметным стеклом. Устройство снабжено юстирующими приспособлениями, позволяющими плавно изменять взаимное расположение излучающих торцов световодов. Одно из изображений, восстановленное с помощью обратного преобразования Фурье, приведено на Рис. 2

В работах [9–14] представлены результаты наших последних исследований по цифровой микроскопии и её применениям.

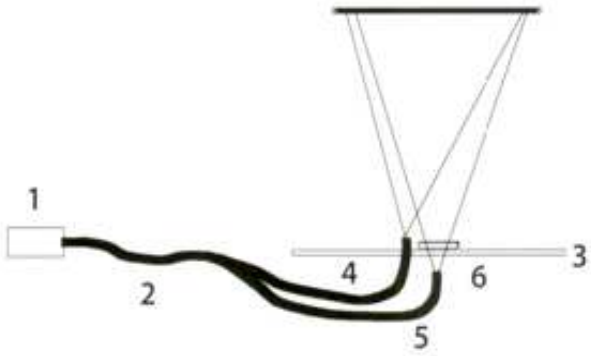


Рис. 1. Оптическая схема записи цифровых голограмм микрообъектов с использованием волоконно-оптических элементов.

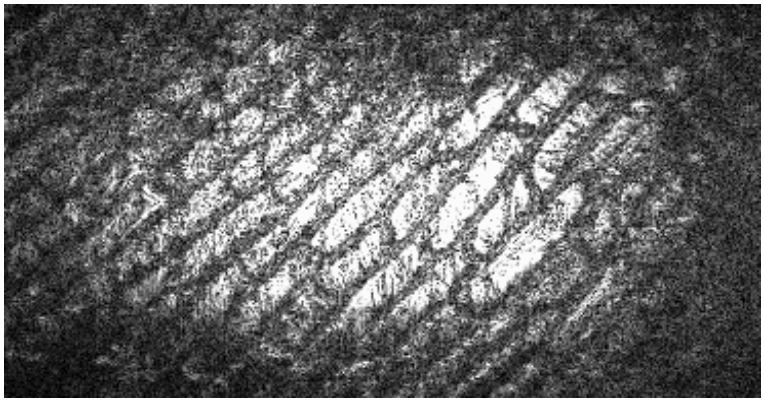


Рис. 2. Восстановленное изображение клеток лука.

Работа выполнена при поддержке РФФИ, проект № 06-07-89304.

Литература

[1] *Reid G. T.* Automatic Fringe Pattern Analysis // A Rewiew, Optics and Lasers Engineering — 1986. — V. 7. — Pp. 53–68.

[2] *Ryszard T.* Review of methods for automatic analysis of fringes in hologram interferometry // SPIE, Interferometric Metrology — 1987. — V. 816. — Pp. 140–148.

[3] *Власов Н. Г., Штанько А. Е.* Метод фазовых шагов // 23-я Школа по когерентной оптике и голографии— 1995. — С. 5–11.

[4] *Власов Н. Г., Каленков С. Г., Сажин А. В.* Решение фазовой проблемы на основе модифицированного метода фазового контраста и фазовых шагов // 23-я Школа по когерентной оптике и голографии— 1995. — С. 13–16.

[5] *Vlasov N. G., Kalenkov S. G., Sajin A. V.* Solution of phase problem // Laser phys.

[6] *Marron J. C., Schioeder K. C.* // Appl. Opt. — 1992. — V. 31. — P. 255.

[7] *Кольер Р., Беркхарт К.* Оптическая голография. — М.: Мир, 1973. — 686 с.

[8] *Yamaguchi I., Kato J., Ohta S., Mizuno J.* Image formation in phase-shiffing digital microscopy // Appl. Opt. — 2001. — V. 40. № 34. — Pp. 6177–6186.

[9] *Vlasov N. G., Kalenkov S. G., Krilov D. V., Shtanko A. E.* Non-lens Digial Microscopy // Proceedings of SPIE— 2005. — V. 40. — Pp. 158–163.

[10] *Каленков С. Г., Власов Н. Г., Крылов Д. В., Штанько А. Е.* Безлинзовая цифровая микроскопия // Естественные и технические науки — 2004. — Т. 3. № 12.— С. 117–120.

[11] *Власов Н. Г., Дугин В. В., Каленков С. Г.* Новый подход к улучшению продольного разрешения оптических систем // Сб. тр. Научн. сессии МИФИ — 2004. — № 4. — С. 240–242.

[12] *Власов Н. Г., Штанько А. Е.* Ассоциативная память с произвольной опорной волной // Сб. тр. Научн. сессии МИФИ — 2005. — Т. 4. — С. 240–241.

[13] *Власов Н. Г., Штанько А. Е., Воробьев С. П.* Способ голографической защиты от подделки малотиражных документов // Голография ЭКСПО-2005 — 2005. — С. 25–26.

[14] *Власов Н. Г., Каленков Г. С., Штанько А. Е.* Одномерный метод фазовых шагов // Научн. конф. МГТУ «Станкин» — 2005. — С. 229–231.

Представление полутоновых объектов с многоуровневым разрешением для ускоренного распознавания образов

Ганебных С. Н., Ланге М. М.

lange_mm@ccas.ru

Москва, ВЦ РАН

Соотношение эффективности и сложности является важной характеристикой большинства методов распознавания образов. Один из перспективных подходов к решению задачи распознавания в терминах соотношения «эффективность-сложность» базируется на применении представлений образов с многоуровневым разрешением [1, 3]. В настоящей работе предлагается способ построения пирамидальных представлений для широкого класса образов, заданных на изображениях двумерными объектами с неоднородной яркостной окраской. Приводятся оценки вычислительной сложности распознавания с использованием пирамидальных

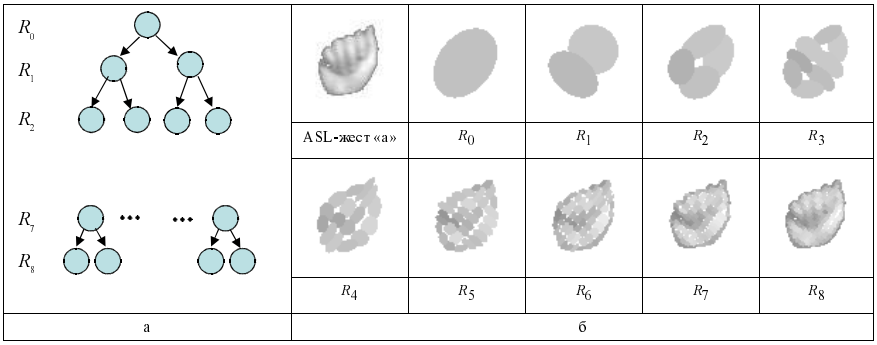


Рис. 1. Пирамидальное представление образа эллиптическими примитивами: а — структура представления; б — уровни представления ASL-жеста.

представлений и стратегии поиска решений по схеме последовательных приближений [2, 4]. Эффективность развитого подхода продемонстрирована экспериментальными результатами распознавания жестов языка ASL (American Sign Language).

Представление образов и база эталонов

В рассматриваемой модели образ определяется множеством пикселей $P = \{p_k \colon k = 1, \ldots, N\}$, яркости которых принимают значения $z_k = z(p_k) = 1, 2, \ldots, q$, а уровень $z = 0$ соответствует яркости фона изображения. В декартовых координатах (X, Y) любой образ P рассматривается как двумерное твердое тело (возможно многосвязное) с тензором инерции $\mathbf{G} = \left(\begin{smallmatrix} g_{yy} & -g_{xy} \\ -g_{yx} & g_{xx} \end{smallmatrix}\right)$, который образован ковариациями g_{xx} и g_{yy} и корреляционными моментами $g_{xy} = g_{yx}$, в виде соответствующих математических ожиданий по всем пикселям $p_k \in P$ с распределением «масс» $w(z_k) = z_k \Big(\sum_{k=1}^{N} z_k\Big)^{-1}$. Класс допустимых образов ограничен объектами, удовлетворяющими условию $(g_{xx} - g_{yy})^2 + 4 g_{xy} g_{yx} > 0$.

Пирамидальное представление R любого образа P из указанного класса строится посредством рекурсивного разбиения образа на сегменты (на два сегмента на каждом шаге рекурсии) и аппроксимации сегментов эллиптическими примитивами [3, 5]. В результате формируется пирамида представлений $R = \{R_l \colon l = 0, \ldots, L\}$, содержащая $L + 1$ уровней, в которой представление R_l l-го уровня содержит 2^l примитивов (Рис. 1а). Пример пирамиды представлений ASL-жеста, содержащей девять уровней ($L = 8$) дан на Рис. 1б.

Если заданы M пирамидальных представлений эталонных образов, где представления однотипных эталонов образуют семантические группы $\mathbf{R}_i^m = \{R_{ij}\colon j = 1, \ldots, m\},\ i = 1, \ldots, M/m$ (по m реализаций в каждой группе), то база эталонов строится на объединении всех семантических групп $\mathbf{R}^m = \bigcup_i \mathbf{R}_i^m$. Благодаря пирамидальной структуре представлений, база эталонов образует многослойную иерархическую структуру $\mathbf{R}^m = \{\hat{\mathbf{R}}_l^m\colon l = 0, \ldots, L\}$, в которой каждый l-ый слой $\hat{\mathbf{R}}_l^m$ содержит M эталонов, объединенных в M/m семантических групп, и каждый эталон слоя $\hat{\mathbf{R}}_l^m$ представлен 2^l примитивами.

Стратегия поиска решений

Решение о распознавании образа P по его представлению R основано на нахождении в базе эталонов $\mathbf{R}^m$ семантической группы $\mathbf{R}_s^m \in \mathbf{R}^m$, которая удовлетворяет заданному критерию и идентифицирует принадлежность образа к s-му классу. Критерий распознавания предполагает, что для любой пары $(R, \mathbf{R}_i^m)$ определена мера

$$D_i^{(l)}(R) = \frac{1}{m}\sum_{j=1}^{m} D^{(l)}(R, R_{ij}), \quad l = 0, 1, \ldots, L, \tag{1}$$

где $D^{(l)}(R, R_{ij}) \geqslant 0$ — заданная мера различия R и R_{ij}, вычисляемая по соответствующим представлениям l-го уровня. С учетом (1) критерий распознавания сводится к нахождению для предъявленного R группы эталонов $\mathbf{R}_s^m$, удовлетворяющей условию

$$D_s^{(L)}(R) = \min_i D_i^{(L)}(R) \leqslant D_s^*, \tag{2}$$

где $D_i^{(L)}(R)$ — мера отклонения представления R от группы $\mathbf{R}_i^m$, вычисляемая по L-му уровню с максимальным разрешением, а $D_s^* > 0$ — заданный допустимый порог отклонения для группы эталонов $\mathbf{R}_s^m$.

Поиск решения по критерию (2) основан на последовательном отборе семантических групп в слоях $\hat{\mathbf{R}}_0^m, \hat{\mathbf{R}}_1^m, \cdots, \hat{\mathbf{R}}_L^m$ базы $\mathbf{R}^m$ и нахождении ближайшей группы в слое $\hat{\mathbf{R}}_L^m$. Число отбираемых групп в слое $\hat{\mathbf{R}}_l^m$ определяется экспоненциально убывающей функцией с параметром α:

$$\mathcal{N}_l = \begin{cases} \mathcal{N}2^{-\alpha l}, & l = 0, \ldots, L, \quad \alpha \leqslant \frac{1}{L}\log\mathcal{N}; \\ N2^{-\alpha(l-L+k)}, & l = L-k, \ldots, L, \quad \alpha > \frac{1}{L}\log\mathcal{N}; \end{cases} \tag{3}$$

где $k = \lfloor \frac{1}{\alpha}\log\mathcal{N} \rfloor,\ \mathcal{N} = M/m$, а логарифм берется по основанию 2. В l-ом слое отбираются $\mathcal{N}_l$ групп, ближайших к R по мере (1), причем в $l+1$-ом слое отбор $\mathcal{N}_{l+1}$ ближайших групп ведется среди $\mathcal{N}_l$ групп, отобранных в l-ом слое.

Стратегия (3) позволяет оценить вычислительную сложность C_α поиска решения в терминах числа обрабатываемых примитивов в представлениях базы $\mathbf{R}^m$. Указанная сложность определяется соотношениями

$$C_\alpha = \begin{cases} M(2^{(1-\alpha)(L+1)} - 1)(2^{1-\alpha} - 1)^{-1}, & \alpha \leqslant \frac{1}{L}\log\frac{M}{m}, \\ M2^{L-k}(2^{(1-\alpha)(k+1)} - 1)(2^{1-\alpha} - 1)^{-1}, & \alpha > \frac{1}{L}\log\frac{M}{m}. \end{cases} \tag{4}$$

При большом числе эталонов ($M \to \infty$) асимптотические оценки сложности (4) имеют вид

$$C_\alpha \leqslant \begin{cases} O(M^{1/\alpha}), & \alpha < 1, \\ O(M\log M), & \alpha = 1, \\ O(M), & \alpha > 1. \end{cases} \tag{5}$$

Для сравнения вычислительная сложность поиска решения по критерию (2), полученного перебором всех семантических групп эталонов в L-ом слое $\hat{\mathbf{R}}_L^m$ базы $\mathbf{R}^m$, определяется величиной $C = M2^L$, которая при $L \leqslant \frac{1}{\alpha}\log\frac{M}{m}$ и $M \to \infty$ дает оценку $C \leqslant O(M^{1+1/\alpha})$.

Результаты распознавания ASL-жестов

Разработанный подход опробован для распознавания ASL жестов, доступных по адресу: `www.vision.auc.dk/~tbm/Gestures/database.html`. Исходные ASL жесты соответствуют 25-и буквам английского алфавита, в котором каждая буква задана 30-ю реализациями одного и того же жеста, образующими семантические классы. Эксперименты проводились с пирамидальными представлениями, содержащими $L+1 = 6, 7, 8, 9$ уровней разрешения и базами эталонов $\mathbf{R}^m$ с параметрами $m = 2, 3, 4$. Распознавание проводилось по критерию (2) с оптимизированными значениями пороговых коэффициентов D_s^* для каждой семантической группы эталонов. При различных значениях α вычислялась доля ложных решений ($\mathrm{P}_{\mathrm{false}}$) и доля отказов ($\mathrm{P}_{\mathrm{reject}}$), не удовлетворяющих критерию (2). Значения величин $\mathrm{P}_{\mathrm{false}}$ и $\mathrm{P}_{\mathrm{reject}}$ при $\alpha = 1$ представлены в таблице.

L		5	6	7	8
$\mathbf{R}^2$	$\mathrm{P}_{\mathrm{false}}$	0,005	0,000	0,005	0,011
	$\mathrm{P}_{\mathrm{reject}}$	0,112	0,101	0,066	0,059
$\mathbf{R}^3$	$\mathrm{P}_{\mathrm{false}}$	0,013	0,011	0,008	0,005
	$\mathrm{P}_{\mathrm{reject}}$	0,053	0,040	0,032	0,019
$\mathbf{R}^4$	$\mathrm{P}_{\mathrm{false}}$	0,005	0,000	0,005	0,000
	$\mathrm{P}_{\mathrm{reject}}$	0,027	0,032	0,016	0,016

Для указанных баз эталонов и уровней разрешения реальный вычислительный выигрыш $\gamma = C/C_{\alpha=1}$ лежит в диапазоне от 7 до 15 и увеличивается с ростом уровня максимального разрешения L.

Работа выполнена при поддержке РФФИ, проект № 06-01-00524.

Литература

[1] *Berretti S., Del Bimbo A.* Multiresolution spatial partitioning for shape representation // IEEE Proceedings of ICPR-2004, Cambridge, England: IAPR, 2004. — Vol. 2, — pp. 775–778.

[2] *Equitz W. E., Cover T. M.* Successive refinement of information // IEEE Transactions on Information Theory. — 1991. — Vol. 37, — pp. 269–275.

[3] *Lange M. M., Ganebnykh S. N.* Tree-like Data Structures for Effective Recognition of 2-D Solids // IEEE Proceedings of ICPR-2004, Cambridge, England: IAPR, 2004. — Vol. 1, — pp. 592–595.

[4] *Ланге М. М., Ганебных С. Н.* Многоуровневая структура данных и быстрый поиск на основе последовательных приближений // всеросс. конф. ММРО-12. — Москва, 2005. — Pp. 153–156.

[5] *Lange M. M., Ganebnykh S. N.* Moment-Based Pattern Representation Using Shape and Grayscale Features // Proc. of the Iberian Conf. on Patt. Recogn. and Image Analysis, IbPRIA-2007, Spain. — Springer, 2007. — Pp. 523–530.

Селекция аномальных ошибок установления соответствия в монокулярном режиме

Гришин В. А.

vgrishin@iki.rssi.ru

Москва, Институт космических исследований РАН

В системах технического зрения, предназначенных для использования в составе систем управления полетом беспилотных летательных аппаратов, находят применение монокулярные датчики оптического потока. Величина ошибок и устойчивость процесса измерений существенно зависят от эффективности селекции аномальных ошибок установления соответствия изображения одной и той же точки поверхности на разных кадрах [1, 2].

Уменьшение количества аномальных ошибок достигается как путем выбора точек изображения, обладающих хорошими свойствами привязки, так и выбором и оптимизацией метода установления соответствия. Тем не менее, полностью исключить аномальные ошибки установления соответствия не представляется возможным. Положение осложняется тем, что сетка отсчетов оптического потока является не регулярной и сильно разреженной, что связано с ограничениями на вычислительную мощность процессора, который должен осуществлять обработку изобра-

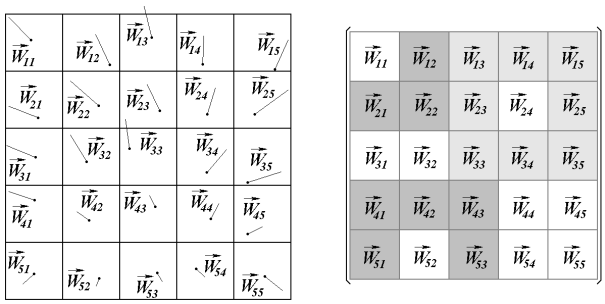


Рис. 1. Точки, по которым измеряется оптический поток

жений в реальном времени. Это существенно затрудняет использование классических методов анализа векторных полей.

Применение известных в статистике методов селекции аномальных ошибок [3, 4] осложняется рядом обстоятельств.

Во-первых, помимо нормальных и аномальных ошибок измерения имеется ярко выраженный тренд, которым, собственно, и является измеряемый оптический поток. Этот тренд должен быть исключен для применения статистических методов.

Во-вторых, как область определения функции потока, так и область ее значений является двумерными пространствами.

Для селекции аномальных ошибок предлагается анализировать структуру оптического потока по степени её локальной «согласованности». Предлагаемый алгоритм следует рассматривать как алгоритм предварительной селекции, позволяющий отбрасывать только грубые ошибки и требующий для своей реализации небольших ресурсов. Тем не менее, отбрасывание грубых ошибок существенно ускоряет сходимость алгоритмов оценивания линейных и угловых перемещений. Более тонкая селекция может осуществляться по величине невязок алгоритмов оценивания линейных и угловых координат, а также путем раздельной обработки безвихревой и соленоидальной компонент оптического потока, порождаемых, соответственно, линейными и угловыми перемещениями.

Алгоритм оценивания потока работает следующим образом.

Все поле зрения разбивается на $N \times K$ непересекающихся областей (см. Рис. 1). В каждой из областей осуществляется поиск точки, отличающейся наиболее хорошими свойствами привязки. Для каждой такой точки устанавливается соответствие на следующем кадре.

Рассмотрим наиболее простой случай определения векторов перемещения $\overrightarrow{W}_{i,j}$ только по двум соседним кадрам.

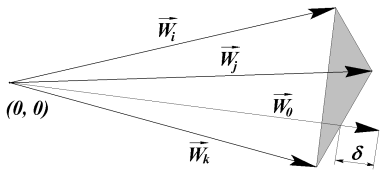


Рис. 2. Выпуклая оболочка трех векторов

Для каждой проверяемой точки формируется окрестность, которая в идеальном случае может содержать от 3 до 8 отсчетов оптического потока. На Рис. 1 справа выделены окрестности угловой точки $\overrightarrow{W}_{1,1}$, точки на границе $\overrightarrow{W}_{5,2}$ и внутренней точки $\overrightarrow{W}_{2,4}$. Если число отсчетов меньше 3, то такая точка отбрасывается как заведомо ненадежная, т. к. нет возможности проанализировать локальную структуру оптического потока. Из указанного количества отсчетов может быть сформировано от 1 до 56 «троек» векторов (число сочетаний из 8 по 3). Поскольку в окрестности проверяемой точки также могут находиться аномальные ошибки оценивания потока, то проверяется каждая тройка векторов. Если конец вектора потока в проверяемой точке попадает в выпуклую оболочку векторов потока какой-либо из троек $\overrightarrow{W}_i$, $\overrightarrow{W}_j$, $\overrightarrow{W}_k$, то считается, что аномальная ошибка определения потока в данной точке отсутствует, и процесс проверки троек прекращается. Если после обработки всех троек конец проверяемого вектора $\overrightarrow{W}_0$ не попал ни в одну из них, то величина δ (Рис. 2) сравнивается с порогом. Если δ не превосходит порога, то считается, что в проверяемой точке нет аномальной ошибки. Величина порога зависит от степени разреженности сетки отсчетов, профиля поверхности, характера движения объекта, и определяется экспериментально, либо адаптивно подстраивается в процессе работы системы технического зрения.

Результаты моделирования работы алгоритма представлены в Таблице 1. Средняя длина векторов оптического потока менялась от 40 до 101 пикселя (в зависимости от пары кадров). Величина порога была установлена равной 10 пикселям. Для данной величины порога доля неправильно обнаруживаемых аномальных ошибок составила $10,6\%$. В основном, в данную категорию попали угловые точки.

Работа выполнена при поддержке РФФИ, проекты №06-08-01497-а и №06-01-00524-а.

Величина аномальных ошибок в пикселях	20	30	40	50
Правильно обнаружено аномальных ошибок	9	15	24	24
Пропущено аномальных ошибок	15	9	0	0

Таблица 1. Результаты моделирования процесса селекции.

Литература

[1] *Johnson A. E., Matthies L. H.* Precise Image-Based Motion Estimation for Autonomous Small Body Exploration // 5th Int. Symp. on Artificial Intelligence, Robotics and Automation in Space (iSAIRAS'99), 1999. — Pp. 627–634.

[2] *Гришин В. А.* Системы технического зрения в решении задач навигации и терминального управления // Космическое приборостроение. Программа и тезисы докладов выездного семинара. Россия. Таруса. 7–9 июня 2006, Москва: Институт космических исследований РАН, 2006.— С. 22–23.

[3] *Кобзарь А. И.* Прикладная математическая статистика. Для инженеров и научных работников. — Москва: ФИЗМАТЛИТ, 2006. — 816 с.

[4] *Крянев А. В., Лукин Г. В.* Математические методы обработки неопределенных данных. — Москва: ФИЗМАТЛИТ, 2003. — 216 с.

Эффективное распознавание взаимосвязанных объектов на основе ациклических марковских моделей

Двоенко С. Д., Савенков Д. С.

dsd@uic.tula.ru, DenisSavenkov@home.tula.net

Тула, Тульский государственный университет

Эффективная процедура распознавания основана на представлении решетчатого графа соседства взаимосвязанных объектов набором ациклических графов. Потери от сужения множества взаимосвязей между объектами компенсируются путём расширения множества ациклических графов.

Древовидная марковская модель

Массив T взаимосвязанных объектов $t \in T$ представлен как случайное поле (X, Y) со скрытой компонентой $X = (x_t, t \in T)$ значений классов $x_t \in \mathcal{X} = \{1, \ldots, m\}$ и наблюдаемой компонентой $Y = (\mathbf{y}_t, t \in T)$. Взаимосвязанность объектов выражена неориентированным графом G без петель, ребра которого соединяют соседние элементы массива. Предположение об условной независимости наблюдений относительно реализации скрытого поля классов $\psi_t(\mathbf{y}_t|X) = \psi_t(\mathbf{y}_t|x_t)$ позволяет на этапе обучения остаться в рамках классической теории распознавания образов [1, 2]. Необходимо восстановить скрытое поле классов X для реализации поля Y, например, на основе байесовского решающего правила

$\hat{X} = \arg\max\limits_{X \in \mathcal{X}^{|T|}} p(X|Y)$. Применение таких моделей для зависимых наблюдений с произвольным характером соседства приводит к трудоемким процедурам типа simulating annealing, например, при обработке растровых текстурных изображений [3, 4, 5].

Восстановим скрытое поле классов X на основе другого байесовского правила в виде $\hat{X} = (\hat{x}_t, t \in T)$, $\hat{x}_t = \arg\max\limits_{x_t \in \mathcal{X}} p_t(x_t|Y)$. Для древовидного (т. е. ациклического) графа G построен эффективный алгоритм распознавания [2]. Марковское свойство поля X и древовидность графа G позволяют перейти от полученных на этапе независимого обучения апостериорных маргинальных распределений $p_t(x_t|\mathbf{y}_t)$ к апостериорным маргинальным распределениям $p_t(x_t|Y)$, $t \in T$. В [1] показано, что априорное поле X является односторонним марковским $q_t(x_t|X_{(t)}) = q_t(x_t|x_r)$, где $X_{(t)}$ — скрытое поле X без элемента x_t, а t является потомком вершины r относительно дерева G. Апостериорное поле X остается односторонним марковским с тем же графом G и условными распределениями $p_t(x_t|X_{(t)}, Y) = p_t(x_t|x_r, Y_t^+)$, где часть Y_t^+ поля Y образует поддерево с корнем в $\mathbf{y}_t$. Распознавание классов X выполняется всего за два просмотра дерева G. Распределения $p_t(x_t|Y_t^+)$ вычисляются при восходящем просмотре, начиная с терминалов, где $p_t(x_t|Y_t^+) = p_t(x_t|\mathbf{y}_t)$, и завершая в корне, где $p_t(x_t|Y_t^+) = p_t(x_t|Y)$ для всего дерева. Распределения $p_t(x_t|Y)$ вычисляются при нисходящем просмотре из корня [2].

Представление решетки ациклическими графами

Очевидно, что произвольный граф соседства G нельзя заменить древовидным без потери его фундаментального свойства нести полную информацию о положении каждого элемента массива T относительно других его элементов. Решетка представляет отношение соседства на растре и не является ациклической структурой. Корректная аппроксимация исходного графа соседства обычно требует разработки специального алгоритма, сопоставимого по сложности с алгоритмом распознавания [6].

Предлагается заменить исходный граф соседства элементов массива набором ациклических графов для компенсации редуцированного множества взаимосвязей в древовидном графе расширенным множеством различных ациклических графов. Предполагается, что достаточно сохранять только локальные взаимосвязи элементов массива. Например, для текстурного изображения удобно использовать графы как на Рис. 1.

Итерационный алгоритм распознавания скрытого поля X построим следующим образом. Зададим ациклический граф из набора на Рис. 1, и однократным применением алгоритма из [2] перейдем от распределений $p_t(x_t|\mathbf{y}_t)$ к апостериорным распределениям $p_t(x_t|Y)$, соответствующим реализации значений $x_t, t \in T$ при наблюдении поля Y. Далее, выбе-

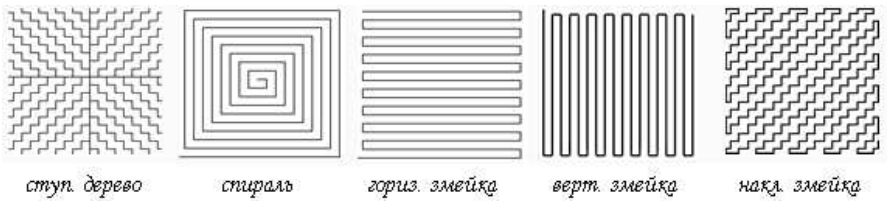


Рис. 1. Ациклические графы

рем другой граф из заданного набора и вновь применим алгоритм из [2], рассмотрев вместо распределений $p_t(x_t|\mathbf{y}_t)$ только что полученные апостериорные распределения $p_t(x_t|Y)$, и перейдем от них к новым апостериорным распределениям, которые снова обозначим как $p_t(x_t|Y)$.

С другой стороны, однократное применение алгоритма из [2] сформирует для каждого ациклического графа свое множество апостериорных распределений $p_t(x_t|Y)$, $t \in T$, и решений о классах $\hat{x}_t(Y)$. Согласно [7], окончательное решение комбинируется для каждого объекта $t \in T$ как среднее распределений $p_t(x_t|Y)$, полученных для каждого графа G.

Предлагается три итерационных алгоритма распознавания: повторение одного графа, чередование графов, повторение комбинирования. Эксперименты показывают, что распределения $p_t(x_t|Y)$ быстро стабилизируются. На Рис. 2 два цвета смешаны пропорционально маргинальным вероятностям классов. Для растра 201×201 показаны: а — классификация учителя, б — независимое распознавание (13548 ошибок), в — повторение ступенчатого дерева дважды (2806 ошибок), г — однократное чередование всех графов (987 ошибок), д — повторение комбинации всех графов дважды (763 ошибки).

Работа поддержана РФФИ, № 06-01-00412, и INTAS, № 04-77-7347.

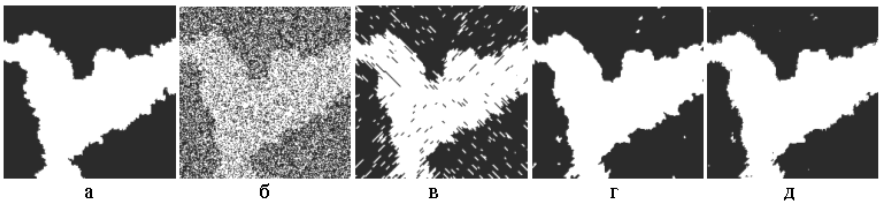


Рис. 2. Распознавание двух классов

Литература

[1] *Двоенко С. Д., Копылов А. В., Моттль В. В.* Задача распознавания образов в массивах взаимосвязанных объектов. Постановка задачи и основные предположения // Автоматика и телемеханика. — 2004. — № 1. — С. 143–158.

[2] *Двоенко С. Д., Копылов А. В., Моттль В. В.* Задача распознавания образов в массивах взаимосвязанных объектов. Алгоритм распознавания // Автоматика и телемеханика. — 2005. — № 12. — С. 162–176.

[3] *Лебедев Д. С., Безрук А. А., Новиков В. М.* Марковская вероятностная модель изображения и рисунка. Препринт. — М.: ИППИ АН СССР, 1983.

[4] *Geman S, Geman D.* Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images // IEEE Trans. on PAMI. — 1984. — V. 6. — Pp. 721–741.

[5] *Li S. Z.* Markov Random Field Modelling in Computer Vision. — Springer-Verlag, 1995.

[6] *Mottl V. V., Dvoenko S. D., Levyant V. B., Muchnik I.B.* Pattern recognition in spatial data: a new method of seismic explorations for oil and gas in crystalline basement rocks // Proc. of 15th ICPR. — 2000. — V. 3. — Pp. 210–213.

[7] *Kittler J, Hatef E., Duin R. P.* Combining classifiers // Proc. of 13th ICPR. — 1996. — V. 2. — Pp. 897–901.

Метод автоматического кадрирования цифровых портретных изображений

Дегтярев С. В., Мирошниченко С. Ю.

ser@vt.kstu.kursk.ru

Курск, Курский государственный технический университет

В настоящее время в связи с бурным развитием технологий ПЗС-фотоприемных матриц, специализированных процессоров и устройств для обработки изображений широкое распространение получила цифровая фототехника. Сегодня цифровые фотоаппараты уже не являются роскошью, доступной узкому кругу профессиональных фотографов. На рынке представлено множество моделей недорогих фотоаппаратов ведущих мировых фирм-производителей: Sony, Kodak, Olympus, Nikon, Samsung и др. Качество оптики, разрешение и физические размеры ПЗС матриц, стоимость фотоаппаратов различных фирм примерно одинаковы, поэтому основной акцент при выборе конкретной модели делается на функциональные возможности и, в том числе, на возможности обработки получаемых изображений.

Одной из наиболее востребованных функций является автоматическое кадрирование портретных изображений с центрированием лица человека в плоскости кадра. Существующие методы выделения лица под-

разделяются на две группы: методы сегментации изображения в различных цветовых пространствах [1] и методы оценки положения лица по наиболее информативной его части — глазам. Методы первой группы обладают невысокой точностью при наличии на изображении сложного фона, тогда как методы второй группы имеют высокую сложность за счет рекурсивной организации вычислений.

Разработан метод автоматического кадрирования цифрового портретного изображения, основанный на распознавании контуров глаз. Контурное описание изображения строится с помощью метода Канни [2], который предполагает свертку исходного изображения с маской КИХ-фильтра, представляющего дискретную аппроксимацию первой производной фильтра Гаусса. Полученное таким образом градиентное изображение подвергается скелетизации и бинаризации [3] и формируется изображение, описывающее положение контурных линий объектов. На основании контурного изображения строится векторное описание [4], каждый элемент которого представляет контур объекта.

Сформированное контурное описание подвергается предварительной классификации для снижения трудоемкости процесса распознавания за счет исключения из рассмотрения заведомо неинформативных контуров.

Для предварительной классификации используются следующие признаки:

— отношение сторон описанного вокруг контура прямоугольника;
— относительные линейные размеры контура;
— относительное расстояние центра масс контура от центра описанного прямоугольника, позволяющее идентифицировать асимметричные контуры.

Те контуры, значение хотя бы одного из указанных параметров которых превышает априорно установленные пределы, исключаются из контурного описания.

По окончании предварительной классификации выполняется распознавание, целью которого является поиск контуров, соответствующих открытым глазам человека. При распознавании применяются следующие признаки [5]:

— среднее относительное расстояние от центра масс контура до каждой его точки;
— относительный смещенный момент инерции;
— относительная плотность контура;
— сумма частот верхней части гистограммы участка исходного изображения, ограниченного описанным вокруг контура прямоугольником.

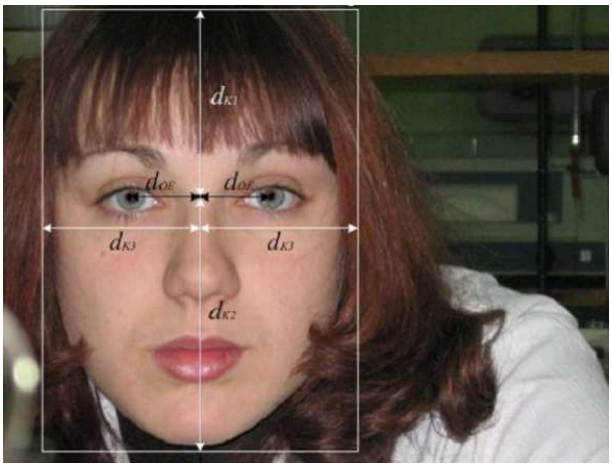


Рис. 1. Определение границ кадрированного изображения.

Перечисленные выше признаки не чувствительны к повороту и масштабированию распознаваемого объекта (однако, линейные размеры глаза должны составлять не менее 40 точек), что обеспечивает распознавание лица при любом его положении в плоскости изображения.

В результате распознавания в контурном описании сохраняются только контуры, соответствующие открытым глазам.

В том случае, если исходное изображение содержит лицо одного человека, кадрирование выполняется следующим образом:

— определяются центры масс контуров, описывающих глаза;
— между центрами масс проводится отрезок, середина которого лежит на оси симметрии лица;
— вычисляется расстояние d_{OE} от глаза до оси симметрии (Рис. 1);
— с помощью d_{OE} рассчитываются размеры кадрированного изображения d_{K1}, d_{K2}, d_{K3} (Рис. 1) на основании априорно определенных зависимостей.

Полученное кадрированное изображение масштабируется до размеров исходного с помощью метода бикубической интерполяции.

Разработанный метод позволяет автоматически кадрировать цифровое портретное изображение вне зависимости от положения лица человека в плоскости кадра и обладает низкой вычислительной сложностью, что позволяет встраивать его непосредственно в цифровые фотоаппараты.

Литература

[1] *Kukharev G., Nowosielski A.* Fast and Efficient Algorithm for Face Detection in Color Images // Machine Graphics and Vision. — Vol. 13, No. 4. — 2004. — Pp. 377–397.

[2] *Canny J. F.* A Computational Approach to Edge Detection // IEEE Trans. Patt. Recogn. and Mach. Intel. — Vol. 8, No. 6, 1986. — Pp.679–698.

[3] *Дегтярев С. В., Садыков С. С., Тевс С. С., Ширабакина Т. А.* Методы цифровой обработки изображений. — Курск: КурскГТУ, 2001. — 167с.

[4] *Фурман Я. А.* Введение в контурный анализ и его приложения к обработке сигналов и изображений. — М.: ФИЗМАТЛИТ, 2002. — 592с.

[5] *Стулов Н. Н.* Способ формирования признаков объектов в СТЗ, инвариантных к повороту, переносу и изменению масштаба // Сист. и мет. обраб. и анализа информ. — М.: Горячая линия–Телеком, 2005. — С. 18–24.

Алгоритмы параметрической идентификации сигнала, использующие обобщенный спектрально-аналитический метод

Долотова Н. С.

ndolotova@rambler.ru

Москва, МГУ им. М. В. Ломоносова, факультет ВМиК

В работе решается задача выявления в сигналах закономерностей достаточно общего вида. Предполагается, что сигнал задан последовательностью $f(t_i)$, где t_i — моменты времени, в которые производились измерения. Ставится задача проверить соответствие сигнала дифференциальному уравнению

$$F\big(\alpha_0, \ldots, \alpha_n, f(t), f'(t), \ldots, f^{(n)}(t)\big) = G\big(\beta_1, \ldots, \beta_n, f'(t), \ldots, f^{(n)}\big), \quad (1)$$

где F, G — заданные функции, $\alpha_0, \ldots, \alpha_n, \beta_1, \ldots, \beta_n$ — параметры модели, которые требуется идентифицировать по имеющимся данным $\{f(t_i)\}$.

Алгоритм параметрической идентификации сигнала

Для решения поставленной задачи предлагается следующий алгоритм, основанный на обобщенном спектрально-аналитическом методе, ОСАМ [1].

1. **Аппроксимация полученного сигнала сплайнами.** При аппроксимации происходит переход от дискретной функции $f(t_i)$ к ее непрерывному аналогу $f(t)$, что дает возможность определить значения функции в узлах гауссовой сетки.
2. **Разложение сигнала в ряд по полиномам Чебышева I рода.** Полученная непрерывная функция $f(t)$ раскладывается в ряд по по-

линомам Чебышева I рода: $f(t) \approx \sum_{i=0}^{N} C_n T_n(t)$, где C_n — коэффициенты разложения в ряд функции $f(t)$. Коэффициенты вычисляются с помощью квадратурной формулы Гаусса по имеющимся значениям функции в узлах гауссовой сетки, вычисленным в п. 1. В дальнейшем почти все вычисления производятся в пространстве коэффициентов разложения.

3. **Выбор параметрического уравнения, описывающего закономерность.** Так как любая непрерывная функция может быть описана некоторым дифференциальным уравнением, то на основании соответствия (не соответствия) данных уравнению, описывающему закономерность, можно делать вывод о том, отвечает или нет сигнал выбранному закону (1).
4. **Подбор оптимальных для сигнала параметров уравнения.** На данном этапе осуществляется подбор оптимальных параметров уравнения. Поиск осуществляется методом градиентного спуска, максимизируя коэффициент корреляции левой и правой частей уравнения (1). Использование OCAM позволяет все вычисления производить в пространстве коэффициентов, что снижает погрешности, возникающие в результате перехода из пространства коэффициентов в пространство значений функции в точках.
5. **Определение наличия закономерности в сигнале.** Для определения присутствия закономерности в сигнале вычисляется коэффициент корреляции левой и правой частей уравнения 1 для подобранных в п. 4 параметров. В зависимости от полученного значения корреляции принимается одно из трёх решений:
   - $\mathrm{cor} > 0.7$ — закономерность присутствует в сигнале;
   - $\mathrm{cor} < 0.4$ — закономерность в сигнале отсутствует;
   - иначе не дается однозначного ответа о наличии или отсутствии закономерности в сигнале.

Основное отличие предлагаемого алгоритма параметрической идентификации сигнала заключается в том, что использование OCAM позволяет значительно упростить вычисления, возникающие на этапе подбора оптимальных параметров, а также обеспечивается устойчивость работы с сильно зашумлёнными сигналами.

Примеры работы алгоритма

Предложенный алгоритм был применен для выявления в сигналах закономерностей, имеющих вид гармонических колебаний или гармонических колебаний с линейной частотной модуляцией, описываемых параметрическим семейством уравнений

$$2\alpha f(t) = (2\alpha t + \beta) f''(t) + (2\alpha t + \beta)^2 f'(t).$$

Рис. 1. Пример работы алгоритма для модели гармонических колебаний с линейной частотной модуляцией.

Рис. 2. Прогнозирование роста численности населения г. Москвы.

Для проверки корректности работы алгоритма использовались искусственно синтезированные данные. Результаты решения задачи предложенным способом показаны на Рис. 1.

Также предлагаемый алгоритм применялся для прогнозирования роста населения г. Москвы. В XIX в. Р. Ф. Ферхьюстом было сделано предположение о том, что численность населения изменяется по логистическому закону $\alpha f(t) = f'(t) + \beta f^2(t)$. Результат параметрической идентификации по данным переписей населения показан на Рис. 2. Оказалось, что при сохранении существующей тенденции и справедливости гипотезы о логистическом росте максимальная численность населения г. Москвы (23 млн. чел.) будет достигнута примерно через 250 лет.

Литература

[1] *Дедус Ф. Ф., Куликова Л. И., Панкратов А. Н., Тетуев Р. К.* Классические ортогональные базисы в задачах аналитического описания и обработки информационных сигналов. — Москва, 2004.

[2] *Суетин П. К.* Классические ортогональные многочлены. — Москва: Наука, 1979.

[3] *Самарский А. А., Николаев Е. С.* Методы решения сеточных уравнений. — Москва: Наука, 1978.

Об одном методе сегментации растровых объектов для задач преобразования формы

Домахина Л. Г.

ludmila.domakhina@gmail.com

Москва, МГУ ВМиК, кафедра Математических методов прогонозирования

Сегментация цифровых изображений является важной задачей современных систем машинного зрения и искусственного интеллекта. Под *сегментацией растровых объектов* будем понимать разбиение области, аппроксимирующей растровый объект, на множество подобластей.

Скелетом многоугольной фигуры [1] называется множество центров максимальных вписанных в нее окружностей. Скелет представляет собой планарный граф, *вершинами* которого являются центры окружностей, касающихся границы в трёх и более точках, а *ребрами* — серединные оси, линии, состоящие из центров окружностей, касающихся границы в двух и более точках. *Гранично-скелетное представление растрового изображения* — это скелет аппроксимирующей фигуры вместе с множеством всех вписанных пустых кругов.

Предлагается метод сегментации растрового объекта, основанный на построении его гранично-скелетного представления. В скелетном графе растрового объекта каждая вершина скелетного графа, кроме листовых, соединяется с ближайшей точкой границы аппроксимирующей фигуры так называемыми *радиальными отрезками*. Минимальное подмножество точек исходного объекта, ограниченное ребром скелетного графа и соответвующими радиальными отрезками, назовем *собственной областью ребра* (Рис. 1).

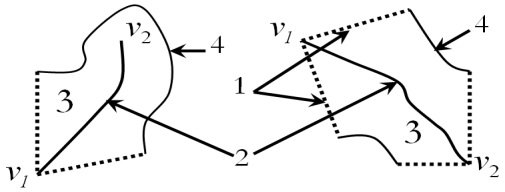


Рис. 1. Собственные области ребер скелета: 1 — радиальные отрезки; 2 — ребро скелета v_1v_2; 3 — собственная область, соответствующая v_1v_2; 4 — граница аппроксимирующей фигуры.

Алгоритм 1. Алгоритм сегментации растрового объекта на основе его гранично-скелетного представления.

Вход: $I_j(x,y) \subseteq I(x,y)$ — растровый объект на изображении $I(x,y)$;
Выход: $Seg_j(x,y)$ — сегментация $I_j(x,y)$;
1: построение $B_j(x,y)$ — граничной аппроксимации фигуры $I_j(x,y)$;
2: построение скелета $Sk_j(x,y)$ фигуры $B_j(x,y)$, n — число его ребер;
3: **для** всех вершин скелета v_k, таких, что $\deg(v_k) \geqslant 3$
4: построить радиальные отрезки из v_k к $B_j(x,y)$
(их число равно $\deg(v_k)$);
5: **для** всех ребер скелета $v_i v_{i+1}$ по всем $i = 0, \ldots, n$
6: найти собственную область $\texttt{SubArea}(v_i v_{i+1})$;
7: $Seg_j(x,y) = \bigcup_{i=0}^{n-1} \texttt{SubArea}(v_i v_{i+1})$ — искомая сегментация $I_j(x,y)$.

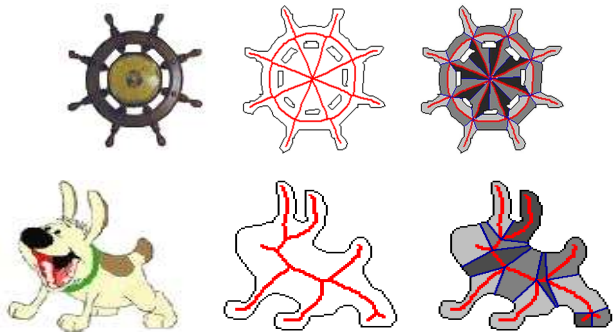


Рис. 2. Сегментация растровых объектов: объект — гранично-скелетное представление — скелетная сегментация.

Теорема 1. *Собственные области двух любых различных ребер скелетного графа растрового объекта не пересекаются.*

Теорема 2. *Объединение собственных областей всех ребер скелетного графа растрового объекта дает весь объект.*

В силу теорем 1 и 2 описанный метод дает разбиение объекта на собственные области, которое и будем считать искомой *скелетной сегментацией* (Рис. 2), а собственные области — его сегментами.

Предложенный метод сегментации может быть применен для решения задач преобразования формы растровых изображений.

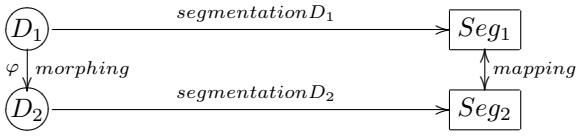
Назовем два растровых объекта *содержательно близкими*, если их скелетные графы изоморфны. Рассмотрим следующие задачи.

Задача 1. Дано: содержательно близкие растровые объекты D_1 и D_2. Необходимо: раскрасить D_2 в соответствии с раскраской D_1.

Задача 2. Дано: содержательно близкие растровые объекты D_1 и D_2. Найти: гомеоморфное преобразование объекта D_1 в D_2.

Задача 2 может быть решена с помощью построения двух промежуточных скелетных сегментаций объектов D_1 и D_2. Основная идея решения заключается в том, что изоморфные сегментации можно непрерывно и взаимно однозначно отобразить друг на друга — построить гомеоморфизм $\varphi\colon D_1 \to D_2$. При этом каждый сегмент прообраза D_1 необходимо отобразить на соответствующий по изоморфизму сегмент образа D_2.

Для решения задачи 1 построим гомеоморфное преобразование φ фигур D_1 в D_2 как решение задачи 2. Тогда цвета точек растра $x \in D_2$ можно интерполировать с помощью цветов соответствующих точек прообраза $\varphi^{-1}(x)$.



Таким образом, с помощью скелетной сегментации можно решать задачи преобразования формы растровых объектов. В практических приложениях использование скелета как множества всех срединных осей сильно сужает класс решаемых задач. Часто используются различные «стрижки» скелета, например, выделение фундаментальной части — базового скелета [3]. В работе [2] описан метод гомеоморфного отображения односвязных объектов с изоморфными базовыми скелетами, который является частным случаем задачи 2 для односвязных объектов.

Работа выполнена при поддержке РФФИ, проект № 05-01-00542.

Литература

[1] *Местецкий Л. М.* Скелет многосвязной многоугольной фигуры. // Труды 15 междунар. конф. ГРАФИКОН-2005, Новосибирск — С. 242–249.

[2] *Местецкий Л. М., Петрова Л. Г.* Расчет гомеоморфизма односвязных многоугольных областей с изоморфными базовыми скелетами. // Искусственный интеллект. — Симферополь, 2006.

[3] *Местецкий Л. М., Рейер И. А.* Непрерывное скелетное представление изображения с контролируемой точностью. // Труды 13 междунар. конф. ГРАФИКОН-2003, Москва — С. 246–249.

Сравнение 3D портретов при распознавании лиц

Дышкант Н. Ф., Местецкий Л. М.
nfd3001@gmail.com, L.Mest@ru.net
Москва, МГУ им. М. В. Ломоносова

Современные системы машинного зрения позволяют в реальном времени (4–5 кадров в секунду) получать трехмерные портреты человеческих лиц в виде облака точек. Задача их анализа при биометрической идентификации требует введения метрики для сравнения портретов.

Поверхность лица представляет собой однозначную функцию двух переменных $z = F(x, y)$ в некоторой системе координат, в которой ось z направлена вдоль оси визирования, Рис. 1. Тогда метрика для измерения сходства поверхностей основывается на сравнении функций двух переменных, заданных на нерегулярных сетках.

Пусть в точках сетки G_1 задана функция F_1, в точках сетки G_2 — функция F_2. Предлагаемое решение задачи включает в себя следующие этапы (см. Рис. 2):

1. Построение триангуляций Делоне на каждой из сеток G_1, G_2;
2. Построение минимальных остовных деревьев этих триангуляций;
3. Локализация точек каждой из триангуляций в треугольниках другой;
4. Интерполяция значений функции F_1 в точках сетки G_2, интерполяция значений функции F_2 в точках сетки G_1;
5. Построение общей триангуляции (на сетке $G_1 + G_2$);
6. Сравнение функций на отдельных ячейках общей сетки.

Сравнение трехмерных портретов

Для построения триангуляций используется алгоритм, в основе которого лежит парадигма рекурсивной декомпозиции «разделяй и властвуй», вычислительная сложность которого $O(N \log N)$.

Для построения минимальных остовных деревьев (МОД) используется алгоритм Черитона и Тарьяна, предложенный в [1], который позволяет построить МОД на основе графа Делоне исходного множества за *линейное* время. При реализации алгоритма используется структура данных фибоначчиева куча, предложенная Фредманом и Тарьяном в [2].

Предлагаемый метод решения задачи локализации двумерной сетки в триангуляции использует МОД, вершинами которого являются узлы данной сетки. Тогда задача сводится к задаче локализации точки в треугольниках триангуляции.

После завершения этапа локализации каждая точка сетки локализована в некотором треугольнике другой сетки, и можно рассмотреть задачу интерполяции функции, заданной в точках одной сетки, в точках другой сетки.

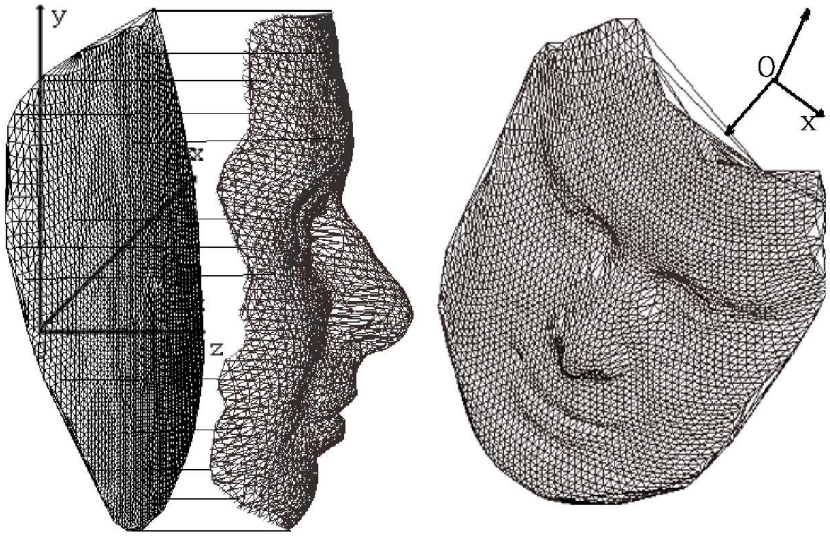


Рис. 1. Триангулированная поверхность лица.

Проведеные эксперименты показали, что вычислительная сложность этапов локализации и интерполяции составляет в среднем $O(N)$.

После интерполяции в каждой точке обеих сеток известны значения двух функций. На узлах обеих сеток строится общая триангуляция Делоне. Тогда в каждом узле общей триангуляции заданы значения двух функций, и можно вычислить меру различия между двумя 3D портретами. Для каждого треугольника считается объем разности между двумя функциями, заданными в его вершинах, затем все результаты суммируются.

Полученные результаты

В Таблице 1 приведены расходы времени для различных этапов алгоритма. Вычислительные эксперименты проводились на машине с процессором AMD Athlon 2 600+ и оперативной памятью 512 Мб для моделей, состоящих из 3 000 точек. В том случае, если одна модель лица находится в базе, общее время уменьшается вдвое. Кроме того, это время можно уменьшить за счет возможности распараллеливания, использования более быстрого процессора и построения общей триангуляции алгоритмом слияния, описанным в [3], с использованием локализации.

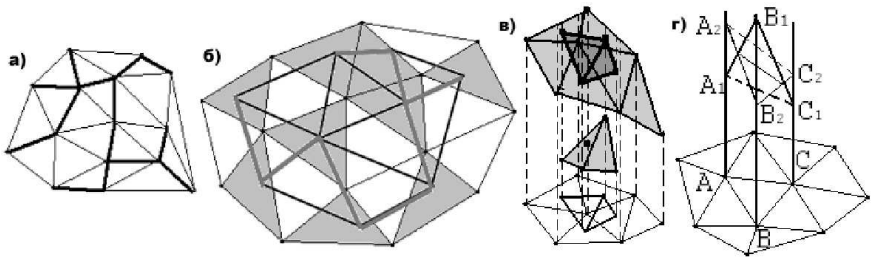


Рис. 2. Основные этапы алгоритма: *а)* триангуляция Делоне и минимальное остовное дерево; *б)* локализация с помощью МОД; *в)* интерполяция; *г)* вычисление объема разности на отдельных ячейках.

Этап алгоритма	**Время (сек)**
Построение триангуляции G_1 и G_2	0,124
Построение МОД G_1 и G_2	0,203
Локализация триангуляциий	0,015
Интерполяция функций	$<$0,001
Построение общей триангуляции	0,109
Вычисление $\int \mid F_1 - F_2 \mid$	0,031
Общее время	**0,497**

Таблица 1. Расход времени для этапов алгоритма.

Алгоритм сравнения оказывается эффективным, что позволяет искать положение, при котором маски совпадают наилучшим образом.

Предложенный метод позволяет эффективно сравнивать портреты, допускает распараллеливание и может быть использован в системах реального времени.

Литература

[1] *Препарата Ф., Шеймос М.* Вычислительная геометрия. — Москва: Мир, 1989. — 314 с.

[2] *Tarjan R. E., Fredman M. L.* Data Structures and Network Algorithms // Society for Industrial and Applied Mathematics. — 1983.

[3] *Местецкий Л. М., Царик Е. В.* Триангуляция Делоне: рекурсия без пространственного разделения точек. // Труды 14 международной конференции ГРАФИКОН, Москва, 2004. — С. 267–270.

Выбор базиса в задаче беспризнакового распознавания личности по фотопортрету

Ермаков А. С.
yermakov@uic.tula.ru
Тула, Тульский государственный университет

Основным носителем информациии о распознаваемых объектах в беспризнаковом подходе к распознаванию образов служит потенциальная функция $K(\omega', \omega'')$, играющая роль скалярного произведения для любых двух объектов распознавания $\omega', \omega'' \in \Omega$.

Потенциальная функция является симметричной действительнозначной функцией, удовлетворяющей условию Мерсера. Последнее условие выражается в неотрицательной определенности матрицы значений потенциальной функции $M = \{m_{ij} = K(\omega_i, \omega_j)\}$ для всякой конечной совокупности объектов $\{\omega_1, \ldots \omega_N\} \subset \Omega$. При этих условиях можно говорить, что потенциальная функция $K(\omega', \omega'')$ погружает множество распознаваемых объектов $\omega \in \Omega$ в линейное пространство.

Потенциальная функция определяет способ измерения несходства объектов между собой, и именно от её выбора зависит, будет ли выполняться гипотеза компактности для решаемой задачи. Именно это соображение является основным при формировании потенциальных функций для решения многих прикладных задач. В то же время, выполнением условий Мерсера для получаемой функции часто пренебрегают. В этом случае двойственная задача оптимизации, к которой сводится задача обучения при использовании метода потенциальных функций, перестает быть выпуклой, что приводит к трудностям получения решения.

Одним из распространенных способов формирования потенциальной функции является скалярное произведение яркостей $\omega_{\mathbf{t}}$ соответствующих пикселов $\mathbf{t} \in T$ изображения. Однако такой способ не учитывает возможности сдвигов и локальных деформаций изображений распознаваемых объектов, в частности, фотопортретов. Поэтому предварительно растры сравниваемых фотопортретов подвергают деформации таким образом, чтобы максимально совместить изображения фотопортретов. Поскольку для каждой пары изображений деформация растра v_t выбирается независимо, то, даже несмотря на то, что, вычислительно, яркостная потенциальная функция

$$K^B(\omega', \omega'') = \sum_{\mathbf{t} \in T} \omega'_{\mathbf{t}+\mathbf{v_t}} \cdot \omega''_{\mathbf{t}-\mathbf{v_t}}$$

представляет собой скалярное произведение яркостей, это уже не гарантирует выполнения условия Мерсера.

Однако введение базиса в линейном пространстве, в которое погружает объекты потенциальная функция $K^B(\omega', \omega'')$, позволяет обойти это препятствие [1].

Наряду с рассмотренной яркостной функцией $K^B(\omega', \omega'')$, также будем рассматривать деформационную потенциальную функцию $K^T(\omega', \omega'') = \exp(-\alpha|T(\omega', \omega'')|)$, где $|T(\omega', \omega'')|$ — мера деформации растров пары изображений ω' и ω'', а также комбинированную потенциальную функцию $K(\omega', \omega'') = \delta K^T(\omega', \omega'') + (1-\delta)K^B(\omega', \omega'')$ где δ — весовой коэффициент, $0 \leqslant \delta \leqslant 1$.

Обычно исследователь имеет дело с некоторым множеством $\Omega^0 = \{\omega_1, \ldots, \omega_n\}$ реально существующих объектов, причем не все из них предполагаются классифицированными. Это множество названо в [1] базисной совокупностью. Базисная совокупность играет роль конечного базиса в линейном пространстве Ω. Таким образом, каждому элементу линейного пространства Ω соответствует вектор его скалярных произведений с объектами базисной совокупности $\mathbf{x}(\omega) = (K(\omega, \omega_1), \ldots, K(\omega, \omega_n))$. Каждый объект ω_i базисной совокупности формирует свой проекционный признак $x_i(w) = K(\omega, \omega_i)$.

Очевидно, при обучении в пространстве проекционных признаков, для потенциальной функции $K(\omega', \omega'')$ уже не требуется выполнение условий Мерсера. Недостатком такого подхода является то, что получаемое при этом решающее правило выражается через все элементы базисной совокупности. В частности, в задаче распознавания личности по фотопортрету роль объектов распознавания играют растровые изображения, и процедура вычисления значения потенциальной функции для пары изображений является вычислительно сложной. Это приводит к большой вычислительной сложности алгоритма на этапе распознавания.

Будем считать, что каждый объект ω_i базисной совокупности формирует свою собственную потенциальную функцию $\tilde{K}_i(\omega', \omega'') = K(\omega', \omega_i)K(\omega_i, \omega'')$. В данной работе подмножество наиболее «информативных» потенциальных функций выбирается из всей базисной совокупности с помощью методологии комбинирования потенциальных функций, предложенной в работе [2], которая позволяет оценить информативность каждой отдельной потенциальной функции в виде веса? с которым она входит в общую функцию. Оценки весов вычисляются итерационным алгоритмом. В результате ненулевыми оказываются лишь небольшое число весов при потенциальных функциях тех объектов, которые и будут составлять искомый базис. Проверка предложенного подхода производилась на базе, состоящей из фотопортретов 295 человек (Таблица 1). Было построено 200 решающих правил, по одному на каждого человека. Результаты верификации личности на тестовой выбор-

295 человек по 8 фотографий			
200 клиентов по 8 фотографий		95 самозванцев по 8 фотографий	
$200 \times 6 = 1200$ для обучения	$200 \times 2 = 400$ для проверки	$25 \times 8 = 200$ для обучения	$70 \times 8 = 560$ для проверки

Таблица 1. Структура базы фотопортретов.

	(1)	(2)
Количество элементов в базисе	1400	11
Яркостная потенциальная функция, % ошибок	2.3	2.6
Комбинированная потенциальная функция, % ошибок	1.4	1.7

Таблица 2. Результаты распознавания на тестовой выборке: (1) все элементы обучающей совокупности в базисе; (2) после выбора базиса. Во втором случае количество элементов в базисе для каждого решающего правила разное, в среднем 11.

ке, приведенные в Таблице 2, показывают, что удаётся выбрать базис со значительно меньшей размерностью при небольшом ухудшении качества распознавания, что позволяет сильно уменьшить вычислительную сложность решающих правил.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00679, № 06-01-08042, № 06-01-00412, № 06-07-89249.

Литература

[1] *Mottl V. V., Dvoenko S. D., Seredin O. S., Kulikowski C. A., Muchnik I. B.* Featureless Pattern Recognition in an Imaginary Hilbert Space and Its Application to Protein Fold Classification. // Second Int. Workshop on Mach. Learn. and Data Mining in Patt. Rec., Leipzig, July 2001. — Pp. 322–336.

[2] *Моттль В. В., Середин О. С., Красоткина О. В., Мучник И. Б.* Комбинирование потенциальных функций в задачах восстановления зависимостей по эмпирическим даннным // Докл. РАН. — 2005. — Т. 401, № 5. — С. 607–612.

Управляемая визуализация спектра изображения

Жарких А. А., Коннов Е. В.

zharkihaa@mstu.edu.ru, wexum@mail.ru

Мурманск, Мурманский Государственный Технический Университет

В данной работе представлены результаты разработки программного модуля БПФ с управляемой визуализацией спектра изображения.

Методы вычисления спектра на основе БПФ достаточно хорошо отработаны и теоретически, и практически. Однако большинство руководств по спектральному анализу не содержат подробной информации о визу-

ализации спектра изображений [1, 2, 3]. В данной работе предлагается один из возможных методов такой визуализации.

Разработанный программный модуль позволяет выполнять следующие функции: прямое и обратное БПФ над изображением в градациях серого цвета; распознавание сдвига, поворота и масштаба изображения на основе корреляционного анализа; визуализация амплитудного и фазового спектров изображения. Программный модуль разработан на языке $C\#$. Работа модуля апробирована на изображениях каллиграфических букв русского рукописного текста. Создан банк спектров этих букв с целью изучения связей изображения буквы с изображением её спектра. Данные связи позволяют объединять сходные спектральные отсчеты отдельных рукописных букв в некоторые классы. В перспективе это позволит формулировать задачу распознавания рукописных букв как задачу распознавания объектов, принадлежащих отдельным классам. Реализована также процедура конкатенации изображений отдельных рукописных букв в целые слова.

Визуализация спектра

Рассмотрим отдельно задачу визуализации спектра. Считается, что, так как структура спектра изображения не важна для конечного пользователя, то она вообще не важна и неинтересна. На взгляд авторов это является заблуждением. Визуализация спектра важна для разработчика системы распознавания образов, т. е. для математика, программиста или специалиста в предметной области.

Визуализация спектра позволяет решить следующие задачи:

1) сформулировать в виде вербального описания требования к геометрической форме и прозрачности частотных коэффициента передачи фильтров;
2) сократить вычислительные ресурсы в задаче оптимизации признакового пространства (включающего подмножества спектральных отсчетов) за счет предварительного эвристического выбора частотного коэффициента передачи фильтра;
3) оптимизировать масштабы элементов признакового пространства (например, увеличить значения спектральных отсчетов, определяющих форму объекта, относительно низкочастотных спектральных составляющих, определяющих фон);
4) ускорить процесс изучения структуры спектра изображения изображения при решении задач 1)–3).

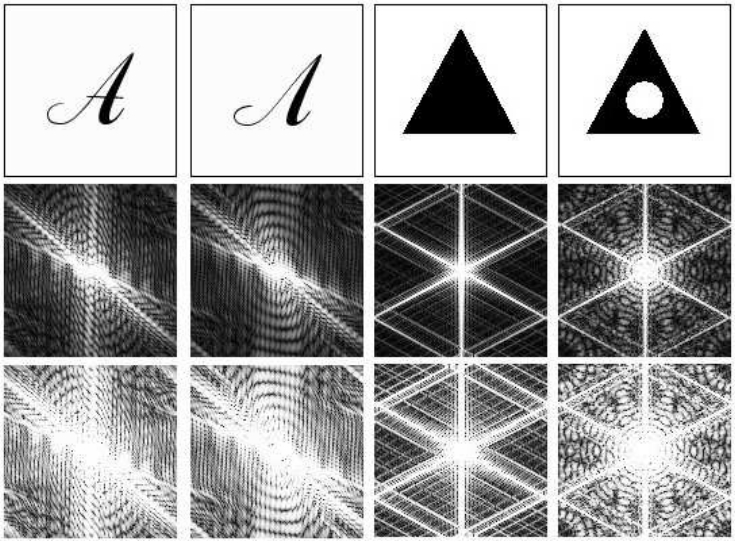


Рис. 1. Пример визуализации амплитудных спектров (средний ряд при $\alpha = 0.0002$, нижний ряд при $\alpha = 0.0008$) изображений схожих по начертанию букв рукописного текста и схожих по начертанию геометрических фигур.

Для визуализации фазового спектра была использована линейная процедура:

$$Y_P(m,n) = \left[255 \cdot \frac{\arg(X(m,n)) + \pi}{2 \cdot \pi}\right], \tag{1}$$

где $m = 0, \ldots, M-1,\ n = 0, \ldots, N-1,\ \arg\left(X(m,n)\right)$ — значение отсчета фазового спектра, $Y_P(m,n)$ — значение пикселя изображения фазового спектра.

Для визуализации амплитудного спектра была использована нелинейная процедура:

$$Y_A(m,n) = \left[255 \cdot \mathrm{th}\left(\alpha \left|X(m,n)\right|\right)\right], \tag{2}$$

где $m = 0, \ldots, M-1,\ n = 0, \ldots, N-1,\ |X(m,n)|$ — значение отсчета амплитудного спектра, $Y_A(m,n)$ — значение пикселя изображения амплитудного спектра, α — параметр для управления визуализацией.

На Рис. 1, 2 показаны результаты визуализации спектра с использованием процедур (1) и (2).

Заключение

Использование представленных методов визуализации спектра позволяет добиться более четкого и детального изображения, а также избавляет от проблемы визуализации участков спектра с малыми (стремящимися к нулю) значениями. Часто используемая для визуализации и анализа амплитудного спектра логарифмическая шкала, на наш взгляд, является непригодной, т. к. не позволяет вычислить значения отсчетов, равных нулю. Предложенный метод визуализации можно использовать как инструмент спектрального анализа, а также при решении задач распознавания образов. Как известно, преобразование Фурье позволяет реализовать аффинно-инвариантные процедуры распознавания образов. Однако конкретные реализации таких процедур тесно связаны с предметной областью. Создание банка видимых изображений спектров под конкретную предметную область позволит скорректировать процедуру уже на этапе выбора алгоритма. Другим важным аргументом использования именно выражения (2) для визуализации амплитудного спектра является следующее. Как показано в [4], функция $X = \mathrm{th}(x)$ реализует гомоморфное отображение поля вещественных чисел на интервал $(-1, +1)$. Операция сложения вещественных чисел переходит в следующую операцию на интервале $(-1, +1)$: $Z = (X + Y)/(1 + XY)$. Эта операция обладает гибридными свойствами. При малых значениях аргументов она переходит в обычную операцию сложения, а на любом из двухэлементных множеств $\{0, +1\}$ или $\{-1, 0\}$ она представляет собой логическое «или». Качественно действие этой операции для произвольных аргументов интервала $(-1, +1)$ сводится к тому, что аргумент больший по модулю «поглощает» аргумент меньший по модулю. Если, с учетом нормировки, применить эту операцию к двум изображениям одинакового размера, то изображения как бы «склеиваются». В контексте задачи распознавания указанная операция может быть использована для формирования амплитудного спектра как признака целого множества или класса изображений. Простое перемножение спектров для этого менее пригодно, так как разрыв между большими и малыми отсчетами только увеличивается.

Представляется перспективным использование предложенного инструмента спектрального анализа и визуализации в задачах обработки измерений (особенно уникальных, редко повторяющихся) физических величин, зависящих от двух переменных.

Открываются дополнительные возможности в решении прикладных задач распознавания образов, таких, как распознавание рукописного текста, отпечатков пальцев, фотографий лиц, изображений средств передвижения и др.

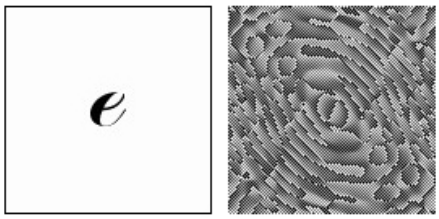


Рис. 2. Пример визуализации фазового спектра изображения буквы рукописного текста.

Литература

[1] *Залманзон Л.А.* Преобразования Фурье, Уолша, Хаара и их применение в управлении, связи и других областях. — М: Наука, 1989. — 496 с.

[2] *Марпл С.Л.* Цифровой спектральный анализ и его приложения. — М: Мир, 1990. — 584 с.

[3] *Даджион Д., Мерсеро Р.* Цифровая обработка многомерных сигналов. — М: Мир, 2006. — 488 с.

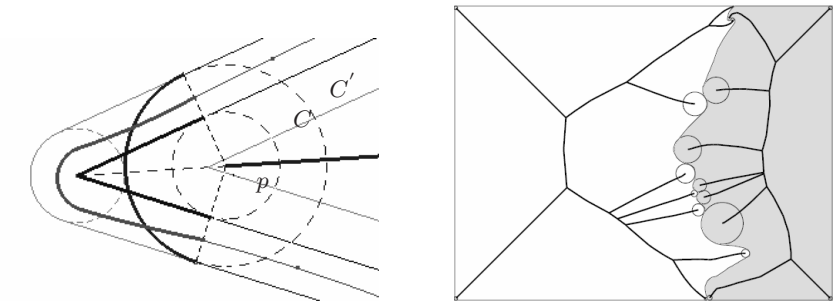
[4] *Жарких А. А.* Идентификация линейных стационарных систем при гомоморфных преобразованиях сигналов // 4-я межд. конф. «Идентификация систем и задачи управления». — Москва, SICPRO'05. — С. 321–332.

Выделение линии профиля по опорным точкам с применением базового скелета

Жукова К. В., Рейер И. А.
reyer@forecsys.ru
Москва, Вычислительный Центр РАН

Форма линии лицевого профиля является важной отличительной особенностью, которая используется в системах машинного зрения, решающих широкий круг задач: от идентификации человека [1, 2, 3, 4] до распознавания эмоций и выражений лица [5].

Важной вспомогательной задачей при этом является выделение линии профиля на изображении лица. Линию обычно выделяют по опорным точкам: кончик носа, переносица, лобные дуги, губы, подбородок, и пр. Эти точки удобно рассматривать как экстремумы кривизны контура головы. Традиционными инструментами для определения экстремумов кривизны являются аппроксимация растрового контура сплайнами [1] и масштабируемая модель кривизны границы (curvature scale space) [6, 3].



а) б)

Рис. 1. Гранично-скелетная модель и особенности кривизны границы.

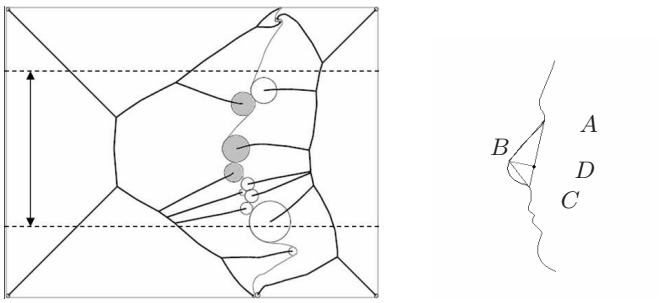
В работе рассматривается метод выделения линии профиля по особенностям кривизны, позволяющий работать с полигональным представлением границы. Метод основан на использовании непрерывной граничноскелетной модели изображения, состоящей из границы полигональной фигуры, аппроксимирующей растровый образ, и скелета этой фигуры. Для выявления особенностей кривизны используется скелетная составляющая модели.

Гранично-скелетная модель изображения

Скелет [7] является эффективным средством для описания структуры плоской области. Изучению скелета области с кусочно-гладкой границей посвящено большое количество работ, установлены различные свойства такого скелета [8]. В частности, известно, что форма скелета чрезвычайно чувствительна к локальным свойствам границы: с каждой точкой локального максимума кривизны (и точкой излома) границы связана отдельная терминальная ветвь скелета. Предлагаемый метод обнаружения и анализа экстремумов кривизнына основан на этом соображении.

Для обнаружения экстремумов кривизны предлагается использовать так называемый базовый скелет, описанный в [9]. Базовый скелет — это подмножество скелета многоугольной фигуры, которое не содержит ветвей, определяемых изменениями границы в пределах точности аппроксимации. Таким образом, базовый скелет можно рассматривать как модель существенной части скелета любой аппроксимирующей объект фигуры.

С каждой точкой базового скелета связан так называемый базовый круг. Это круг с центром в данной точке и радиусом $r + \varepsilon$, где r — радиус максимального вписанного в фигуру круга с центром в данной точке, а ε — точность аппроксимации. Дуга базовой окружности с центром в терминальной вершине внутреннего базового скелета аппроксимирует с известной точностью соответствующий участок границы (Рис. 1а). Та-



а) б)

Рис. 2. Базовые круги, соответствующие носовой части лица.

ким образом, этот участок можно рассматривать в качестве локального максимума кривизны контура в пределах точности аппроксимации.

Аналогично, если рассмотреть внешний базовый скелет многоугольной фигуры, дуги базовых окружностей с центрами в его терминальных вершинах указывают на локальные минимумы кривизны контура в пределах точности аппроксимации.

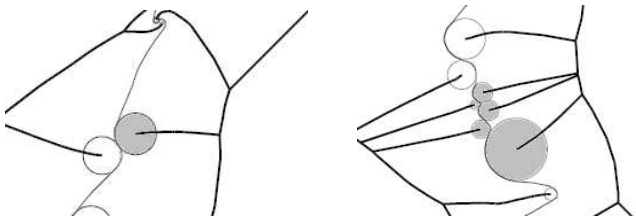
Для поиска опорных точек профиля предлагается использовать непрерывную модель изображения, которая состоит из границы многоугольной фигуры, разбивающей плоскость дискретного изображения на фигурную (содержащую все точки объекта) и фоновую (содержащую все точки фона) компоненты, и базовых скелетов фигурной и фоновой компонент (Рис. 1б).

Выделение линии профиля

Предположим, что нам известно в какую сторону повернута голова на изображении (например, влево). Соответственно, искать линию профиля будем в левой части контура (от самой левой точки с максимальной ординатой до самой левой точки с минимальной ординатой). Рассмотрим множество терминальных вершин внутреннего и внешнего базовых скелетов, максимальные вписанные круги которых касаются этого фрагмента границы, и множество соответствующих им базовых кругов.

Комбинацию базовых кругов, соответствующую носу, будем искать в средней части фрагмента. Эта комбинация должна состоять из трех последовательно расположенных вдоль границы кругов: внешнего (переносица), внутреннего (кончик носа) и снова внешнего (основание носа) (Рис. 2а). Таких троек в указанной области может оказаться несколько. Поэтому используется дополнительная проверка.

Для каждого из трех кругов рассматриваем соответствующий фрагмент контура и находим вершину наиболее удаленную от центра круга.



а) б)

Рис. 3. Базовые круги верхней и нижней частей лица.

Пусть A — точка, соответствующая переносице, B — кончику носа, C — основанию носа. Треугольник ABC (Рис. 2б) должен удовлетворять следующим условиям: (а) проекция D точки B на сторону AC делит ее таким образом, что длина CD меньше длины DA; (б) длина стороны AC не должна быть очень малой по сравнению с высотой средней части фрагмента. Если треугольник ABC не обладает этими свойствами, проверяем следующую тройку кругов.

После того, как найден нос, над кругом переносицы находим внутренний круг, соответствующий надбровным дугам (Рис. 3а), и, поднимаясь по фрагменту контура, ищем первый внешний круг (место, где начинается прическа). Если ни один внешний круг не найден, то определяем верхнюю точку профиля по следующему правилу — высота получившейся лобной части равна высоте носа.

Фрагмент контура, соответствующий губам и подбородку, в идеальном случае описывается последовательностью из пяти кругов, трех внутренних и двух внешних (Рис. 3б). Внутренний круг, соответствующий подбородку — последний в этой последовательности. Также возможны случаи, когда один или несколько кругов, предшествующих кругу подбородка, отсутствуют (всего 9 вариантов взаимного расположения кругов). Чтобы найти подбородок, рассмотрим последовательность из пяти кругов, расположенных ниже круга основания носа. Если эта последовательность оканчивается одним или несколькими внешними кругами, мы удаляем эти круги до тех пор, пока последним не станет внутренний. По чередованию внутренних и внешних кругов в полученной последовательности определяем, к какому типу она относится.

Для проведения экспериментов по выделению линии профиля использовалась база изображений Бернского университета [10]. Эта база содержит 150 полутоновых изображений профилей 30 человек. Также использовалась подготовленная авторами база, состоящая из 152 цветных изображений профилей 17 человек.

В 293 случаях линия профиля на изображении была выделена корректно. На 6 изображениях была неправильно выделена нижняя часть из-за наличия бороды. В 3 случаях лобная часть была ошибочно выделена как носовая.

Рассмотренная схема выделения основана на упрощенной модели линии профиля: не учитываются движения лицевых мышц, искажения формы линии такими деталями, как ресницы, брови, борода, и др. Тем не менее, эта схема может быть использована в качестве отправной точки для дальнейшего развития и модернизации. В частности, для борьбы с искажениями, вызванными ресницами, бровями и бородой, можно использовать базовый скелет с более низкой точностью аппроксимации.

Как видим, применение описаной гранично-скелетной модели изображения позволяет эффективно выделять фрагменты контура объекта, содержащие требуемую комбинацию особенностей кривизны, без построения кусочно-гладкой аппроксимации контура.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00542 и № 05-07-90395.

Литература

[1] *Wu C. J., Huang J. S.* Human face profile recognition by computer // Pattern Recognition. — 1990. — V. 23, No. 3-4. — Pp. 255–259.

[2] *Chellappa R., Wilson C. L., Sirohey S.* Human and machine recognition of faces: A survey // Proceedings of the IEEE. — 1995. — V. 83, No. 5. — Pp. 705–740.

[3] *Liposcak Z. Loncaric S.* A scale-space approach to face recognition from profiles // Lecture Notes in Computer Science. — 1999. — V. 1689. — Pp. 243–250.

[4] *Wallhoff F., Muller S., Rigoll G.* Recognition of Face Profiles from the MUGSHOT Database Using a Hybrid Connectionist/HMM Approach // IEEE Int. Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, Utah, 2001.

[5] *Pantic M., Rothkrantz L. J. M.* Facial action recognition for facial expression analysis from static face images // IEEE Transactions on Systems, Man and Cybernetics. — 2004. — V. 34, No. 3. — Pp. 1449–1461.

[6] *Campos J. C., Linney A. D., Moss J. P.* The analysis of facial profiles using scale space techniques // Pattern Recognition. — 1993. — V. 26, No. 6. — Pp. 819–824.

[7] *Blum H.* A transformation for extracting new descriptors of shape // Symposium on Models for the Perception of Speech and Visual Form, MIT Press, 1967.

[8] *Choi H. I., Choi S. W., Moon H. P.* Mathematical theory of medial axis transform // Pacific. J. of Math. — 1997. — V. 181, No. 1. — Pp. 57–88.

[9] *Местецкий Л. М., Рейер И. А.* Непрерывное скелетное представление изображения с контролируемой точностью // Труды 13-й международной кон-

ференции по компьютерной графике и машинному зрению «Графикон-2003», Москва, 2003.

[10] *Achermann B.* University of Bern Face Database. — Copyright 1995, University of Bern, all rights reserved. — `ftp://iamftp.unibe.ch/pub/Images/FaceImages/`.

Спектральные свойства искаженных изображений и системы распознавания

Карнаухов В. Н., Милюкова О. П., Чочиа П. А.

`victor.karnaukhov@iitp.ru, milukova@iitp.ru, chochia@iitp.ru`

Москва, Институт проблем передачи информации им. А. А. Харкевича РАН

Достаточно полной математической моделью неидеальных изображающих систем служит линейное интегральное уравнение I-ого рода. Решение этого уравнения (задача восстановления) — типичная обратная задача с неполными данными, которой посвящено огромное количество исследований. Однако каким бы эффективным ни был метод восстановления, успех в решении задачи определяется точностью математической модели изображающей системы.

Пусть формирование изображения в линейной системе описывается однородным интегральным уравнением типа свертки. Рассмотрим одномерный аналог задачи

$$Av = \int_{-\infty}^{\infty} h(x-\xi)v(\xi)\,d\xi = u(x).$$

Для апостериорного определения типа искажающего оператора и оценки параметров искажения будем использовать Фурье спектры искаженных изображений.

Искажающий оператор A в рамках данной модели задается ядром интегрального оператора $h(x)$. Рассмотрим спектральные свойства ядра. В реальных физических задачах $h(x)$ есть функция с конечным носителем, ядро $h(x)$ обращается в нуль при $|x| \geqslant a$. Тогда из теоремы Пэли-Винера следует, что преобразование Фурье $S(z) = \int_{-a}^{a} h(t)e^{-izt}\,dt$ — целая аналитическая функция от $z = x + iy$ экспоненциального типа, т. е. первого порядка $\rho = 1$ и конечного типа $\sigma \leqslant a$ [2].

Кроме функций с конечным носителем для описания искажающего оператора часто используют Гауссову функцию распределения и некоторые другие функции, спектры которых — целые функции. Как известно, целые функции обладают многими замечательными свойствами [1], которые (например, изолированность нулей) и определяют структурные особенности спектров Фурье искаженных изображений. В качестве примера на Рис. 1 показано некоторое тестовое изображение и его кепструм

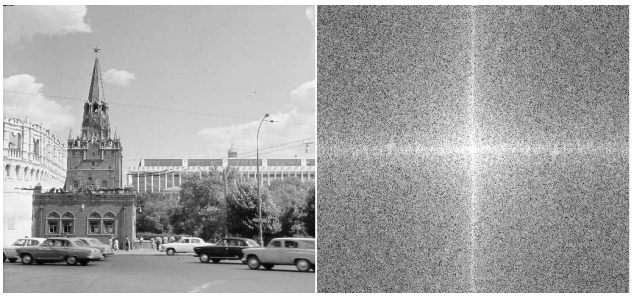


Рис. 1. Исходное изображение и его кепструм (логарифм модуля спектра).

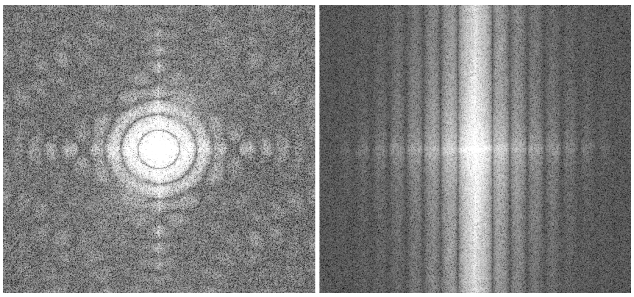


Рис. 2. Кепструмы расфокусированного и смазанного изображения.

(логарифм модуля спектра), а на Рис. 2 показаны кепструмы того же изображения, дефокусированного и смазанного. На основе проведенного анализа спектров искаженных изображений было предложено использовать специальную систему распознавания образов для определения типа искажающего оператора и оценки параметров искажения [3]. В качестве распознающей системы использовалась система, описанная в работе [4].

Кроме того, структурные особенности спектров искаженных изображений в некоторых случаях позволяют решать задачу распознавания искаженного изображения без предварительной процедуры восстановления. Пусть спектр $S(z)$ искажающего оператора с ядром $h(x)$ — целая функция. В точке z_0, где функция $S(z)$ равна нулю, $\arg S(z)$ не определен. Можно показать, что $\arg(z - z_0)$ при переходе через точку z_0 изменяется на $-\pi$ при обходе точки сверху или на $+\pi$ при обходе точки z_0 снизу. Если ядро искажающего оператора есть действительная четная функция, то все нули лежат на действительной оси. В каждой точке z_i, которая является корнем уравнения $S(z) = 0$, значение аргумента скачком изменяется на $+\pi$ или $-\pi$ в зависимости от выбранного нами направ-

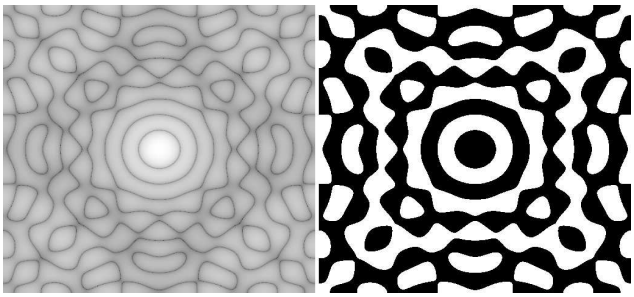


Рис. 3. Кепструм и фазовый спектр дефокусирующей функции

ления обхода. Для иллюстрации на Рис. 3 показаны кепструм и фазовый спектр искажающей дефокусирующей функции.

Следовательно, фаза исходного изображения совпадает с фазой искаженного в случае, когда спектр искажающего оператора не содержит нулей. Если ядро искажающего оператора — четная функция с конечным носителем, то фаза искаженного изображения может отличаться от фазы исходного на соответствующих интервалах действительной оси. Точки, где происходит скачок фазы, определяются местоположением нулей спектра на действительной оси и могут быть определены с помощью систем распознавания. Многие распознающие алгоритмы в качестве признаков используют компоненты фазового спектра Фурье изображения. Причем для эффективного распознавания может оказаться достаточным небольшого числа первых фазовых компонент [4]. Из сказанного выше следует, что в этих случаях можно распознать исходное неискаженное изображение по фазе искаженного изображения без предварительной процедуры восстановления.

Успех в решении перечисленных задач определяется эффективностью используемых методов и систем распознавания. Следовательно, если распознающая система эффективна при заданном уровне шума, то она может быть применена для выполнения поставленных задач при соответствующих ограничениях на уровень суммарной погрешности.

Работа выполнена при поддержке Программы фундаментальных научных исследований ОИТВС РАН, проект «Алгоритмическое и программное обеспечение инфокоммуникационных сетей».

Литература

[1] *Титчмарш Е.* Теория функций. — Москва: Наука, 1980.

[2] *Винер Н., Пэли Р.* Преобразование Фурье в комплексной области. — Москва: Наука, 1964.

[3] *Aizenberg I., Butakoff C., Karnaukhov V., Merzlyakov N., Milyukova O.* Blurred Image Restoration Using the Type of Blur and Blur Parameters Identification on the Neural Network // Proceedings of SPIE, Image Processing: Algorithms and Systems, 2002. — Vol. 4667. — Pp. 460–471.

[4] *Aizenberg I., Aizenberg N., Butakoff C., Farberov E.* Image recognition on the Neural Network based on Multi-Valued Neurons // Proc. of the Int'l Conf. on Patt. Recogn., Spain. — IEEE Computer Society Press, 2000. — Pp. 993–996.

Распознавание плоских и объемных изображений на основе дискретно-геометрических методов

Козлов В. Н.

vnkozlov@mail.ru

Москва, МГУ им. М. В. Ломоносова, мех.-мат. факультет

Работа относится к распознаванию образов и ориентирована на приложения в компьютерном зрении и стереовосприятии.

Изображение определяем как конечное (непустое) множество точек на плоскости (или в трехмерном пространстве, в случае объемных изображений). Содержательным обоснованием этому может служить то, что любое реальное (нецветное) изображение можно аппроксимировать изображением из точек, причем градации «серого цвета» передаются разной плотностью точек в разных частях изображения. Возможно рассмотрение и цветных изображений, поскольку, как известно, цветное изображение можно представить тремя нецветными. Наконец, если говорить о реальном зрении, то изображение из среды проецируется на сетчатку глаз, что приводит к возбуждению части рецепторных клеток, т. е. в конечном счете — к формированию на сетчатке аналога составленного из точек изображения.

Рассматриваемый подход к распознаванию существенным образом опирается на введение внутренней кодировки изображений, инвариантной к аффинным их преобразованиям.

В плоском и объемном случаях внутренний код изображений, для наглядности — фигур, вводится так. Нумеруются точки фигуры; с учетом ее размерности рассматривается множество всех симплексов, образованных точками фигуры; для каждого симплекса вычисляется мера. Код фигуры образует множество всех троек, состоящих из двух симплексов и числа, являющегося отношением их ненулевых мер.

Для каждой из размерностей доказано, что фигуры, с точностью до перенумерации их точек, имеют один и тот же код тогда и только тогда, когда они аффинно эквивалентны.

Сравнение и распознавание произвольных фигур A и B основывается на следующем. Порождаются множества A^* и B^* всех фигур, по-

лучаемых из A и B преобразованиями из некоторого класса (в общем случае аффинными). Рассматривается множество величин $r(A', B')$, где A' из A^*, B' из B^*, являющихся расстоянием Хаусдорфа между множествами A' и B'. Показывается, что минимум на этом множестве достигается на конечном его подмножестве, что и позволяет его вычислить. Этот минимум и служит мерой сходства и различия фигур. Содержательно это можно представить как такое наложение фигур друг на друга, при котором минимизируется степень их несовпадения по форме, причем независимо от первичной взаиморасположенности и взаимоориентации фигур, их размеров, растянутости или сжатия, локальных погрешностей.

Изображения в памяти распознающего устройства и предъявляемые для распознавания могут содержать множество мельчайших и ненужных деталей. Как упростить изображение, оставив на нем существенные для распознавания черты, если заранее не известно, что есть это существенное? Описывается, как по изображению построить его более простой аналог — эскиз, и доказывается, что сходство эскизов определенным образом связано со сходством оригиналов.

Рассмотрено (на доказательном уровне) восстановление объемных фигур по их плоским проекциям (моделирование стереовосприятия).

К настоящему времени имеются две компьютерные реализации подхода: по распознаванию произвольных фигур и по стереовосприятию.

Работа поддерживается РФФИ, проект № 07-01-00433.

Литература

[1] *Козлов В. Н.* Элементы математической теории зрительного восприятия — М.: ЦПИ мех.-мат. МГУ, 2001. — 128 стр.

[2] *Kozlov V. N.* Visual pattern and geometric transformation of images // Pattern recognition and Image Analysis. — 2000. — V. 10, No. 3. — Pp. 321–342.

Динамическое программирование с построчным комбинированием переменных для обработки изображений

Копылов А. В.

kopylov@uic.tula.ru

Тула, Тульский технический университет

Множество задач анализа изображений, таких как сглаживание, локальный текстурный анализ, сегментация, совмещение изображений в стереовидении или распознавании образов, и т. д., могут быть представлены как задачи преобразования исходного изображения $Y = (y_{\mathbf{t}},\, \mathbf{t} \in \mathbf{T})$, которое обычно принимает значения на непрерывной или дискретной оси уровня яркости $y_{\mathbf{t}} \in \mathbf{Y}$ и определено на подмножестве $\mathbf{T} =$

$= \{\mathbf{t} = (t_1, t_2)\colon t_1 = 1, \ldots, N_1,\ t_2 = 1, \ldots, N_2\}$ двумерного пространства (плоскости изображения), во вторичный массив $X = (x_{\mathbf{t}}, \mathbf{t} \in \mathbf{T})$. Данный массив определен на том же множестве аргументов $\mathbf{t} \in \mathbf{T}$ и принимает значения $x \in \mathbf{X}$ из множества, специфичного для каждой задачи.

Основная идея оптимизационного подхода к анализу изображений состоит в том, что алгоритм анализа данных строится как алгоритм оптимизации, условно минимизации, подходящей целевой функции, определенной на множестве всех возможных вариантов вторичного массива данных. Такая целевая функция практически всегда может быть выбрана в так называемой парно-сепарабельной форме как сумма элементарных функций двух видов, а именно узловых функций и функций связи.

$$J(X\,|Y\,) = \sum_{\mathbf{t}\in\mathbf{T}} \psi_{\mathbf{t}}(x_{\mathbf{t}}|Y_{\mathbf{t}}) + \sum_{(\mathbf{t}',\mathbf{t}'')\in G} \gamma_{\mathbf{t}',\mathbf{t}''}(x_{\mathbf{t}'}, x_{\mathbf{t}''}). \tag{1}$$

Узловые функции $\psi_{\mathbf{t}}(x_{\mathbf{t}}|Y_{\mathbf{t}})$, несущие информацию о данных, играют роль меры несходства между искомым локальным значением $x_{\mathbf{t}}$ и некоторой окрестностью $Y_{\mathbf{t}}$ точки $\mathbf{t}$ обрабатываемого массива. Каждая функция связи $\gamma_{\mathbf{t}',\mathbf{t}''}(x_{\mathbf{t}'}, x_{\mathbf{t}''})$, основанная на модельных представлениях, накладывает штраф на различие значений пары соседних переменных на соответствующем ребре неориентированного графа G, который образован множеством пар $G \subset T \times T$ соседних элементов изображения.

В случае, когда целевые переменные принимают значения из конечного множества $x \in \mathbf{X} = \{1, \ldots, m\}$, данная оптимизационная задача является одним из частных случаев так называемых задач разметки, который известен как (max, +) или (min,+) задачи.

Для произвольного графа соседства G задача оптимизации парно-сепарабельной целевой функции (1) является NP-полной, но в частном случае (min, +) задачи, когда соседство переменных определяется при помощи произвольного ациклического графа, принцип последовательного исключения переменных приводит к оптимизационной процедуре [1], имеющей много общего с принципом динамического программирования Беллмана. Этот принцип состоит в разложении исходной задачи на последовательность подзадач с уменьшающимся числом переменных, вплоть до единственной переменной на последнем шаге.

В том случае, когда граф соседства имеет вид решетки, принцип динамического программирования Беллмана не может быть непосредственно применен, тем не менее представляется очень заманчивым сохранить вычислительные преимущества процедуры динамического программирования, даже ценой некоторых эвристических компромиссов.

Прежде всего мы можем провести разбиение исходного графа на совокупность поддеревьев. Затем, следуя итерационному методу

Гаусса-Зайделя, можно оптимизировать целевую функцию путем поиска на каждом шаге глобального минимума частичных целевых функций, связанных с поддеревьями, при помощи процедуры древовидного динамического программирования [1]. Для того, чтобы достичь более глубокого минимума, можно изменять разбиение графа на каждом шаге процедуры. Возможно также построение неитерационного приближенного алгоритма на основе удаления некоторых ребер графа соседства [2], [3].

Тем не менее, существует класс задач разметки, для которых граф смежности переменных не удается разложить на поддеревья, для которых частичная оптимальная разметка совпадает с глобально оптимальной. Такие задачи не могут быть решены подобными методами.

В данной работе рассматривается совершенно другой подход к задаче парно-сепарабельной оптимизации с графом смежности переменных в виде решетки. Основная идея данного подхода состоит в представлении исходной целевой функции с решетчатой смежностью переменных (1) как функции, переменные которой представляют уже не отдельные узлы решетки, а целые строки этих узлов. При этом строки решетки естественным образом выстраиваются в вертикальную цепочку, образуя тем самым цепочечный граф смежности для нового типа агрегированных переменных, каждая из которых представляет собой всю совокупность элементарных переменных в соответствующей строке.

Оптимизация функции с цепочечным графом смежности переменных формально является классической задачей динамического программирования, и может быть выполнена за число шагов, равное числу строк решетки N_1. Однако, являясь линейной относительно числа строк, вычислительная сложность процедуры оптимизации оказывается экспоненциальной относительно числа целевых переменных. Для преодоления этой трудности в данной работе рассматривается приближенная процедура оптимизации парно-сепарабельных целевых функций с решетчатым графом смежности, основанная на эвристической замене каждой функции Беллмана относительно группы переменных в отдельной строке подходящей парно-сепарабельной функцией, что позволяет снизить вычислительную сложность процедуры до линейной относительно числа элементов массива $N_1 N_2$.

Для экспериментальной проверки точности разработанных алгоритмов были использованы некоторые классы (min, +) задач разметки, для которых известно точное решение [4]. Первый класс задач представлен субмодулярными задачами, возникающими при построении трехмерной модели человеческого лица. Решение этих задач получено при помощи алгоритма минимального сечения графа [5]. Второй класс составляют задачи сегментации, для которых существуют эквивалентные тривиаль-

ные задачи, и которые решены при помощи алгоритма линейной релаксации [6]. Третий класс задач составляют модельные задачи, полученные из тривиальных путем случайных эквивалентных преобразований [6].

Для сравнения применялись алгоритмы на основе древовидной декомпозиции и принципа Гаусса-Зайделя [3], который использует решение, полученное алгоритмом с аппроксимацией решетчатого графа соседства последовательностью деревьев [1], в качестве начального приближения. Эксперименты показали, что алгоритмы с ациклической оптимизацией имеют достаточную точность для субмодулярных задач, так же как и для задач остальных двух классов.

Работа выполнена при поддержке РФФИ, проекты №06-01-08042 и №06-01-00412.

Литература

[1] *Mottl V., Blinov A., Kopylov A., Kostin A.* Optimization techniques on pixel neighborhood graphs for image processing // Graph-Based Representations in Pattern Recognition. — Springer–Verlag/Wien, 1998. — Pp. 135–145.

[2] *Veksler. O.* Stereo Correspondence by Dynamic Programming on a Tree // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). — 2005. — Vol. 2. — Pp. 384–390.

[3] *A. Kopylov.* Acyclic pair-wise separable optimization for image processing // Proceedings of the Ninth International Conference on Pattern Recognition and Information Processing. Vol 1, 2007. — Pp. 203–207.

[4] *Schlesinger M. I., Flach B.* Some solvable subclasses of structural recognition problems // Proc. of Czech Patt. Recogn. Workshop, Praha, 2000. — Pp. 55–62.

[5] *Boykov Yu., Veksler O., Zabih R.* Fast approximate energy minimization via graph cuts // PAMI, 23(11), November 2001. — Pp. 1222–1239.

[6] *Schlesinger M. I.* Syntactic analysis of two-dimensional visual signals in noisy conditions // Kibernetika, 4, 1976. — Pp. 113–130.

Создание математических методов, параллельных алгоритмов и программ для решения задач анализа изображений и задач управления в системах высокоточной навигации и наведения движущихся объектов

Костоусов В. Б., Кандоба И. Н., Перевалов Д. С.

vkost@imm.uran.ru

Екатеринбург, Институт математики и механики УрО РАН

Доклад посвящен обзору результатов исследований в области обработки и анализа изображений, проводимых в ИММ УрО РАН, и их применения для решения ряда прикладных задач. Основное внимание

уделяется математическим методам обработки и анализа черно-белых полутоновых изображений. Рассматриваются вопросы, связанные с построением на основе этих методов автоматических и автоматизированных алгоритмов анализа таких изображений. Обсуждаются проблемы эффективного распараллеливания этих алгоритмов, а также результаты их программной реализации на высокопроизводительных параллельных вычислительных системах. Приводятся результаты практического применения разработанного математического, алгоритмического и программного обеспечения для решения задач управления в системах высокоточной навигации и наведения движущихся объектов по геофизическим полям, а также построения специализированных систем технического зрения.

Дешифрирование аэрокосмических снимков

В задачах навигации по геофизическим полям эффективность их решения во многом зависит от качества эталонной навигационной информации. Часто в качестве эталонной информации используются данные о природных и искусственных объектах на подстилающей земной поверхности. Источником данных об объектах являются космические снимки высокого разрешения. Такие снимки играют исключительно важную роль в процессе создания новых и обновления существующих цифровых карт местности, которые являются первичным источником информации для подготовки эталонной информации для навигационных систем. В докладе обсуждаются некоторые автоматические и полуавтоматические алгоритмы дешифрирования полутоновых черно-белых космических снимков высокого разрешения. Здесь под дешифрированием снимка понимается процесс получения информации об объектах местности (расположение и взаимная ориентация) и отнесение этих объектов к некоторому множеству классов. Общая технологическая схема дешифрирования космического снимка высокого разрешения состоит из следующих основных этапов: предварительная обработка изображения, обнаружение и выделение объекта на изображении, определение качественных и количественных характеристик объекта, непосредственное распознавание (идентификация) объекта. В докладе представляются некоторые методы выделения отдельных классов объектов — городской и промышленной застройки, дорожной и речной сети, и др. Также рассматривается ряд методов автоматической и автоматизированной идентификации различных классов топографических объектов [1, 2].

Разработка и анализ систем технического зрения

При построении систем технического зрения реального времени одной из ключевых проблем является достижение высокой скорости обработки

изображений с сохранением качества получаемого результата. Показано, что в ряде задач этого можно достичь, используя специальные способы представления входных данных, например, порядковые шкалы измерений [3]. Другой возможностью является разработка узкоспециализированных алгоритмов анализа изображений [4]. Эффективность таких алгоритмов обусловлена тем, что они существенным образом используют априорную информацию о геометрии и цвете интересующих объектов на изображении. При этом одной из основных проблем является оценка качества и устойчивости их работы в изменяющихся условиях съёмки. Обсуждается способ построения оценки устойчивости алгоритма с помощью вероятностной модели, описывающей работу детекторов элементарных признаков. Приводится пример анализа системы технического зрения, предназначенной для наведения и контроля автоматической системы расцепки железнодорожных вагонов.

Работа выполнена при поддержке РФФИ, проект № 06-01-00229.

Литература

[1] *Kandoba I., Kostousov V., Skripnuk V., Shabanov G.* The system for automated deciphering of cosmic earth surface photographs // 19th Cong. of Int'l Society of Photogrammetry & Remote Sensing (ISPRS), Amsterdam, 2006. — P. 196.

[2] *Kandoba I. N., Kostousov V. B., Vlasova M. V., Skripnyuk V. V.* The system for automated interpretation of satellite photographs for digital map revision // 10th Int'l Conf. of Integrated Navigation Systems, S.-Petersburg, 2003. — Pp. 174–175.

[3] *Перевалов Д. С.* Использование матриц сравнений в задаче поиска по эталону // Материалы IX Межд. конф. «Интеллектуальные системы и компьютерные науки», МГУ, Москва, 2006. — С. 226–228.

[4] *Перевалов Д. С.* Вероятностная модель и эффективный алгоритм распознавания изображения автосцепки // 37-я регион. мол. конф. «Проблемы теор. и приклад. мат-ки», ИММ УрО РАН, Екатеринбург, 2006. — С. 468–472.

Сегментация цветных телевизионных изображений лиственного покрова в задачах лесной таксации

Кревецкий А. В., Ипатов Ю. А.

inf@marstu.mari.ru

Йошкар-Ола, Марийский гос. тех. университет

В лесном хозяйстве при проведении исследований в области анализа тенденции роста многолетних растений необходимо знать проективную зону лиственного покрова. Ручные методы анализа фотографических изображений лиственного покрова являются очень трудоемкими, длительными и экономически затратными. В настоящей работе предла-

Рис. 1. Типовое изображение лиственного покрова.

гается один из путей автоматизации анализа таких цифровых изображений, который, в сочетании с возможностью ручной коррекции ошибок обнаружения фрагментов интересующих растений, обеспечивает более высокую точность измерений и на два порядка более высокое быстродействие по сравнению с применяемым ручным методом.

Для вычисления относительной площади проективного лиственного покрова по его цифровому изображению необходимо относительно каждого элемента разрешения принять обоснованное решение – отнести элемент к фрагменту листвы интересующего растения или к мешающим растительным объектам. Типовое изображение исследования представлено на Рис. 1.

При использовании в качестве дискриминационных признаков данных о яркости и цвете пикселей, для построения оптимального или квазиоптимального (в байесовском смысле) алгоритма принятия решения, важно знать законы распределения вероятностей цвета полезных и мешающих пикселов в цветовом пространстве. Обучающая выборка формируется путём выделения характерных областей, относящихся к лиственному покрову и фону с помощью специально разработанной программы.

На Рис. 2 приведены выборочные условные законы распределений цвета $W(\bar{I}\,|\,H_1)$ и $W(\bar{I}\,|\,H_2)$ для обеих гипотез в RGB пространстве. Видно, что отсчеты статистически неоднородного фона и проективного покрова выделяются в слабо перекрывающиеся кластеры. Их вытянутый вдоль диагонали цветового куба характер объясняется неравномерной освещенностью полезных объектов и фона, и поэтому учет яркостной информации мало информативен.

В связи с этим, для упрощения алгоритма сегментации предлагается использовать проекции данных распределений на секущую плоскость, перпендикулярную вектору (255,255,255), см. Рис. 3.

Для разрабатываемого алгоритма сегментации указанные распределения $W(\bar{I}\,|\,H_1)$ и $W(\bar{I}\,|\,H_2)$ аппроксимируются функциями

$$K(\mathbf{x}, i) = \exp\bigl(-\alpha\,\|\mathbf{x} - \mathbf{x}_i\|^2\bigr), \tag{1}$$

где α — декремент затухания, $\|\mathbf{x}\|$ — норма вектора $\mathbf{x}$ в двухмерном пространстве, i — номер проверяемой гипотезы. Для выбранной формы ап-

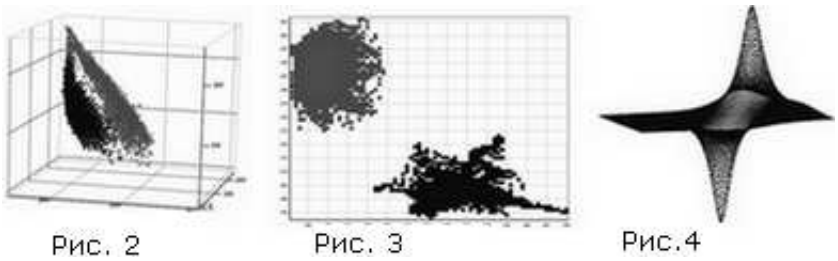


Рис. 2. Кластерная модель лиственного покрова и фона в RGB пространств.**Рис. 3.** Проекция на секущую плоскость. **Рис. 4.** Двухмерная потенциальная функция.

проксимации распределений оптимальный по критерию максимального правдоподобия (или минимального расстояния в цветовом пространстве) сводится к следующим шагам:

1. Определение проекции цвета текущей точки на выбранную плоскость цветового пространства.
2. Вычисление для нее величины отношения правдоподобия, где (x, y) — координаты пикселя в кадре изображения:

$$\bar{\lambda}(x, y) = W(\bar{I} \,|\, H_2) \;/\; W(\bar{I} \,|\, H_1).$$

3. Нормировка поля отношений правдоподобия к 255 градациям серого для возможности визуализации, Рис. 5.
4. Пороговая обработка нормированного изображения $\lambda\,(x, y)$, Рис. 6:

$$U(x, y) = \begin{cases} 1, & \text{если } \lambda(x, y) \geqslant \lambda; \\ 0, & \text{если } \lambda(x, y) < \lambda. \end{cases}$$

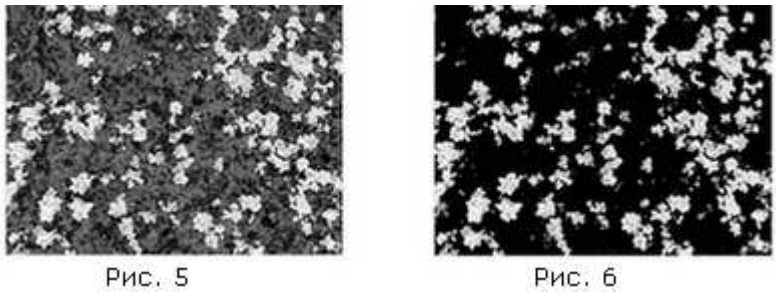


Рис. 5. Нормировка изображения потенциальной функцией вида (1).
Рис. 6. Результирующее изображение после пороговой обработки.

С целью снижения вероятности ошибок принимаемых решений, связанных с особенностями видов растений, условиями наблюдений и другими неучтенными факторами, в программном комплексе введена возможность ручной коррекции результатов сегментации.

Выводы

Разработанная программа позволяют автоматизировать процесс анализа и распознавания фотографических изображений с выигрышем по скорости и эффективности проведения исследований.

Реализуемый в программе алгоритм является оптимальным по критерию максимального правдоподобия. Для исключения выбросов, связанных с неоднородность фона, на котором производиться распознавание, в программе предусматривается ручная коррекция результатов автоматического анализа.

Точность измерений превышает ручные методы, используемые для проведения такого рода исследований. Данная программа имеет свидетельство об официальной регистрации программ для ЭВМ № 2007610624.

Литература

[1] *Прэтт У.* Цифровая обработка изображений. — М.: Мир, 1982.

[2] *Гонсалес Р., Вудс Р.* Цифровая обработка изображений. — М.: Техносфера, 2005. — 1072 с.

[3] *Ту Дж., Гонсалес Р.* Принципы распознавания образов. — М.: Мир, 1978.

Векторизация бинарных изображений на многоядерном процессоре

Кудинов П. Ю., Местецкий Л. М.
pkudinov@gmail.com, l.mest@ru.net
Москва, МГУ

К задаче векторизации граничных контуров (аппроксимации их многоугольниками) предъявляются очень высокие требования по скорости. Появление многоядерных процессоров является ресурсом для повышения эффективности этих алгоритмов. Предлагаемый метод параллельной векторизации включает в себя четыре шага: разделение изображения на части, прослеживание каждой части параллельно, объединение результатов прослеживания, вытягивание многоугольных фигур минимального периметра [2].

Алгоритм прослеживания

Разработанный метод позволяет использовать произвольные алгоритмы, прослеживающие граничные пары точек контуров на изображении.

Алгоритм должен выявить все пары соседних по горизонтали разноцветных точек, относящихся к одному контуру, и упорядочить эти пары вдоль контура. Примером является метод симплексного прослеживания, описанный в [2]. На Рис. 1 показан пример работы такого алгоритма. Разноцветными кружками обозначены пиксели изображения, а крестиками — точки, входящие в след трассировки.

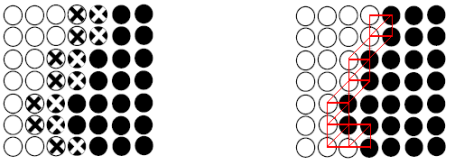


Рис. 1. След трассировки контура (слева) и маршрут движения симплекса (справа).

Описание метода

Разделение изображения на части. Изображение разделяется на N частей таким образом, чтобы части имели три общие строки.

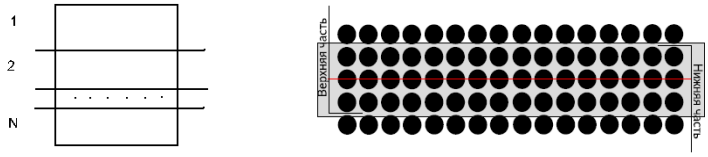


Рис. 2. Разделение изображения на части.

Среднюю из этих трех строк будем называть *линией разреза*. Интерес представляют контуры, в следе трассировки которых есть точки, общие с линией разреза, так как именно эти контуры оказались *разрезанными*. На Рис. 3 (справа) изображены разрезанные контуры К1, К2, К3, по которым нужно восстановить два контура: внешний и внутренний для фигуры серого цвета.

После того, как изображение разделено на части, проводится анализ линий разреза. В каждой линии регистрируются пары разноцветных точек — *места склейки*, на Рис. 3 (слева) они обведены прямоугольниками.

Прослеживание частей. Каждая часть прослеживается параллельно выбранным алгоритмом. В процессе прослеживания регистрируются выходы на линии разреза в парах разноцветных точек. Это поз-

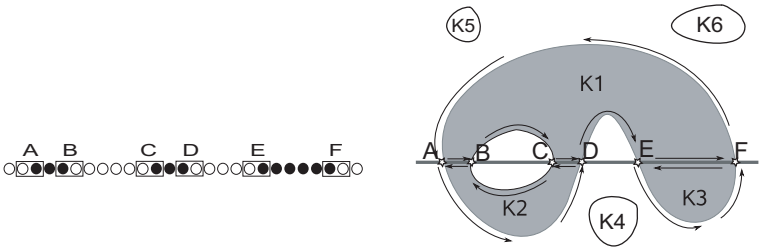


Рис. 3. Пары разноцветных точек в линии разреза (слева), разрезанные и независимые контуры (справа).

воляет установить соответствие между местами склейки и контурами, которым они принадлежат. Каждому месту склейки соответствует два контура — в верхней и нижней частях, а каждому контуру — последовательность мест склейки, упорядоченных по направлению прослеживания.

Операция склейки. Прослеживание разрезанных частей вообще говоря меняет топологию контуров. На Рис. 3 изображена фигура серого цвета, образованная двумя контурами: внешним черного цвета и внутренним — белого. После разреза образовалось три контура: K1, K2, K3 из которых и нужно получить исходные контуры.

Операция заключается в последовательном просмотре массива мест склейки, в котором места склейки упорядочены слева направо. На каждой итерации из установленного на этапе прослеживания соответствия выбираются два контура из верхней и нижней части и соединяются в месте склейки (происходит перенаправление ссылок в списках прослеживания). Затем определяется следующее место склейки (после A следующим будет место склейки, отмеченное как D, Рис. 3), выбираются соответствующие контуры K1, K2 и производится соединение. Аналогично будут обработаны E, F, после чего следующим местом склейки станет A. На этом процедура восстановления внешнего контура завершена. Осталось две пары разноцветных точек: B, C. По ним так же восстанавливается контур.

Операция восстановления контуров линейна по количеству мест склейки, что позволяет рассчитывать на небольшое время ее выполнения на большинстве изображений.

Вытягивание многоугольника минимального периметра.

Воспользовавшись независимостью следов трассировки контуров, можно запустить процедуры восстановления непрерывных границ параллельно на нескольких процессорах. Речь идёт о построении многоуголь-

ника минимального периметра, лежащего целиком в коридоре, образованном разноцветными точками вдоль контура.

Результаты

Предложенный метод прослеживания и склейки контуров является эффективным и не влияет на скорость обработки частей изображения, что позволяет рассчитывать на N-кратное увеличение производительности вычислений при использовании N процессоров (или N-ядерного процессора). Метод был реализован в полном объеме с использованием технологии OpenMP [1] на языке C++. Результаты были подтверждены экспериментально.

Литература

[1] *Воеводин В. В., Воеводин Вл. В.* Параллельные вычисления. — СПб.: БХВ-Петербург, 2002. — 608 с.

[2] *Местецкий Л. М.* Непрерывный скелет бинарного растрового изображения // Тр. конф. «Графикон-98», Москва, 1998.

Определение темпа музыкального произведения методом конкурирующих гипотез.

Курганский Д. А.
dmitry@widisoft.com
Москва, WIDISOFT

Ритмическая структура является одной из важнейших характеристик музыкального произведения. Музыкальное произведение представляет собой последовательность акустических событий, расположенных в узлах достаточно равномерной ритмической сетки (tatum grid) [1]. Важной составляющей задачи распознавания музыки является нахождение узлов этой сетки и определение их метрической роли (сильная или слабая доля).

Алгоритм, позволяющий определять положения музыкальных долей, может применяться не только для нахождения границ тактов, но и для определения точек вероятной смены гармонии [2], а также для предсказания положений последующих нот. В данной работе рассматривается алгоритм, определяющий положения узлов ритмтческой сетки и границ музыкальных долей. В работе [1] автором был рассмотрен алгоритм определения сетки татумов с помощью функции ошибки остатка, и продемонстрировано, что этот алгоритм применим только для случая квазистационарного темпа. В настоящей работе для решения поставленной задачи используется метод конкурирующих гипотез и показывается его приме-

нимость для случаев быстрой смены темпа и сложной ритмической фактуры.

Метод конкурирующих гипотез

В алгоритме производится последовательный анализ ряда, членами которого являются отсчеты времени, соответствующие началам нот. На основе этих отсчетов строятся локальные гипотезы G, содержащие начальное время t_1, период p и штраф Z. Алгоритм строит последовательности локальных гипотез, создавая дочерние гипотезы от уже существующих. Наследование заключается в том, что начальная точка наследника совпадает с предсказанием родительской гипотезы, а штраф включает в себя штраф родительской гипотезы. Число наследников у одной гипотезы не ограничено, то есть формируется дерево гипотез. В случае, когда две гипотезы совпадают, выбирается имеющая меньший штраф. После анализа всего ряда формируются несколько цепочек гипотез, среди которых выбирается гипотеза с наименьшим штрафом.

Построение цепочки гипотез и присвоение штрафа

Имея в своем распоряжении всего два члена ряда, x_i и x_{i+1}, можно построить гипотезу $G(x_{i+1}, p_i, 0)$, где $p_i = x_{i+1} - x_i$. Гипотеза G предсказывает появление следующей ноты в точке $x_{i+2} = x_{i+1} + p_i$. Проверка гипотезы производится сравнением с истинным положением ноты. Предсказание гипотезы считается правильным, если предсказанная величина совпадает с началом ноты, с точностью до достаточно малой величины δp. Эта величина соответствует максимальному интервалу между началами нот, которые слышны как одновременные. В данной работе использовалось значение δp, равное 20 мс.

Если предсказание отличается от реальной ноты на величину, большую δp, то вычисляется ошибка предсказания, и гипотезе присваивается штраф за несовпадение $Z_{unmatch}$, зависящий от этой ошибки. В этом случае от точки x_{i+1} строится вторая гипотеза G_2, с периодом p_{i2}, правильно предсказывающая поступившую ноту, которая также является наследником гипотезы G. Этой гипотезе присваивается штраф за изменение периода $Z_{irreg}(p_i/p_{i2})$.

Помимо описанных случаев, гипотеза может правильно предсказать не $i+1$ ноту, а $i+2$ или более позднюю. В этом случае гипотезе присваивается штраф за пропуск ноты Z_{skip}. Таким образом, штраф каждой гипотезы состоит из следующих слагаемых:

$$Z = Z_{inhert} + Z_{unmatch} + Z_{irreg} + Z_{skip}, \tag{1}$$

где Z_{inhert} — это величина штрафа, унаследованного от родительской гипотезы. Очевидно, что штраф может иметь нулевое значение, только если

исходная последовательность времён является строго равномерной. Но и при переменном темпе выбор цепочки гипотез с наименьшим штрафом позволяет построить хорошее приближение сетки татумов из совокупности узлов всех гипотез.

Применение алгоритма

В настоящее время алгоритм реализован в двух версиях, осуществляющих нахождение узлов ритмической сетки в реальном времени, а также для случая предварительно записанной последовательности нот. Использование данных о громкости нот позволяет определить метрическую роль узлов сетки. Алгоритм демонстрирует хорошую устойчивость по отношению к изменениям темпа и сложным ритмическим картинам.

Качество работы алгоритма определяется в существенной мере подбором параметров штрафа, оптимальный набор которых зависит от музыкального жанра. Важно, что алгоритм показывает удовлетворительный результат, даже если исходная последовательность нот содержала существенное количество ошибок, например, была получена в результате автоматического распознавания музыкального произведения.

Учет глобальных характеристик, таких как средний темп и его дисперсия, соотношение общего числа нот и числа найденных узлов, и. т. д., позволит улучшить поведение алгоритма в ритмически сложных случаях.

Литература

[1] *Курганский Д. А.* Анализ ритма и определение темпа музыкальных произведений // Доклады конференции ММРО-12, Москва: Макспресс, 2005. — С. 362–363.

[2] *Masataka Goto, Yoichi Muraoka* Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions // Speech Communication. — 1999. — Т. 1, № 27. — С. 311–335.

Алгоритм синтеза фазокодированных последовательностей с заднным уровнем боковых лепестков циклической АКФ

Леухин А. Н.
inf@marstu.mari.ru
Йошкар-Ола, ГОУ ВПО Марийский государственный технический университет

Предложен алгоритм синтеза шумоподобных фазокодированных (ФК) последовательностей с заданным уровнем боковых лепестков циклической автокорреляционной функции (АКФ). Приводятся выражения для синтеза всех возможных кодовых последовательностей для произвольного числа кодовых интервалов.

Постановка задачи синтеза

Проблемой синтеза шумоподобных сигналов, начиная с 50-х годов прошлого столетия, занимаются многочисленные научные коллективы у нас в стране и за рубежом [1, 2, 3, 4]. Особый практический интерес представляют фазокодированные дискретные последовательности, обладающие нулевым уровнем боковых лепестков циклической АКФ. Решение задачи для «базисных» случаев рассмотрено в работах [5, 6, 7, 8].

Дискретную фазокодированную последовательность $\mathbf{\Gamma} = \{\gamma_n\}_{0,N-1}$ можно представить в виде:

$$\gamma_n = \exp\left(i\varphi_n\right), \quad n = 0, \ldots, N-1. \tag{1}$$

Циклическую АКФ можно определить на основе выражения:

$$\eta_\tau = \sum_{n=0}^{N-1} \gamma_{n+\tau(\mathrm{mod}\ N)}\gamma_n^*, \quad \tau = 0, \ldots, N-1. \tag{2}$$

Нулевой отсчет циклической АКФ должен быть равен размерности кодовой последовательности $\eta_0 = N$, а все остальные отсчеты (боковые) должны принимать одинаковое значение a: $\eta_1 = \eta_2 = ... = \eta_{N-1} = a$. Значение уровня боковых лепестков a может быть любым вещественным числом из диапазона $a \in [a_{\min}, N]$, где верхняя граница диапазона $a_{\max} = N$, а нижняя граница $a_{\min}$ зависит от размерности кодовой последовательности N и удовлетворяет условию $a_{\min} \geqslant \frac{N}{1-N}$.

На основании выражений (1), (2) задачу синтеза ФК последовательностей с одноуровневой АКФ сводится к решению системы уравнений:

для четных $N,\ K=\frac{N}{2}-1,\ n=1,\ldots,K$:

$$\begin{cases} \cos(\varphi_n)+\cos(\varphi_{N-n})+\sum\limits_{m=1}^{N-n-1}\cos(\varphi_m-\varphi_{m+n})+ \\ \qquad\qquad +\sum\limits_{m=1}^{n-1}\cos(\varphi_m-\varphi_{m+N-n})=a; \\ \cos(\varphi_K)+\sum\limits_{m=1}^{N-K-1}\cos(\varphi_m-\varphi_{m+K})=a/2; \\ \sin(\varphi_n)-\sin(\varphi_{N-n})-\sum\limits_{m=1}^{N-n-1}\sin(\varphi_m-\varphi_{m+n})+ \\ \qquad\qquad +\sum\limits_{m=1}^{n-1}\sin(\varphi_m-\varphi_{m+N-n})=0. \end{cases} \tag{3}$$

для нечетных $N,\ K=\frac{N-1}{2},\ n=1,\ldots,K$:

$$\begin{cases} \cos(\varphi_n)+\cos(\varphi_{N-n})+\sum\limits_{m=1}^{N-n-1}\cos(\varphi_m-\varphi_{m+n})+ \\ \qquad\qquad +\sum\limits_{m=1}^{n-1}\cos(\varphi_m-\varphi_{m+N-n})=a; \\ \sin(\varphi_n)-\sin(\varphi_{N-n})-\sum\limits_{m=1}^{N-n-1}\sin(\varphi_m-\varphi_{m+n})+ \\ \qquad\qquad +\sum\limits_{m=1}^{n-1}\sin(\varphi_m-\varphi_{m+N-n})=0. \end{cases} \tag{4}$$

В такой постановке задачи синтез ФК последовательностей с заданным уровнем боковых лепестков циклической АКФ сводится к поиску решений системы уравнений (3), (4), где неизвестными являются углы поворотов элементов кода $\varphi_1,\varphi_2,\ldots,\varphi_{N-1}$, а произвольное решение системы уравнений будет иметь вид:

$$\boldsymbol{\Psi}=\left[\varphi_0=0^\circ,\ \varphi_1,\ \varphi_2,\ \ldots,\ \varphi_{N-1}\right]. \tag{5}$$

Отметим, что известные методы синтеза фазокодированных последовательностей позволяют синтезировать коды лишь с определенными значениями уровня боковых лепестков ($a=0$ или $a=-1$), кроме того, значения фаз всегда кратны π/N. Предлагаемый алгоритм позволяет синтезировать кодовые последовательности без этих ограничений.

Алгоритм синтеза ФК последовательностей

Алгоритм основан на развитом в работах [5, 6, 7, 8] методе решения системы уравнений (3), (4). Отметим, что всё множество получаемых кодовых последовательностей можно разбить на три класса.

1. Решения [6, 7], получаемые из «базисных» решений вида

$$\varphi_n=n^2(\mathrm{mod}\ N),\quad n=1,\ldots,N. \tag{6}$$

2. Решения [5], основанные на разностных множествах $D(N,k,\lambda) = \{d_1, d_2, \ldots, d_k\}$, состоящих из k различных элементов $d_1, d_2, \ldots, d_k \in G$ группы G порядка N, λ — параметр разностного множества, определяющих число упорядоченных пар (d_i, d_j), где $d_i, d_j \in D$, таких что $d_i * d_j^{-1} = d$ или $d_i^{-1} * d_j = d$, $d \in G$, $d \neq e$.

Если на позициях вектора (5) с порядковыми номерами $d_i \in D$ разместить значения фаз, равных 0, а на остальных $N-k$ позициях — значения фаз, равных некоторому значению α, то получим N-позиционную кодовую последовательность с одноуровневой циклической АКФ. Решение системы уравнений (3), (4) в этом случае будет иметь вид:

$$\mathbf{\Psi} = \left[\varphi_0 = 0^\circ,\ \alpha,\ \varphi_{d_2} = 0^\circ,\ \alpha,\ \alpha,\ \ldots,\ \varphi_{d_k} = 0^\circ,\ \alpha,\ \alpha\right]. \tag{7}$$

Угол α определится как:

$$\alpha = \pi \pm \arccos\left(\frac{N^2 + 2k^2 - 2kN + a - N - Na}{2k(N-k)}\right), \tag{8}$$

где a — уровень боковых лепестков из диапазона $a \in [a_1, N]$, где $a_1 = \frac{N^2+4k^2-4kN-N}{N-1}$.

3. Решения [8] общего вида (5), не формируемые ни на основе «базисных» решений, ни на основе разностных множеств. Например, в случае размерности кодовой последовательности $N = p - 1$, где p — простое число, решения системы уравнений (3) будут иметь следующий вид:

$$\begin{aligned} &\varphi_0 = 0^o, \quad \varphi_1 = \tfrac{360^\circ}{p}; \\ &v_1 = 1, \quad v_{n+1} = \theta_l v_n + 1 \ (\mathrm{mod}\ p), \quad v_n = v_{n+1}; \\ &\varphi_{l,n+1} = \tfrac{360^\circ}{p} v_n, \quad n = 1, \ldots, p-2, \quad l = 0, \ldots, L-1; \end{aligned} \tag{9}$$

где $L = \varphi(p-1)$ — функция Эйлера.

Алгоритм синтеза ФК последовательностей с заданным уровнем боковых лепестков реализует выше рассмотренные решения с использованием алгоритмических методов теории конечных полей и теории чисел.

Работа выполнена при поддержке РФФИ, проект №07-07-00285, и гранта Президента РФ МД-63.2007.9.

Литература

[1] *Кук Ч., Бернфельд М.* Радиолокационные сигналы. Теория и применение. — Москва: Сов. радио, 1971.

[2] *Свердлик М. Б.* Оптимальные дискретные сигналы — Москва: Сов. радио, 1975.

[3] *Варакин Л. Е.* Системы связи с шумоподобными сигналами — Москва: Радио и связь, 1985.

[4] *Гантмахер В. Е., Быстров Н. Е., Чеботарев Д. В.* Шумоподобные сигналы, анализ, синтез, обработка — СПб: Наука и техника, 2005.

[5] *Leukhin A. N.* Algebraic solution of the synthesis problem for coded sequences // Quantum Electronics. — 2005. — V. 35, № 8. — P. 688–692.

[6] *Леухин А. Н.* Полное решение задачи синтеза фазокодированных сигналов с равномерным энергетическим спектром // Известия Самарского научного центра РАН. — 2005. — Т. 7, № 1. — С. 163–168.

[7] *Леухин А. Н., Тюкаев А. Ю., Бахтин С. А.* Синтез и анализ сложных фазокодированных последовательностей // Электромагнитные волны и электронные системы. — 2007. — № 4. — С. 32–37.

[8] *Леухин А. Н., Тюкаев А. Ю., Бахтин С. А., Корнилова Л. Г.* Новые фазокодированные последовательности с хорошими корреляционными характеристиками // Электромагнитные волны и электронные с-мы. — 2007. — № 6.

Исследование автокорреляционных функций ортогональных фазокодированных последовательностей

Леухин А. Н., Тюкаев А. Ю.

inf@marstu.mari.ru

Йошкар-Ола, ГОУ ВПО Марийский гос. тех. университет

Исследованы взаимнокорреляционные свойства фазокодированных дискретных последовательностей, обладающих нулевым уровнем боковых лепестков циклической автокорреляционной функции. Дана сравнительная оценка автокорреляционных свойств ортогональных и квазиортогональных в широком смысле фазокодированных дискретных последовательностей.

Введение

Сигналы, имеющие циклическую автокорреляёционную функцию (АКФ) с нулевым уровнем боковых лепестков, идеальны для решения таких задач радиолокации как обнаружение, разрешение и оценка параметров [1, 2].

В то же время, задача распознавания наилучшим образом решается с применением ортогональных в широком смысле сигналов, то есть таких сигналов, у которых циклическая взаимная корреляционная функция (ВКФ) равномерна и имеет нулевой уровень отсчётов.

С развитием цифровой техники всё большее значение стали приобретать дискретные сигналы, которые можно различать по законам модуляции. Особое место среди дискретных сигналов занимают фазокодированные дискретные последовательности $\mathbf{\Gamma} = \{\gamma_n\}_{0,N-1}$ (ФКП), которые

можно определить на основании выражения:

$$\gamma_n = \exp\left(i\varphi_n\right), \quad n = 0,\ 1\ldots, N-1, \tag{1}$$

где значение фазы на каждом n-ом кодовом интервале определяется из диапазона $\varphi_n \in [0,\ 2\pi]$, модуль каждого кодового элемента $|\gamma_n| = 1,\ N$ — количество кодовых элементов в последовательности, i — мнимая единица.

Синтез и анализ ФКП с хорошими корреляционными свойствами является важной задачей теории синтеза сигналов.

Ортогональные системы в широком смысле

Нормированную циклическую ВКФ двух ФКП $\mathbf{\Gamma} = \{\gamma_n\}_{0,N-1}$ и $\mathbf{N} = \{\nu_n\}_{0,N-1}$ размерностью N определим на основании выражения:

$$\eta_\tau = \frac{1}{N}\sum_{n=0}^{N-1} \gamma_{n+\tau(\mathrm{mod}\ N)} \cdot \nu_n^*, \quad \tau = 0, 1, \ldots, N-1, \tag{2}$$

Последовательности $\mathbf{\Gamma} = \{\gamma_n\}_{0,N-1}$ и $\mathbf{N} = \{\nu_n\}_{0,N-1}$ назовем *ортогональными в широком смысле*, если все отсчеты их нормированной ВКФ равны нулю. Семейство всех взаимноортогональных ФКП размерности N назовем *ортогональным алфавитом*, а количество элементов алфавита (объем) обозначим через L.

Примером известных ортогональных в широком смысле ФКП являются базисные функции дискретного преобразования Фурье (элементарные контуры) [3] и функции Радемахера [4]. Семейство всех элементарных контуров размерности N образует алфавит ортогональных символов объемом $L = N$. Система всех функций Радемахера с порядком k и размерностью $N = 2^k$ также образует алфавит ортогональных символов объемом $L = k$.

Нормированную циклическую автокорреляционную функцию дискретной последовательности $\mathbf{\Gamma} = \{\gamma_n\}_{0,N-1}$ определим на основании выражения:

$$r_\tau = \frac{1}{N}\sum_{n=0}^{N-1} \gamma_{n+\tau(\mathrm{mod}\ N)} \cdot \gamma_n^*, \quad \tau = 0, 1, \ldots, N-1, \tag{3}$$

где γ_n^* — комплексно сопряженный кодовый элемент дискретной последовательности $\mathbf{\Gamma} = \{\gamma_n\}_{0,N-1}$.

Квазиортогональные системы в широком смысле

Введём понятие квазиортогональных в широком смысле фазокодированных дискретных последовательностей. Для любых двух квазиортогональных в широком смысле ФКП $\mathbf{\Gamma} = \{\gamma_n\}_{0,N-1}$ и $\mathbf{N} = \{\nu_n\}_{0,N-1}$

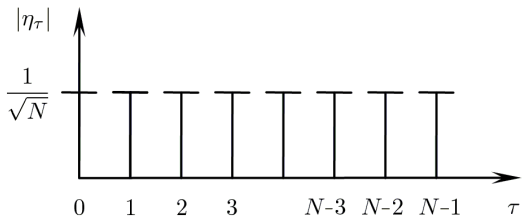


Рис. 1. Примерный вид нормированной циклической ВКФ синтезированных квазиортогональных ФКП.

должно выполняться равенство:

$$|\eta_\tau| = \frac{1}{N}\left|\sum_{n=0}^{N-1} \gamma_{n+\tau(\mathrm{mod}\ N)} \cdot \nu_n^*\right| = c, \quad \tau = 0, 1, \ldots, N-1. \tag{4}$$

Равенство (4) должно выполняться при условии: $c \ll N$, где c — некоторое неотрицательное вещественное число. Семейство всех возможных взаимноквазиортогональных ФКП размерности N назовем *квазиортогональным алфавитом*, а количество элементов алфавита (объем) обозначим через L.

В работе [5] разработан метод синтеза ФКП, позволяющий получить все возможные дискретные кодовые последовательности, нормированная циклическая АКФ которых имеет нулевой уровень боковых лепестков. Исследования показали, что синтезированные в работе [5] ФКП могут обладать равномерной нормированной циклической ВКФ с уровнем модулей отсчётов равным $\frac{1}{\sqrt{N}}$ (Рис. 1), в случае, когда размерность дискретных последовательностей N является нечётным числом [6], т. е.:

$$|\eta_\tau| = \frac{1}{\sqrt{N}}, \quad \text{при нечётном } N, \quad \tau = 0, 1, \ldots, N-1, \tag{5}$$

где $|\eta_\tau|$ — 0 1 2 3 N−3 N−2 N−1 модуль нормированной циклической ВКФ, N — размерность сигнала, τ — временной сдвиг.

Заключение

Синтезированные в работе [6] фазокодированные дискретные последовательности, в отличие от ортогональных сигналов (элементарные контуры и функции Радемахера), могут обладать равномерной нормированной циклической ВКФ с уровнем модулей отсчётов равным $\frac{1}{\sqrt{N}}$ (Рис. 1), в том случае, если размерность N данных фазокодированных последовательностей — нечётное число. При больших значениях размерности N та-

кие последовательности можно считать квазиортогональными, т. к. уровень модулей отсчётов их нормированной циклической ВКФ будет стремиться к нулю, т. е. $\lim\limits_{N\to\infty} \frac{1}{\sqrt{N}} \to 0$. Циклическая АКФ таких фазокодированных последовательностей, в отличие от ортогональных сигналов, при любом значении размерности N обладает идеальными свойствами, т. е. нулевым уровнем боковых лепестков.

Работа выполнена при поддержке РФФИ, проект № 07-07-00285 и гранта Президента РФ МД-63.2007.9.

Литература

[1] *Woodward P. M.* Probability and Information Theory with Applications to Radar. — Pergamon Press, N.Y., 1953.

[2] *Кук Ч., Бернфельд М.* Радиолокационные сигналы. Теория и применение. — Москва: Сов. радио, 1971.

[3] Введение в контурный анализ и его приложения к обработке изображений и сигналов / под ред. *Фурмана Я. А.* — Москва: Физматлит, 2002.

[4] *Гоноровский И. С.* Радиотехнические цепи и сигналы — М: Радио и связь, 1986.

[5] *Leukhin A. N.* Algebraic solution of the synthesis problem for coded sequences // Quantum Electronics. — 2005. — V. 35, No. 8. — Pp. 688–692.

[6] *Леухин А. Н., Тюкаев А. Ю., Бахтин С. А.* Синтез и анализ сложных фазокодированных последовательностей // Электромагнитные волны и электронные системы. — 2007. — № 4. — С. 32–37.

Синтез фазокодированных дискретных последовательностей системы Гаусса, образующих квазиортогональный алфавит

Леухин А. Н., Тюкаев А. Ю.

inf@marstu.mari.ru

Йошкар-Ола, ГОУ ВПО Марийский гос. тех. университет

Разработан регулярный метод синтеза алфавита квазиортогональных в широком смысле фазокодированных дискретных последовательностей с идеальными свойствами циклической автокорреляционной функции.

Введение

Особый интерес среди синтезируемых кодовых последовательностей с хорошими корреляционными характеристиками представляют фазокодированные дискретные последовательности $\mathbf{\Gamma} = \{\gamma_n\}_{0,N-1}$ (ФКП), обладающие идеальными свойствами циклической автокорреляционной функции (АКФ) т. е. нулевым уровнем боковых лепестков циклической

АКФ, которую можно определить на основании выражения:

$$r_\tau = \sum_{n=0}^{N-1} \gamma_{n+\tau(\mathrm{mod}\ N)} \cdot \gamma_n^*, \quad \tau = 0, 1, \ldots, N-1, \tag{1}$$

где N — количество кодовых элементов в последовательности, γ_n^* — комплексно сопряженный кодовый элемент ФКП $\mathbf{\Gamma} = \{\gamma_n\}_{0,N-1}$, которую можно представить в следующем виде:

$$\gamma_n = \exp(i\varphi_n), \quad n = 0, 1, \ldots, N-1, \tag{2}$$

где значение фазы на каждом n-ом кодовом интервале определяется из диапазона $\varphi_n \in [0, 2\pi]$, модуль каждого кодового элемента $|\gamma_n| = 1$, N — количество кодовых элементов в ФКП, i — мнимая единица.

В работе [1] разработан метод синтеза ФКП, позволяющий получить все возможные ФКП заданной размерности N с нулевым уровнем боковых лепестков циклической АКФ. Важное прикладное значение имеют не только ФКП, обладающие идеальными свойствами циклической АКФ, но и ортогональные сигналы, т. е. такие сигналы, у которых циклическая взаимная корреляционная функция (ВКФ) равномерна и имеет нулевой уровень отсчётов. Исследования показали, что синтезированные в работе [1] ФКП с идеальными свойствами циклической АКФ могут обладать равномерной нормированной циклической ВКФ с уровнем модулей отсчётов равным $\frac{1}{\sqrt{N}}$, в том случае, если размерность N данных ФКП — нечётное число [2]. При больших значениях N такие последовательности можно считать квазиортогональными, т. к. уровень модулей отсчётов их нормированной циклической ВКФ будет стремиться к нулю, т. е. $\lim\limits_{N\to\infty} \frac{1}{\sqrt{N}} \to 0$. Таким образом, данные дискретные последовательности образуют алфавит квазиортогональных ФКП. Поэтому актуальной является задача поиска ортогональных сигналов из общего объёма ФКП, синтезированных в работе [1].

Синтез алфавита квазиортогональных фазокодированных последовательностей системы Гаусса

Задача получения алфавита квазиортогональных ФКП заданной размерности N из общего объёма дискретных последовательностей, синтезированных в работе [1], сводится к нахождению таких ФКП $\mathbf{\Gamma}$ и $\mathbf{N}$, для которых выполняется условие:

$$|\eta_\tau| = \frac{1}{N} \left| \sum_{n=0}^{N-1} \gamma_{n+\tau\ (\mathrm{mod}\ N)} \cdot (\nu_n^*) \right| = \frac{1}{\sqrt{N}}, \quad \tau = 0, 1, \ldots, N-1, \tag{3}$$

где $\mathbf{\Gamma} = \{\gamma_n\}_{0,N-1}$ и $\mathbf{N} = \{\nu_n\}_{0,N-1}$ — ФКП, принадлежащие одному алфавиту, N — размерность ФКП, $|\eta_\tau|$ — модуль нормированной циклической ВКФ.

На основании (2) выражение (3) можно записать в следующем виде:

$$|\eta_\tau| = \frac{1}{N}\left|\sum_{n=0}^{N-1}\exp\left(i\left(\varphi^{(j)}_{n+\tau(\mathrm{mod}\ N)} - \varphi^{(l)}_n\right)\right)\right| = \frac{1}{\sqrt{N}}, \tag{4}$$

где $\tau = 0, 1, \ldots, N-1,\ n = 0, 1, \ldots, N-1,\ \varphi^{(j)}_n$ и $\varphi^{(l)}_n$ — значения фаз ФКП $\mathbf{\Gamma} = \{\gamma_n\}_{0,N-1}$ и $\mathbf{N} = \{\nu_n\}_{0,N-1}$ соответственно.

Все возможные ФКП системы Гаусса с идеальными свойствами циклической АКФ можно представить в виде [2]:

$$\mathbf{\Gamma}^{(l)} = \left\{\exp\left(i\frac{2\pi}{N_1}\lambda_l n^2\right)\right\}_{0,N-1} \tag{5}$$

где $N_1 = 2N$ для $N\,(\mathrm{mod}2) \equiv 0$ и $N_1 = N$ для $N\,(\mathrm{mod}2) \equiv 1,\ \lambda_l$ — вычеты по модулю N, взаимно простые с $N,\ l = 0, 1, \ldots, \varphi(N) - 1,\ \varphi(N)$ — функция Эйлера от числа N.

На основании (4) и (5) выражение для модуля циклической ВКФ двух ФКП системы Гаусса представим в следующем виде:

$$\begin{aligned}|\eta_\tau| &= \frac{1}{N}\left|\sum_{n=0}^{N-1}\exp\left(i\frac{2\pi}{N}\lambda_j(n+\tau)^2\right)\exp\left(-i\frac{2\pi}{N}\lambda_l n^2\right)\right| = \\ &= \frac{1}{N}\left|\sum_{n=0}^{N-1}\exp\left(i\frac{2\pi}{N}\left((\lambda_j - \lambda_l)n^2 + 2\lambda_j\tau n\right)\right)\right|,\end{aligned} \tag{6}$$

где λ_j и λ_l — числа, взаимно простые с $N,\ \tau = 0, 1, \ldots, N-1,\ N$ — нечетное число.

Пусть $a_1 = 2\lambda_j\tau,\ a_2 = \lambda_j - \lambda_l$. При нечетном N и a_2 взаимно простом с N выполняется равенство [3]:

$$\left|\sum_{n=0}^{N-1}\exp\left(i\frac{2\pi}{N}\left[a_1 n + a_2 n^2\right]\right)\right| = \sqrt{N}. \tag{7}$$

Выражение (7) при a_2, взаимно простом с N, является полной тригонометрической суммой Гаусса $S(N)$. Для модуля суммы Гаусса выполняется равенство [3]:

$$|S(N)| = \begin{cases}\sqrt{N}, & \text{если } N \equiv 1\ (\mathrm{mod}2);\\ \sqrt{2N}, & \text{если } N \equiv 0\ (\mathrm{mod}4);\\ 0, & \text{если } N \equiv 2\ (\mathrm{mod}4).\end{cases} \tag{8}$$

Таким образом, нормированная циклическая ВКФ двух ФКП, задаваемых с помощью выражения (5), будет квазиортогональной и равной $|\eta_\tau| = \frac{1}{\sqrt{N}}$.

Алгоритм синтеза всех возможных квазиортогональных алфавитов системы Гаусса можно представить в следующем виде:

1. Определяется система вычетов по модулю нечетного числа N, взаимно простых с N, $\left\{\lambda_0, \lambda_1, \ldots, \lambda_{\varphi(N)-1}\right\}$, где $\varphi(N)$ — функция Эйлера.
2. Определяется наименьшее число p_1 в разложении $\left\{\lambda_0, \lambda_1, \ldots, \lambda_{\varphi(N)-1}\right\}$ числа N.
3. Среди всех $C_{\varphi(N)}^{p_1-1}$ сочетаний по p_1-1 вычетов по модулю N, взаимно простых с N, из $\varphi(N)$ возможных вычетов отбираются k-ые сочетания вычетов $\left\{\lambda_0^{(k)}, \lambda_1^{(k)}, \ldots, \lambda_{p_1-2}^{(k)}\right\}$, которые удовлетворяют условию: $\lambda_j^{(k)} - \lambda_l^{(k)} \equiv 1 \,(\mathrm{mod} N),\ j, l = 0, 1, \ldots, p_1 - 2,\ j \neq l$.
4. Полная система ФКП заданной размерности N, образующих искомый квазиортогональный алфавит имеет вид:

$$\mathbf{\Gamma}^{(l)} = \left\{\exp\left(i\frac{2\pi}{N}\lambda_l^{(k)} n^2\right)\right\}_{0,N-1}, \tag{9}$$

где $l = 0, 1, \ldots, p_1 - 2,\ n = 0, 1, \ldots, N-1$.

Заключение

В работе предложен метод формирования алфавита квазиортогональных последовательностей системы Гаусса. Цклическая автокорреляционная функция каждого символа этого алфавита имеет нулевые боковые лепестки.

Работа выполнена при поддержке РФФИ, проект №07-07-00285 и гранта Президента РФ МД-63.2007.9.

Литература

[1] *Leukhin A.N.* Algebraic solution of the synthesis problem for coded sequences. — // Quantum Electronics. – 2005. – V.35, № 8. – p. 688 - 692 35.

[2] *Леухин А.Н., Тюкаев А.Ю., Бахтин С.А.* Синтез и анализ сложных фазокодированных последовательностей. — // Электромагнитные волны и электронные системы 2007. №4. – с. 32 – 37.

[3] *Коробов Н.М.* Тригонометрические суммы и их приложения. — М.: Наука, 1989. – 240 с.

Сокращение размерности пространства спектральных признаков в многоклассовой задаче распознавания сигналов

Манило Л. А., Немирко А. П.
APNemirko@mail.eltech.ru
Санкт-Петербург, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

Распознавание сигналов в частотной области, как правило, основано на анализе спектральных признаков, получаемых при вычислении функции спектральной плотности мощности (СПМ). Это описание достаточно полно отражает частотные свойства представленных групп сигналов, но приводит к необходимости построения дискриминантных функций в пространстве большой размерности. Снизить размерность признакового пространства можно путем формирования упорядоченного набора признаков на основе группировки спектральных коэффициентов, а также отображения полученного описания в пространство меньшей размерности с применением множественного дискриминантного анализа. Основой для построения решающих функций является анализ линейного дискриминанта Фишера J, максимизация которого приводит к выбору наилучшего для разделения c классов сигналов набора $(c-1)$ векторов [1]. Однако не всегда его оптимизация обеспечивает надежное распознавание сигналов.

Критерий J, оценивающий степень разделения исходных классов сигналов, можно представить скалярной величиной, задаваемой следом матрицы в виде:

$$J = \mathrm{tr}(\mathbf{S}_2^{-1}\mathbf{S}_1), \tag{1}$$

где $\mathbf{S}_1$ — матрица рассеяния между классами; $\mathbf{S}_2$ — обобщенная матрица рассеяния внутри классов.

В случае c классов проекции объектов при переходе из L-мерного пространства сформированных спектральных признаков $\mathbf{G} = (G_1, G_2, \ldots, G_L)^{\mathrm{T}}$ в $(c-1)$-мерное пространство могут быть найдены с помощью матричного преобразования $\mathbf{Y} = \mathbf{W}^{\mathrm{T}}\mathbf{G}$, где $\mathbf{W}$ — матрица размера $L \times (c-1)$, нахождение которой сопряжено с максимизацией J. Недостаток применения выражения (1) связан с тем, что при увеличении числа классов J становится индикатором больших межгрупповых расстояний и слабо отражает взаимное расположение близко расположенных в частотной области классов.

Оптимизация построения решающих правил

Оптимизировать процедуру построения решающих правил можно путем сведения ее к набору задач попарной классификации с введением

весовых коэффициентов $a_{i,j}$, усиливающих влияние на критерий J близко расположенных классов. В этом случае обобщенное выражение для критерия J принимает вид:

$$J = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} n_i n_j a_{i,j} \ \mathrm{tr}\left(\left(\mathbf{W}^T \mathbf{S}_2 \mathbf{W}\right)^{-1} \left(\mathbf{W}^T \mathbf{S}_1^{(i,j)} \mathbf{W}\right)\right), \qquad (2)$$

где n_i и n_j — частота появления объектов, образующих классы ω_i и ω_j.

Весовую функцию $a_{i,j}$ можно связать с ценой ошибки распознавания каждой пары классов ω_i и ω_j. В работе [2] предлагается использовать веса в виде некоторого представления функции ошибок $\mathrm{erf}\left(\frac{\eta - t}{\sigma}\right)$, где t — граница решающего правила, а η и σ — параметры распределений, вычисляемые для заданных групп объектов, исходя из предположений о нормальном законе распределений с равными ковариационными матрицами. Этот подход представляется эффективным, поскольку критерий J может быть приближен к оценке достоверности распознавания объектов путем суммирования вероятностей правильного решения при попарной классификации. В данной работе развивается идея приближения критерия J к оценке точности классификации объектов в пространстве спектральных признаков, представленных нормированными значениями СПМ. Задачу нахождения $a_{i,j}$ в (2) предлагается решить следующим образом.

Метод вычисления весовых функций

Рассмотрим в двумерном пространстве (x_1, x_2) два класса объектов ω_i и ω_j с нормальным законом распределения и единичными матрицами ковариации. Если расстояние между центрами этих классов обозначить как $\Delta_{i,j} = \|\mathbf{m}_i - \mathbf{m}_j\|$, где $\mathbf{m}_i$ и $\mathbf{m}_j$ — векторы средних значений, то, проецируя эти классы на новое направление $\mathbf{V}$, расстояние между ними будет изменяться в зависимости от угла α между направлением, соединяющим центры классов, и вектором $\mathbf{V}$. Эту зависимость можно представить как $\Delta_{i,j}^{(\nu)} = \Delta_{i,j} \cos\alpha$. При равных априорных вероятностях появления объектов обоих классов вероятность правильного распознавания определится в виде:

$$\gamma_{i,j} = \frac{1}{2} + \gamma_{i,j}^{'} = \frac{1}{2} + \frac{1}{2} \,\mathrm{erf}\left(\frac{\Delta_{i,j}^{(\nu)}}{2\sqrt{2}}\right),$$

где $\mathrm{erf}(\cdot)$ — функция ошибок.

Тогда для случая c классов с идентичными распределениями критерий $J^{(\gamma)}$, оценивающий среднюю точность распознавания, можно представить в виде:

$$J^{(\gamma)} = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} n_i n_j \gamma_{i,j}, \tag{3}$$

а критерий J, оценивающий степень расхождения классов, примет вид:

$$J = \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} n_i n_j a_{i,j} \operatorname{tr}\left(\mathbf{V}^T \mathbf{S}_1^{(i,j)} \mathbf{V}\right). \tag{4}$$

Сравнивая (3) и (4), можно предложить аппроксимацию переменной составляющей $J^{(\gamma)}$ выражением, используемым для нахождения J (4), задав веса $a_{i,j}$ в виде:

$$a_{i,j} = \frac{\gamma_{i,j}^{'}}{tr\left(\mathbf{V}^T \mathbf{S}_1^{(i,j)} \mathbf{V}\right)}$$

для случая наилучшего взаимного расположения двух классов (ω_i, ω_j), что соответствует случаю совпадения направлений векторов $\mathbf{V}$ и $\mathbf{m}_{i,j} = = (\mathbf{m}_i - \mathbf{m}_j)$. При этом $\alpha = 0$; $\operatorname{tr}\left(\mathbf{V}^T \mathbf{S}_1^{(i,j)} \mathbf{V}\right) = (\Delta_{i,j})^2$, а параметр $a_{i,j}$ в области малых значений $\Delta_{i,j}$, используя приближение функции ошибок полиномиальной функцией, можно задать в виде:

$$a_{i,j} \approx \frac{1}{8\sqrt{\pi} x_{i,j}} \left(1 - \frac{x_{i,j}^2}{3} + \frac{x_{i,j}^4}{2!\,5}\right), \tag{5}$$

где $x_{i,j} = \left(\dfrac{\Delta_{i,j}}{2\sqrt{2}}\right)$, $\Delta_{i,j} \leqslant \sqrt{2}$, $x_{i,j} \leqslant 0,5$.

Этот способ нахождения весовых функций $a_{i,j} = a(\Delta_{i,j})$ можно применить и для многоклассовой задачи, исходя из предположения, что каждый из c классов имеет матрицу внутригруппового рассеяния, задаваемую обобщенной матрицей разброса $\mathbf{S}_2 = \sum_{i=1}^{c} n_i \Sigma_i$, где Σ_i — выборочная ковариационная матрица i-го класса. Тогда для каждой пары классов в исходном L-мерном пространстве спектральных признаков необходимо найти евклидово расстояние между центрами соответствующих классов и определить веса $a_{i,j}$, используя выражение (5). Максимизация критерия (2) приводит к процедуре нахождения собственных векторов и анализу распределений групп объектов в пространстве признаков пониженной размерности.

Применение для распознавания опасных аритмий

Эффективность введения весовых функций (5) оценивалась по результатам экспериментов, выполненных на реальных электрокардиосиг-

налах (ЭКС). Решалась задача распознавания 3-х классов опасных аритмий. В качестве исходного описания объектов, представленных фрагментами ЭКС длительностью 2 с, использован упорядоченный набор 28 спектральных признаков, полученных в частотной области, ограниченной 15 Гц, с применением перекрывающихся сегментов [3]. Некоррелированные оценки СПМ получены с шагом $\Delta f = 0,976$ Гц, но при этом шаг по частотной оси выбран вдвое меньше этой величины и составлял 0,488 Гц. В этом случае удается сохранить особенности формы спектра анализируемых сигналов при относительной устойчивости получаемых оценок СПМ. В ходе экспериментов были построены разделяющие функции, определены границы областей решений и найдены ошибки классификации, являющиеся критерием надежности распознавания. Как показал результат линейного дискриминантного анализа, в этом случае средняя ошибка классификации может быть уменьшена с 8,2% до 4,6%, что является показателем эффективности применения этой процедуры оптимизации.

Работа выполнена при поддержке РФФИ, проекты № 06-01-00546, № 07-01-00569.

Литература

[1] *Дуда З., Харт П.* Распознавание образов и анализ сцен: Пер. с англ. М.: Мир, 1976. — 511 с.

[2] *Loog M., Duin R. P. W., Haeb-Umbach R.* Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria // IEEE Trans. on Pattern Analysis and Machine Intelligence. — 2001. — Vol. 23, № 7. — Pp. 762–766.

[3] *Манило Л. А.* Упорядочение спектральных признаков по эмпирическим оценкам межгруппового расстояния в задачах классификации биосигналов // Изв. вузов России. Радиоэлектроника. — 2006. — Вып. 3, — С. 20–29.

Восстановление в реальном времени пространственных характеристик гибкого объекта по стереопаре изображений

Местецкий Л. М., Цискаридзе А. К.
l.mest@ru.net, AchikoTsi@gmail.com

Рассматривается задача восстановления пространственных характеристик объектов с осесимметрическими элементами. Предлагается подход, основанный на концепции пространственного гибкого объекта. Пространственный гибкий объект определяется как семейство эллипсоидов с центрами на графе древовидной структуры. Предлагается метод идентификации пространственного гибкого объекта в реальном времени по стереопаре изображений силуэтов объекта. Метод основан на по-

строении непрерывных скелетов силуэтов. Рассматривается приложение к задаче распознавания жестов в реальном масштабе времени.

Постановка задачи

Задача восстановления формы пространственного объекта по нескольким двумерным изображениям хорошо известна и имеет множество приложений. Особенность рассматриваемой нами постановки этой задачи состоит в том, что двумерные изображения являются бинарными и представляют собой лишь силуэты пространственного объекта. Такая задача, в частности, возникает при распознавания жестов с помощью стандартного недорогого оборудования. Исходными данными служат изображения низкого разрешения формата 480×640, полученные с использованием обычных WEB-камер. Такие камеры плохо передают текстурные особенности изображений и позволяют с достоверностью выявить лишь силуэты представленных на изображении объектов. Для распознавания жеста требуется восстановить пространственную форму столь сложного и изменчивого объекта, как человеческая ладонь. Актуальность такой постановки задачи объясняется тем, что круг потенциальных пользователей систем распознания жестов включает в себя большое число людей, не способных приобрести дорогое оборудование, но весьма нуждающихся в системах, способных понимать жесты. Речь идёт об инвалидах, слабослышащих. Известны работы, целью которых является создание систем, понимающих азбуку глухонемых [4]. Также известны разработки соответствующих систем управления компьютером с помощью жестов [5].

Невозможность анализа изображений на уровне текстур не позволяет применить для решения задачи хорошо известные методы восстановления формы объектов по стереопаре изображений, основанные на автоматическом выявлении общих точек, присутствующих на обоих изображениях. Очевидно, что если на изображении представлен лишь силуэт объекта, то достоверно на нём можно выявить лишь точки границы этого объекта. Но на двух картинках в стереопаре изображений силуэты полностью различаются, т. е. все точки границы одного силуэта отличаются от всех граничных точек другого силуэта.

Предлагаемый подход к решению задачи основан на идее выявления в структуре объекта таких точек, которые, хотя и не видны на изображениях стереопары, но могут быть вычислены на каждом изображении в результате анализа представленного на нём силуэта. Таким образом, в роли общих реперных точек стереопары предлагается использовать некоторые (невидимые) точки, не являющиеся граничными точками силуэтов. В качестве такого множества реперных точек предлагается рассмотреть множество серединных осей силуэтного изображения, составляющих его скелет.

Реализация предлагаемого подхода ставит несколько сложных задач. Во-первых, нужно построить скелеты силуэтов в такой форме, которая позволит идентифицировать точки разных скелетов. Во-вторых, нужно по результатам идентификации пары скелетов восстановить пространственную форму всей ладони. В-третьих, нужно обеспечить регистрацию динамических изменений формы ладони с целью распознавания жестов. Особо следует отметить, что решение этих задач должно осуществляться в рамках системы машинного зрения в реальном времени работы этой системы, т. е. требуемая скорость обработки должна составить несколько стереопар изображений в секунду. Это предъявляет высокие требования к вычислительной эффективности разрабатываемых алгоритмов.

Метод решения

В работе [1] введено понятие плоского гибкого объекта и предложен эффективный метод сравнения гибких объектов на основе гранично-скелетной модели. В настоящей работе предлагается обобщение плоского гибкого объекта на пространственный случай.

Пространственный гибкий объект определяется как семейство эллипсоидов различной формы с центрами на графе древовидной структуры. Анализ стереопары позволяет воссоздать пространственную структуру объекта. Восстановление пространственных характеристик объекта даёт возможность отслеживать динамику перемещения объекта, а также изменение его формы. В частности, применительно к ладони это позволяет отслеживать жесты.

Реализация данного подхода предполагает решение следующих задач:

1. Калибровка камер.
2. Выделение силуэтов ладоней на изображениях.
3. Построение непрерывного скелета для силуэта ладони.
4. Идентификация реперных точек на скелетах.
5. Трёхмерная визуализация.
6. Оценка параметров пространственного положения ладони (жеста).

Калибровка камер [3] включает: съемку двумя камерами эталонного объекта — картонной модели «куба» с нанесённой на грани прямоугольной сеткой, Рис. 1; ручную идентификацию точек на полученных снимках; вычисление коэффициентов для определения в дальнейшем пространственных координат произвольных точек на стереопарах изображений.

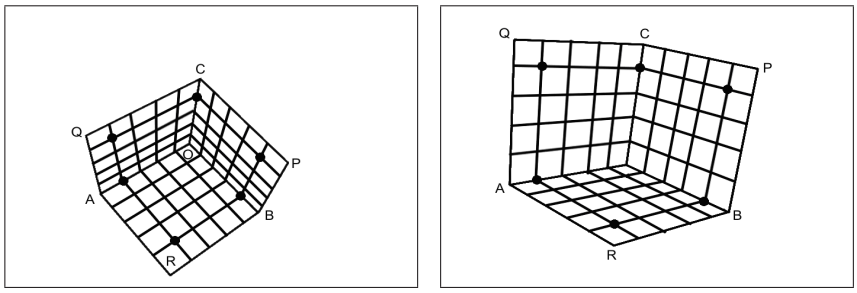


Рис. 1. Стереопара калибровочного куба.

Получение стереопары бинарных изображений. Каждое изображение отдельно сегментируется, выделяется силуэт, который затем представляется в виде бинарного растрового изображения.

Построение непрерывных скелетов обоих силуэтных (бинарных) изображений осуществляется методом, описанным в [2]. Скелет представляет собой геометрическое место точек — центров вписанных в силуэт окружностей.

Идентификация точек скелетов. Предполагается, что на плоском изображении образ осей гибкого объекта совпадает со скелетом силуэта, что позволяет вычислить оси. Пусть C — некоторая точка на одном из плоских изображений, образующих стереопару. В пространстве ей соответствует прямая, которая проектируется в эту точку. Образ этой прямой на другом изображении является эпиполярной линией точки C. Для заданной точки на скелете её стереопара находится на пересечении другого скелета с эпиполярной линией этой точки, Рис. 2.

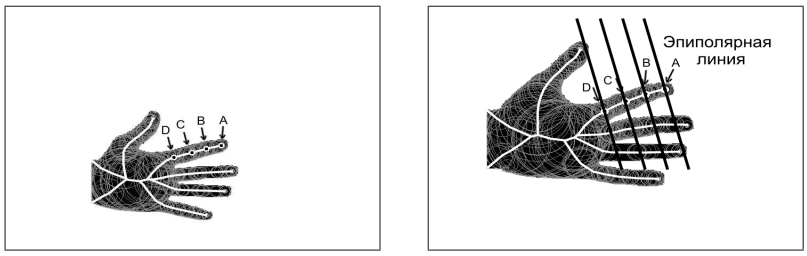


Рис. 2. На скелетах найдены одинаковые точки.

Построение пространственного гибкого объекта. Построив оси, можно вычислить формы эллипсоидов и восстановить пространственный образ гибкого объекта, Рис. 3.

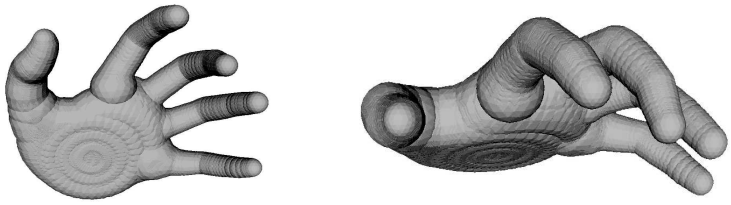


Рис. 3. Визуализация пространственного гибкого объекта.

Работа выполнена при поддержке РФФИ, проект № 05-01-00542.

Литература

[1] *Mestetskiy L.* Shape comparison of flexible objects // Int. Conf. on Computer Vision Theory and Applications, Barcelona, 2007.

[2] *Местецкий Л. М.* Непрерывный скелет бинарного растрового изображения // Труды межд. конф. Графикон–98. — Москва, 1998.

[3] *Форсайт Д. Понс Ж.* Компьютерное зрение. — Вильямс, 2004.

[4] *Burger T., Caplier A., Mancini S.* Cued speech hand gestures recognition tool // Int. Conf. on Computer Vision Theory and Applications, VISAPP 2007.

[5] *Keskin C., Aran O., Akarun L.* Real time gestural interface for generic applications // European Signal Processing Conference, EUSIPCO 2005.

Оптимизация псевдоградиента в задаче псевдоградиентного оценивания межкадровых геометрических деформаций изображений

Минкина Г. Л., Самойлов М. Ю., Дикарина Г. В., Захаров А. А.

tag@ulstu.ru

Ульяновск, Ульяновский государственный технический университет

Исследуются вопросы оптимизации объема локальной выборки отсчетов изображений, используемой для нахождения псевдоградиента целевой функции, в частности, влияние объема и плана локальной выборки отсчетов изображений на процесс сходимости оценок параметров и признаки локальных экстремумов оценки целевой функции. Предложены методики априорной и апостериорной оптимизации объема локальной выборки по различным критериям качества оценивания.

Оценивание параметров межкадровых геометрических деформаций изображений (МГДИ) является одной из ключевых задач при представлении и обработке последовательностей изображений [1]. При решении указанной задачи хорошо себя зарекомендовали псевдоградиентные про-

цедуры (ПГП) [2]. Формирование с помощью ПГП оценки вектора $\bar{\alpha}$ параметров МГДИ может быть описано соотношением

$$\hat{\bar{\alpha}}_{t+1} = \hat{\bar{\alpha}}_t - \mathbf{\Lambda}_{t+1}\bar{\beta}_{t+1}\left(Z_{t+1},\, \hat{\bar{\alpha}}_t\right), \tag{1}$$

где $\bar{\beta}$ — псевдоградиент целевой функции (ЦФ), характеризующей качество оценивания; $\mathbf{\Lambda}_t$ — матрица усиления, задающая приращение оценок параметров на итерации; Z_{t+1} — локальная выборка отсчетов наблюдаемых изображений, используемая для нахождения $\bar{\beta}$ на $(t+1)$-й итерации.

Недостатками ПГП (1) обработке изображений являются наличие локальных экстремумов оценки ЦФ и сравнительно небольшой рабочий диапазон. Поэтому весьма актуальной является оптимизация процедур по скорости сходимости и вычислительным затратам. При этом заметим, что характер сходимости оценок и вычислительные затраты во многом определяются объемом локальной выборки (ОЛВ) μ_t, используемым на итерациях процесса псевдоградиентного оценивания для нахождения псевдоградиента ЦФ. Тем не менее, вопросы оптимизации ОЛВ в литературе практически не исследованы.

В качестве исходной информации для нахождения скорости сходимости вектора оценок $\hat{\bar{\alpha}}$ исследуемых параметров $\bar{\alpha}$ к оптимальному значению $\bar{\alpha}^*$ представляется целесообразным использование плотности распределения вероятностей (ПРВ) этих оценок на соответствующих итерациях. Целесообразно исследовать различные величины, характеризующие скорость сходимости оценок: математическое ожидание, вероятность превышения порогового значения, доверительный интервал, и др. При оценивании одного параметра эти характеристики непосредственно применимы к его оценке. Если же оценивается совокупность параметров, то в общем случае, на одной и той же итерации для каждого i-го параметра может получиться свое значение ОЛВ μ_{it}, обеспечивающее выполнение заданного критерия. Это неприемлемо, поскольку на каждой итерации локальная выборка должна формироваться один раз. Соответственно и для критерия необходима единая мера. Для задачи оценивания МГДИ в качестве такой меры предлагается использовать плотность распределения вероятностей (ПРВ) $w_{t-1}(r)$ расстояний между одноименными точками изображений с опорного и деформированного кадров, вошедшими в локальную выборку Z_t. Это расстояние r между истинным положением точки и его оценкой назовем евклидовым расстоянием оценки (ЕРО). Разработана методика нахождения ПРВ ЕРО, основанная на предварительной дискретизации области определения параметров. В качестве величин, характеризующих скорость сходимости ЕРО на конкретной t-й итерации, исследовалось, например, математическое ожидание изменения оценки при ОЛВ $\mu = m$

$$\mathsf{E}\,[r]\big|_{\mu=m} = \int_0^\infty r\Big(w_{t-1}(r)\big|_{\mu=m} - w_t(r)\big|_{\mu=m}\Big)\,dr. \tag{2}$$

Аналогично (2) может быть найдено и матожидание $\mathsf{E}\,[\Delta r(+k)]$ улучшения вектора оценок параметров при увеличении ОЛВ μ на k:

$$\mathsf{E}\,[\Delta r(+k)] = \int_{-\infty}^\infty r\Big(w_t(r)\big|_{\mu=m+k} - w_t(r)\big|_{\mu=m}\Big)\,dr.$$

Кроме приведенного параметра, исследовалось использование в качестве меры скорости сходимости ЕРО условия превышения ею с заданной доверительной вероятностью некоторого порогового значения, доверительный интервал при заданной доверительной вероятности и ряд других параметров.

Для нахождения вероятностных свойств ЕРО и влияния на них ОЛВ требуется найти вероятностные характеристики изменения оценок параметров в процессе их сходимости. Это удается сделать, применив в качестве величины, позволяющей при заданной ЦФ качества оценивания комплексно характеризовать параметры исследуемых изображений и мешающих шумов, коэффициент

$$\mathfrak{R}_i = \rho_i^+\left(\bar{\varepsilon}\right) - \rho_i^-\left(\bar{\varepsilon}\right), \tag{3}$$

характеризующий улучшение оценки, где $\rho_i^-\left(\bar{\varepsilon}_t\right)$ — вероятность того, что изменение оценки i-го параметра направлено от оптимального значения, а $\rho_i^+\left(\bar{\varepsilon}_t\right)$ — к оптимальному, $\bar{\varepsilon} = \hat{\bar{\alpha}} - \bar{\alpha}^*$ — рассогласование оценки и оптимального значения.

С использованием ЕРО и коэффициента (3) разработана методика оптимизации ОЛВ, по критериям минимума вычислительных затрат, минимума числа итераций оценивания при ограничении на вычислительные затраты и обеспечения заданной скорости сходимости оценок параметров [3]. Методика позволяет найти оптимальный ОЛВ для каждой итерации оценивания при заданном распределении вероятностей начального рассогласования оценок параметров. Для априорной оптимизации ОЛВ по различным критериям разработано алгоритмическое и программное обеспечение. Для этого получены выражения для расчета коэффициента улучшения оценки при характерных ЦФ. Проанализировано соответствие теоретических результатов, полученных при априорной оптимизацией ОЛВ и экспериментальным результатам, полученных на различных классах имитированных и реальных изображений.

Оптимальный по заданному критерию ОЛВ, рассчитанный априорно, обеспечивает выполнение этого критерия лишь в среднем. При конкретной реализации изображения оценка ЦФ кроме глобального экстремума может содержать еще и множество ложных локальных экстремумов.

Последние могут быть вызваны, например, коррелированностью отдельных протяженных объектов на изображении и проявляются, если большая часть отсчетов локальной выборки попадает в эти области, т. е. обусловлены ограниченностью ОЛВ. Поэтому увеличение объема или замена локальной выборки Z_t способствует выводу ПГП из локального экстремума. Таким образом, на каждой итерации оценивания целесообразна проверка признаков локальных экстремума ЦФ, а при их наличии — увеличение объема или смена локальной выборки. При этом ОЛВ становится адаптивной величиной.

Найдены признаки локальных экстремумов ЦФ и на их основе синтезированы процедуры оценивания параметров МГДИ, в которых ОЛВ в ходе выполнения процедуры автоматически адаптируется на каждой итерации. При этом очередная итерация проводится при выполнении некоторого условия. Если при минимальном ОЛВ μ_{min} условие не выполняется, то ОЛВ увеличивается (до некоторого предела) до выполнения условия. Это позволяет для сложившейся на данной итерации Z_t минимизировать ее объем и, соответственно, сократить вычислительные затраты. При этом в условии выполнения итерации не используются дополнительные отсчеты изображений.

Отметим, что предлагаемый подход позволяет синтезировать процедуры для совместного решения задач оценивания параметров МГДИ и идентификации с решающим правилом, основанным на значениях ЦФ. В частности, как пример рассмотрена задача адаптации ОЛВ для ПГП идентификации фрагмента на изображении с одновременным определением параметров его местоположения.

Работа выполнена при поддержке РФФИ, проект № 07-01-00138-а.

Литература

[1] *Tashlinskii Alexander.* Computational Expenditure Reduction in Pseudogradient Image Parameter Estimation // Computational Scince — ICCS 2003. — 2003. — Part II (vol. 2658). — Pp. 456— 462.

[2] *Цыпкин Я. З.* Информационная теория идентификации. — Москва: Наука. Физматлит, 1995. — 336 с.

[3] *Minkina G. L., Samojlov M. U., Tashlinskii A. G.* Employment of the Objective Functions in Pseudogradient Estimation of Interframe Geometric Deformations of Image // Patt. Rec. and Im. Anal. — 2005. — No. 1, vol. 15. — Pp. 247–248.

Обработка изображений и потоков видео с целью выделения линейных элементов (Метод Локара)

Михайлов П. И.
cs11mih@mail.ru
Омск, Омский государственный университет им. Ф. М. Достоевского

В работе рассматривается построение алгоритмов обработки изображений и видеопотоков с целью выделения интегрально наиболее близких к прямой элементов изображения (при работе с изображениями) и следа от перемещения объектов (при работе с видеопотоком).

Приложения подобных алгоритмов естественным образом возникают при проверке гипотезы французского криминалиста Локара о том, что углы, образованными линейными элементами подписи, для каждого человека являются психофизиологическими инвариантами. В работе [2] описаны процедуры выделения линейных элементов, углов между ними, проведен статистический анализ результатов применения алгоритма к сериям подписей. Сделан вывод о том, что предположение Локара верно. В данной работе также рассматривается применение разработанных алгоритмов к видеоизображениям с целью обнаружения следа прямолинейного перемещения объектов.

Предположение Локара

Эдмонд Локар, французский криминалист, заметил, анализируя подписи заключенных, что между углами, образованными линейными элементами (интегрально наиболее близких к прямой участков автографа) подписи для каждого человека есть закономерность — углы сохраняются от подписи к подписи. Разработан алгоритм выделения линейных элементов, углов между ними, проведен статистический анализ результатов применения алгоритма к сериям подписей [2]. Алгоритм последовательно проходит все точки изображения. Из каждой точки, принадлежащей автографу, алгоритм пытается «провести» линейный элемент. Алгоритм находит начальную точку, вычисляет угол и продвигается по изображению из найденной точки в направлении угла, пока на пути есть точки, принадлежащие подписи. Затем найденный элемент запоминается и стирается с исходного изображения. Заметим, что если в исходном изображении линейные элементы пересекаются, то в результате работы алгоритма один из этих элементов будет разделен на два или более. Поэтому после того, как все линейные элементы найдены, необходимо произвести «склейку» близких по углу и координатам элементов. Преобразование Радона [1] для выявления прямолинейных участков на данных изображениях не дает удовлетворительного результата, т. к. с помощью этого преобразования сложно различить непрерывный линейный участок и уча-

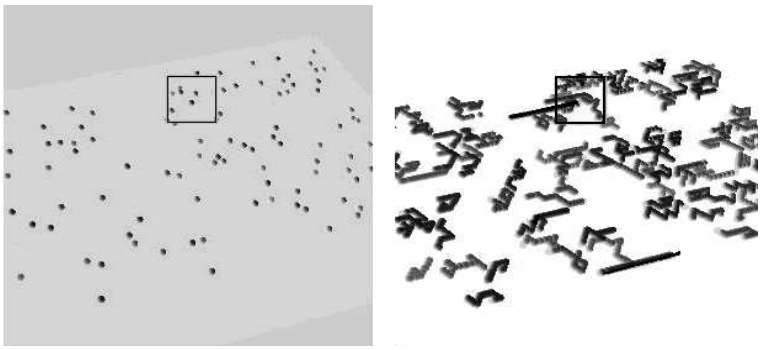


Рис. 1. Результат обработки потока видео на начальном этапе детализации.

сток, состоящий из двух отрезков, лежащих на одной прямой (т. е. прямая содержит разрыв). Для изображения подписи также возможна ситуация, когда угол между прямыми, проходящими через этот отрезок, невелик. Для подписи эти отрезки будут представлять собой различные линейные элементы. С помощью преобразования Радона распознать такую ситуацию достаточно сложно.

Выявление координированного движения

Также рассматривается применение алгоритма Локара и к задаче выявления координированного движения. Разработан программный комплекс для моделирования координированного движения объектов в местах массового скопления людей. Рассматривается открытая местность, по которой перемещаются с различными скоростями и направлениями объекты (люди). Предполагается, что имеется система видеонаблюдения, которая предназначена для обнаружения возможных координированных перемещений и действий злоумышленников. Требуется выделить среди этих объектов (людей) те, которые движутся прямолинейно и/или в определенном направлении. Моделирование проводилось в несколько этапов в зависимости от степени детализации. Сначала в качестве объектов были взяты разноцветные сферы, перемещающиеся по плоскости с различными закономерностями, Рис. 1.

Следующими этапами стали добавление более приближеннного к реальному освещения и материалов объектов, изменение положения камеры. Далее вместо сфер в были введены модели людей, затем была добавлена анимация, Рис. 2.

Рассматривались несколько последних кадров сформированного таким образом видеопотока, особым образом вычислялась их разность, к полученному потоку применялась модификация алгоритма Локара с це-

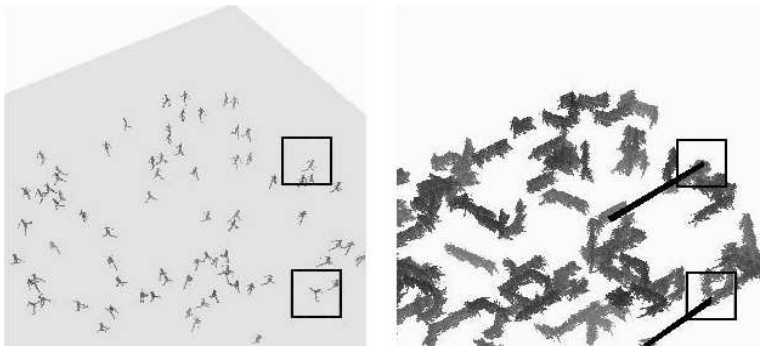


Рис. 2. Результат обработки видео на втором этапе детализации

лью выделения следа перемещения объекта от кадра к кадру и накопления таким образом предполагаемого пути перемещения объекта, Рис. 1, 2. Если путь (или пути) удовлетворяют определенным условиям, например, близки к прямолинейному, или сходятся к определенной точке, то оператору системы выдается сигнал предупреждения. Преобразование Радона применять в данной ситуации тоже затруднительно, поскольку в полученной «разности» кадров требуется выделять отрезки строго фиксированной длины. Для повышения производительности системы при обработки потоков видео использовались возможности современных графических процессоров.

Заключение

Разработанные алгоритмы выделения отрезков, интегрально близких к прямой, оказались применимы в задаче выявления координированного движения в местах массового скопления людей. Следует отметить, что при рассмотрении данной задачи возник ряд специфических проблем, часть из которых удалось решить, а часть требует более детального исследования (синхронизация и обмен данными между CPU и GPU, ограничения, обусловленные графическим конвейером). В дальнейшем предполагается проведение оптимизации алгоритмов и совершенствование их для работы с реальными потоками видео. Рассматривается возможность введения в модель дополнительных «камер» и обрабатываемых потоков видео с целью повышения точности результатов работы алгоритма.

Литература

[1] *Deans, Stanley R.* The Radon Transform and Some of Its Applications. — New York: John Wiley & Sons, 1983. — 314 с.

[2] *Михайлов П. И., Файзуллин Р. Т.* Анализ подписи, основанный на выделении линейных элементов (Метод Локара) // Информационные технологии

моделирования и управления. Международный сборник научных трудов. Выпуск 17, Воронеж: Научная книга, 2004. — С. 145–150.

Построение оценок достоверности результатов распознавания речи с использованием альтернативных моделей

Нгуен Минь Туан
nmtuan@yahoo.com
Москва, ВЦ РАН

Один из основных подходов к выявлению ошибок при распознавании состоит в вычислении для каждого распознанного слова w числовой характеристики $Cm(w)$ — меры достоверности. Величина $Cm(w)$ сравнивается с некоторым порогом τ_w. Если значение больше, то слово считается правильно распознанным. В противном случае соответствующая часть сигнала считается шумом или незнакомым словом. Существует много различных подходов к определению и оценке меры достоверности [1, 2]. В данной статье предлагается метод её вычисления, основанный на оценке достоверности составляющих слова акустических векторов.

Оценка достоверности распознавания слова

Пусть имеется последовательность акустических векторов $Y = (y_1, \ldots, y_N)$ сегмента сигнала, распознанного декодером как слово w. Тогда на выходе из декодера каждый акустический вектор y последовательности Y соответствует некоторому состоянию q марковской модели слова w. Определим оценку достоверности распознавания акустического вектора y как:

$$P_q(y) = \frac{P(y|\Theta_q)}{P(y|\Theta_q) + P(y|\bar{\Theta}_q)},$$

где $P(y|\Theta_q)$ — вероятность принадлежности акустического вектора y классу акустических векторов состояния q; $P(y|\bar{\Theta}_q)$ — вероятность того, что y не принадлежит классу векторов состояния q.

Считаем, что вероятность $P(y|\Theta)$ имеет вид

$$P(y|\Theta) = \left(\sum_{i=1}^{M} c_i N(m_i, v_i, y)\right)^{\alpha},$$

где M — количество смесей; c_i — вес i-ой смеси, $\sum_{k=1}^{M} c_k = 1$, $c_k > 0$; $N(m_i, v_i, y)$ — нормальное распределение с вектором средних m_i и диагональной матрицей ковариации v_i; α — числовой коэффициент, $0 < \alpha \leqslant 1$.

Для любого распознанного слова w с соответствующими ему последовательностями акустических векторов $Y = (y_1, \ldots, y_N)$ и состояний

скрытых марковских моделей $Q = (q_1, \ldots, q_N)$ определим оценку достоверности корректного распознавания как

$$Cm(w) = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log P_{q_i}(y_i)\right). \quad (1)$$

Чтобы пользоваться формулой (1) для вычисления оценки достоверности слов, требуется определить значения параметров моделей Θ_q и $\bar{\Theta}_q$ состояний q скрытых марковских моделей звуков. Эмпирический метод оценки этих параметров описывается в следующем разделе.

Оценка параметров моделей Θ_q и $\bar{\Theta}_q$

Предлагаемый метод оценки параметров моделей Θ_q и $\bar{\Theta}_q$ состояния q состоит в нахождении параметров этих моделей таким образом, чтобы увеличить значение $P_q(y)$ для всех «корректных» акустических векторов и уменьшить значение $P_q(y)$ в противном случае на выбранном наборе акустических векторов. Акустический вектор y будем считать «корректным», если последовательность акустических векторов $Y = (y_1, \ldots, y_N)$, содержащая y, корректно распознана декодером.

Введем функцию стоимости акустического вектора y:

$$F(y, \Theta_q, \bar{\Theta}_q) = \begin{cases} 1 - P_q(y), & y \text{ — корректный;} \\ P_q(y), & y \text{ — некорректный.} \end{cases}$$

Тогда задача оценки параметров моделей Θ_q и $\bar{\Theta}_q$ сводится к решению задачи оптимизации:

$$\arg\min_{\Theta_q, \bar{\Theta}_q} \sum_{y} F(y, \Theta_q, \bar{\Theta}_q). \quad (2)$$

Задачу (2) предлагается решать методом градиентного спуска.

Численные эксперименты

Численные эксперименты выполнялись на речевом корпусе FaVoR [3].

На Рис. 1 показана параметрическая кривая ошибок I и II рода, построенная для слова «два». Для всей тестовой выборки (при выборе оптимального значения порога для каждого слова из словаря системы) удалось корректно определить 92.8% неправильно распознанных слов. При этом ошибка II рода (отказ от правильно распознанного слова) составила 4.7%.

Заключение

В докладе представлен метод выявления ошибок результатов распознавания речи, который основан на оценке достоверности для наблюдаемых акустических параметров. Его применение на представительном

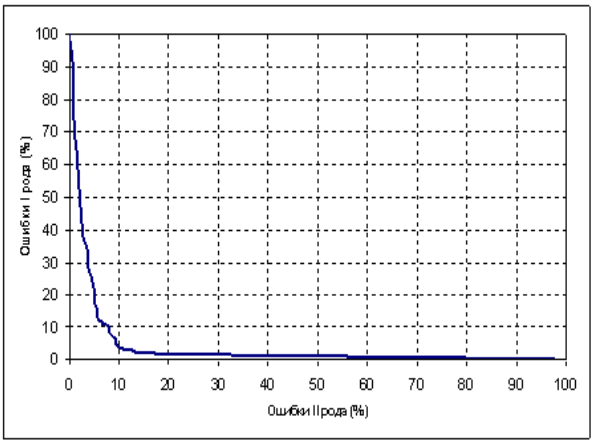


Рис. 1. Параметрическая кривая ошибок I и II рода.

корпусе речевых данных, в котором имеется большое количество различных акустических шумов, как речевой, так и неречевой природы, показало значительное сокращение количества ошибок II рода при умеренном количестве ошибок I рода.

Литература

[1] *Thomas Kemp, Thomas Schaaf.* Confidence measures for spontaneous speech recognition. // Proc. of International Conference on Acoustic, Speech and Signal Processing, 1997. — Pp. 875–878.

[2] *Manhung Siu, Herbert Gish.* Evaluation of word confidence for speech recognition systems // Computer Speech and Language, 1999. Pp. 299–319.

[3] *Десятчиков А. А и др.* Комплекс алгоритмов для устойчивого распознавания человека. — Известия РАН. Теория и системы управления. — 2006.

Расстояния и другие меры близости на множестве черно-белых цифровых изображений

Парфенов П. Г., Каплий И. А., Куликов О. С.
parfenov@uniyar.ac.ru
Ярославль, ЯрГУ

Строится серия функций на множестве пар черно-белых цифровых изображений, являющихся либо расстоянием, либо аналогом расстояния. Предложенные функции позволяют, с одной стороны, решать задачи раз-

личения изображений, с другой стороны, дают некую меру сходства. Рассмотрены приложения такого подхода к текстурам.

Под изображением A будем понимать матрицу $A = (a_{ij})$ размера $m \times n$ с элементами, принимающими значения либо 0, либо 1. Рассматривая матрицу $A = (a_{ij})$ как условный экран и следуя подходу работы [5], введем возможными традиционными способами экранные расстояния между парами пикселов:

$$r_1(a_{ij}, a_{kl}) = \sqrt{(k-i)^2 + (l-j)^2};$$
$$r_2(a_{ij}, a_{kl}) = |k-i| + |l-j|;$$
$$r_3(a_{ij}, a_{kl}) = \max\bigl\{|k-i|, |l-j|\bigr\}.$$

Классическим расстоянием между компактными подмножествами метрического пространства является расстояние Хаусдорфа [1, 2, 5]. Используя соответствующие экранные расстояния $r_p(a_{ij}, a_{kl})$, $p = 1, 2, 3$, стандартным образом определим аналоги расстояния Хаусдорфа на условном экране между изображениями A и B, для $p = 1, 2, 3$:

$$h_p(A, B) = \max\bigl\{\max\{r_p(a_{ij}, B) \colon a_{ij} = 1\}, \max\{r_p(A, b_{kl}) \colon b_{kl} = 1\}\bigr\}.$$

Расстояния h_p, $p = 1, 2, 3$, не характеризуют различие или сходство форм изображений A и B, так как зависят от их взаимного расположения на условном экране. Обозначим через T некоторый класс преобразований изображений, сохраняющих их форму и, естественно, не выводящих изображения за пределы условного экрана. Как представляется, величины $H_p(A, B) = \min\{h_p(\tau(A), B) \colon \tau \in T\}$, характеризуют различия или близость форм изображений A и B. В работах [3, 4] было введено понятие характеристического набора коэффициентов черно-белого цифрового изображения. Для любого такого изображения A строится шестнадцатимерный набор неотрицательных целочисленных коэффициентов $K(A) = \{k_t(A) : t = 0, 1, 2, \ldots, 15\}$, где $k_t(A)$ указывает число фрагментов, размера 2×2, изображения A. Фрагментом, размера 2×2, изображения A назовем любую подматрицу матрицы A, состоящую из элементов $a_{ij}, a_{i+1j}, a_{ij+1}, a_{i+1j+1}$, которые могут принимать значения 0 или 1. Очевидно, что всего существует 16 различных фрагментов. Естественным образом можно определить расстояния на множестве таких наборов для изображений A и B:

$$\rho_1(A, B) = \sqrt{\textstyle\sum_{t=0}^{15}(k_t(A) - k_t(B))^2};$$
$$\rho_2(A, B) = \textstyle\sum_{t=0}^{15} |k_t(A) - k_t(B)|;$$
$$\rho_3(A, B) = \max\bigl\{|k_t(A) - k_t(B)| : t = 0, 1, 2, \ldots, 15\bigr\}.$$

Введенные таким образом функции ρ_p в значительной степени и характеризует различия изображений, но в строгом смысле не является расстоянием между изображениями, так как существуют примеры двух различных изображений A и B, для которых характеристические наборы $K(A)$ и $K(B)$ совпадают.

Но если меру близости ρ_p легко вычислить для любой пары изображений, то вычисление расстояний на основе расстояния Хаусдорфа представляет трудности для текстур, так как множество преобразований T может оказаться пустым в силу того, что любой сдвиг текстуры будет выводить текстуру за пределы условного экрана. Поэтому предлагается сравнивать текстуры, разбивая их на некие блоки W_q; в данном случае текстуры разбиваются на четверти, $q = 1, 2, 3, 4$.

Рассмотрим теперь в качестве T множество целочисленных сдвигов изображений $\tau = (\tau_1, \tau_2) \in T$, где $\tau_1, \tau_2 \in Z$. Дополнительно введем еще изображения $\tau W_q = \{a_{i+\tau_1 j+\tau_2} = a_{ij}\}$, $q = 1, 2, 3, 4$, и положим вне этого блока значения пикселов равными нулю. Безусловным требованием является то, что изображения τW_q не выводятся за пределы условного экрана. Определим еще изображение B_q^τ, совпадающее с B_q на всех пикселах из блока τW_q, и положим вне этого блока значения пикселов равными нулю. Тогда определим расстояния $h_p(\tau W_q, B_q^\tau)$, $q = 1, 2, 3, 4$, $p = 1, 2, 3$.

Вычислим, соответственно, величины

$$M_p^q(A, B) = \min\left\{h_p(\tau W_q, B_q^\tau) \colon \tau \in T\right\}.$$

Определим теперь для текстур A и B расстояния:

$$R_p(A, B) = \max\left\{M_p^q(A, B) \colon q = 1, 2, 3, 4\right\}.$$

Ниже приведены некоторые характерные результаты численных экспериментов для изображений, приведенных на Рис. 1 и Рис. 2, представленных на экране 50×50 пикселов.

Рисунок 1, изображения букв (1,2):
$\rho_1(A, B) = 239.4$ $\rho_2(A, B) = 432$ $\rho_3(A, B) = 183$
$H_1(A, B) = 22.2$ $H_2(A, B) = 31$ $H_3(A, B) = 24$
Рисунок 1, изображения букв (1,3):
$\rho_1(A, A') = 80.2$ $\rho_2(A, A') = 138$ $\rho_3(A, A') = 67$
$H_1(A, A') = 16.4$ $H_2(A, A') = 22$ $H_3(A, A') = 24$
Рисунок 2, текстуры (1,3):
$\rho_1(1, 3) = 80.2$ $\rho_2(1, 3) = 2060$ $\rho_3(1, 3) = 1030$
$R_1(1, 3) = 2.8$ $R_2(1, 3) = 3$ $R_3(1, 3) = 2$
Рисунок 2, текстуры (1,2):

Рис. 1. Изображения букв

Рис. 2. Текстуры

$\rho_1(1,2) = 80.2 \quad \rho_2(1,2) = 2418 \quad \rho_3(1,2) = 1093$
$R_1(1,2) = 7.8 \quad R_2(1,2) = 9 \quad R_3(1,2) = 6$

Полученные результаты, как представляется, показывают хорошее согласование введённых в настоящей работе мер близости ρ_p, $p = 1, 2, 3$, и расстояний, построенных на основе классической метрики Хаусдорфа. Трудоемкость вычисления этих мер близости существенно меньше трудоемкости вычисления для метрик классического типа.

Литература

[1] *Куратовский К.* Топология. — кн. 1, 2 — Москва: Мир, 1966, 1982.

[2] *Келли, Дж.* Общая топология. — Москва: Наука, 1968. — 178 с.

[3] *Кроновер Р. М.* Фракталы и хаос в динамических системах. — Москва: Техносфера, 2006. — 69 с.

[4] *Парфенов П. Г.* О некоторых свойствах характеристического набора коэффициентов черно-белого цифрового изображения // Моделирование и анализ информационных систем. — 2005. — Т. 12, № 1. — С. 52–54.

[5] *Парфенов П. Г.,Назарычев С. Л.* Об одном подходе к различению элементов из больших совокупностей традиционных систем символов // Моделирование и анализ информационных систем. — 2005. — Т. 13, № 1. — С. 46–48.

Об одном алгоритме распознавания числовых матриц

Пролубников А. В., Дудин Д. Л.
prolubnikov@univer.omsk.su
Омск, Омский государственный университет

В докладе предлагается алгоритм распознавания объектов, представляемых числовыми матрицами. В качестве примера рассматриваются растровые изображения, задаваемые матрицами, значения элементов которых соответствуют пикселям изображения. Предполагается, что некоторое изображение из числа эталонных изображений подверглось зашумлению. При этом известно, в каком интервале могло происходить изменение каждого элемента матрицы. Необходимо определить, какое эталонное изображение соответствует зашумленному изображению.

Существует множество подходов к решению данной задачи. Ниже предлагается подход, основанный на оценивании расстояний решений систем линейных уравнений с матрицами, соответствующими зашумленному изображению, от объединенных множеств решений систем линейных уравнений с интервальными матрицами, соответствующими эталонным изображениям.

Формальная постановка задачи следующая. Пусть имеется L квадратных $n \times n$-матриц A_i с элементами a_{ij}. В ходе зашумления одной из матриц — матрицы A_{i_0} — получена некоторая матрица C. Известно, что значение элемента матрицы могло быть изменено в пределах интервала $[a_{ij} - \Delta, a_{ij} + \Delta]$, $\Delta > 0$. Необходимо определить i_0.

Без ограничения общности можно считать матрицы C, A_i квадратными. В противном случае, если имеется L $m \times n$-матриц таких, что $m < n$, то к каждой матрице добавляются $n - m$ нулевых строк.

Пусть $A_i(\delta)$ — интервальная $n \times n$-матрица, с элементами, равными интервалам $[a_{ij} - \delta, a_{ij} + \delta]$, т. е. множество числовых матриц, чьи элементы принадлежат этим интервалам. Тогда $C \in A_{i_0}(\delta)$ для некоторого δ. Для системы уравнений вида $A(\delta)x = b$, где $A(\delta)$ — интервальная матрица, b — некоторый вектор из $\mathbb{R}^n$, объединенное множество решений — это множество $\mathrm{Sol}(A(\delta), b) = \left\{x \in \mathbb{R}^n \;\middle|\; \exists A \in A(\delta) \colon Ax = b\right\}$. Рассматриваемые далее интервальные системы уравнений — системы уравнений с неинтервальной правой частью.

Изменяя δ ($0 < \delta \leqslant \Delta$), будем оценивать расстояние от x_0 до множеств $\mathrm{Sol}(A_i(\delta), b)$, где x_0 — решение системы линейных алгебраических уравнений $Cx = b$. Если для некоторого δ выполняется $x_0 \in \mathrm{Sol}(A_j(\delta), b)$, то матрица C является зашумленной матрицей A_j. Поскольку возможна ситуация, когда имеется несколько значений j таких, что $C \in A_j(\delta)$, и в этой ситуации распознавание таким способом невозможно, в ходе итераций алгоритма будем оценивать не вхождение, а близость значения x_0 к множествам $\mathrm{Sol}(A_i(\delta), b)$ при изменении значения δ.

В общем случае задача нахождения объединенного множества решений интервальной системы линейных уравнений обладает высокой вычислительной сложностью [1]. В предлагаемом алгоритме, благодаря неинтервальности правых частей рассматриваемых интервальных систем уравнений, находится несколько точек из этих множеств (представителей множеств), с помощью которых оценивается близость x_0 к множествам $\mathrm{Sol}(A_i(\delta), b)$. При нахождении представителей множеств производится решение систем линейных алгебраических уравнений с матрицами, модифицированными до матриц со строгим диагональным преобладанием, что обеспечивает геометрическую сходимость приближенных итерационных методов нахождения решения систем линейных алгебра-

Алгоритм 1. Принципиальная схема.

Вход: матрицы $A_i,\ i = 1, \ldots, L,\ C$;
Выход: i_0 — номер матрицы;
1: инициализация: $s_i^j := 0$ для всех $i = 1, \ldots, L,\ j = 1, \ldots, m$;
2: найти x_0 — решение СЛАУ $Cx = b$;
3: **для** $k = 1, \ldots, m$
4: сгенерировать набор матриц $\{\widetilde{A}_i^j(\delta_k)\}_{j=1}^p$ для оценивания расстояния до множества $\mathrm{Sol}(A_i(\delta_k), b)$;
5: решить СЛАУ $\widetilde{A}_i^j(\delta_k) = b,\ i = 1, \ldots, L,\ j = 1, \ldots, p$, $\{x_{ij}^k\}$ — полученные решения;
6: получить представителей $\{\widetilde{x}_{ij}^k\}$;
7: вычислить $\{\rho_i^k\}$: $\rho_i^k := \min\limits_{1 \leqslant j \leqslant p} \{\rho(\widetilde{x}_{ij}^k, x_0)\}$;
8: **для** $k = 1, \ldots, m$
9: **если** q таково, что $\rho_q^k = \min\limits_{1 \leqslant i \leqslant L} \{\rho_i^k\}$ **то**
10: $s_q^k := s_q^k + 1$;
11: найти $i_0,\ k_0$ такие, что $s_{i_0}^{k_0} = \max\limits_{1 \leqslant i \leqslant L,\ 1 \leqslant k \leqslant m} \{s_i^k\}$;

ических уравнений к точному решению. Модифицирование матрицы A с элементами $a_{ij},\ i, j = 1, \ldots, n$ производится следующей заменой их диагональных элементов:

$$a_{ii} := a_{ii} + \sum_{i \neq j}^{n} a_{ij} + 1.$$

В приведенной схеме алгоритма m — число итераций алгоритма, p — число решений из $\mathrm{Sol}(A_i(\delta_k), b)$, рассматриваемого на k-й итерации алгоритма. Значение δ_k определяется на каждой итерации как $\delta_k = (\Delta/m) \times k$. $\rho(x, y)$ — расстояние между векторами x и y.

Генерирование набора $\{\widetilde{A}_i^j(\delta_k)\}_{j=1}^p \in A_i(\delta_k)$ может быть произведено случайным генерированием матриц с элементами в заданных $A_i(\delta_k)$ интервалах. Далее, поскольку в случае неинтервальной правой части объединенные множества решений выпуклы, соответствующие решения $\{x_{ij}^k\}$ могут быть использованы для получения представителей $\{\widetilde{x}_{ij}^k\}$, позволяющих более точно оценить расстояние до объединенных множеств решений.

Предложенный подход показал свою эффективность в ходе проведенного эксперимента с матрицами размеров до 500×500 и значениями элементов от 0 до 255. Изменение элемента матрицы (зашумление) в ходе эксперимента представляло собой равномерно распре-

деленную случайную величину, принимающую значения в диапазоне от 0 до Δ ($100 \leqslant \Delta \leqslant 255$). Количество изменяемых элементов составляло до 80%. Позиции изменяемых элементов выбирались как в соответствии с равномерным распределением по всем позициям элементов матрицы, так и в соответствии с выбором отдельных групп элементов матрицы, подвергаемых зашумлению. Так, при подаче на вход алгоритма изображений букв латинского алфавита и цифр в градациях серого цвета (размер изображений составлял 200×200 пикселей), алгоритм устойчиво распознавал изображения при величине равномерного шума до 60% и одновременном зашумлении до 10 круговых областей, случайно выбираемых в изображении, радиусом до 30 пикселей.

Эффективность алгоритма становится ниже при работе с мнонохромными изображениями. В этом случае при зашумлении происходит инвертирование значений элементов матрицы, тогда как представленный алгоритм использует непрерывность изменения границ множества $\mathrm{Sol}(A(\delta), b)$ при непрерывном изменении элементов A. При проведении вычислительного эксперимента с монохромными изображениями были получены следующие результаты. При наличии небольшого числа случаев, принципиально тяжелых для распознавания алгоритмом (3% от общего числа испытаний), распознавание производилось правильно при уровне равномерного шума до $35,4\%$, тогда как, например, алгоритм типа «Кора» и морфологический метод позволяют устойчиво распознавать изображения вплоть до шума в 42% и 45% соответственно [2].

Литература

[1] *Kreinovich V., Lakeyev A. V., Rohn J., Kahl P.* Computational Complexity and Feasibility of Data Processing and Interval Computations.— Dordrecht: Kluwer, 1997.

[2] *Дюкова Е. В., Кирнос Э. А.* Сравнение алгоритма распознавания типа «Кора» и черно-белой морфологии в задаче распознавания черно-белых изображений // Всеросс. конф. ММРО-9, Москва, 1999. — С. 178–179.

Восстановление трехмерных сцен: первичная модель и способы ее последующего уточнения

Свешникова Н. В., Юрин Д. В.

sveshnikova_n@list.ru

Москва, Московский физико-технический институт

Работа посвящена построению системы восстановления трехмерных сцен по последовательности цифровых изображений. Предполагается, что на изображениях уже выделено и прослежено небольшое количество характеристических отметок. Первичное восстановление осуществ-

ляется итерационным алгоритмом факторизации в перспективной проекции (ИПП), далее результат уточняется двумя различными способами. Первый основан на поиске новых характеристических точек с помощью детектора Харриса (ДХ). Положение соответствующих точек на втором кадре оценивается на основе уже известных соответствий и уточняется с помощью трекера Канаде-Лукаса (КЛ). Второй способ использует стерео подход.

Первичное восстановление

Пусть имеется ($F > 5$) цифровых изображений неподвижной сцены, полученных с обычного фотоаппарата. Пусть также на всей последовательности найдено и прослежено небольшое количество ($P = 10 \div 20$) характеристических точек. Тогда восстановим трехмерные координаты $\{\mathbf{s}_p\}$ этих точек сцены алгоритмом ИПП [1], который также вычисляет все положения $\{\mathbf{t}_f\}$ и ориентации камеры $\{\mathbf{i}_f,\ \mathbf{j}_f,\ \mathbf{k}_f\}$, ее фокусное расстояние g, и предоставляет оценку точности результата, полученную в [2].

Выполнив триангуляцию Делоне (ТД), моделируем поверхность сцены между известными точками плоскостями в соответствии с разбиением на треугольники.

Поиск новых соответствий

Выберем из исходной последовательности изображений кадр L и найдем на нем характеристические точки с помощью ДХ [3].

Для этих точек построим гипотезы их положения на втором кадре R. Пусть найденный на кадре L уголок $\mathbf{x}_L$ находится внутри треугольника τ_L. Предположим, что сцена представляет собой кусочно-гладкую поверхность, это позволяет рассматривать результат ТД как аппроксимацию поверхности сцены. Тогда пара соответствующих треугольников τ_L и τ_R на кадрах L и R определяет аффинное преобразование $\mathbf{A}$, и $\mathbf{A}\mathbf{x}_L$ есть гипотетическое положение выбранного уголка на кадре R. На основе гипотезы ищется соответствие трекером КЛ [4].

Результаты работы трекера должны удовлетворять ограничениям эпиполярной геометрии, которая вычисляется по известным из результатов ИПП параметрам съемки. Далее можно продолжить уточнение или, используя положения и ориентации камер, полученные из алгоритма ИПП, вычислить трехмерные координаты новых точек и дополнить ими модель сцены. Схема алгоритма изображена на Рис. 1.

Уточнение модели с использованием стерео

Как альтернативный способ уточнения сеточной модели предлагается стерео подход [5]. Его важное достоинство в том, что каждое соответствие ищется согласованно с соседями. Однако необходимы малое смещение

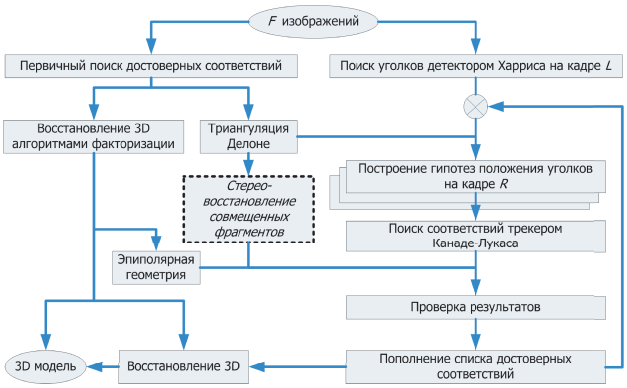


Рис. 1. Схема системы восстановления трехмерных сцен

кадров относительно друг друга и диапазон поиска. Чтобы ограничить диапазон смещений будем решать стерео задачу для фрагментов, на которые ТД разбивает изображение. Каждый треугольник предварительно грубо совмещается в рамках аффинной модели. Далее соответствия ищутся стерео алгоритмом [5]. На схеме Рис. 1 этому этапу соответствует блок, выделенный пунктирной границей.

Результаты

Предложенные подходы тестировались на реальных данных (сцена «Дом») и показали хорошие результаты. На 12 кадрах вручную выделены 16 точек, которые, как и результат ТД, изображены на Рис. 2, в центре. На Рис. 2 слева и справа приведена пара кадров сцены «Дом», выбранная из последовательности для поиска новых соответствий. На Рис. 3 изображены полученные трехмерные модели сцены «Дом». Текстурированная модель слева есть результат восстановления ИПП. Модель в центре дополнена соответствиями, найденными ДХ и трекером КЛ, а справа — результат применения стерео подхода. Сеточное представление уточненных моделей демонстрирует количество восстановленных трехмерных точек сцены.

Работа выполнена при поддержке РФФИ, грант 06-01-00789-а.

Литература

[1] *Свешникова Н. В., Юрин Д. В.* Априорный и апостериорный расчет погрешностей восстановления трехмерных сцен алгоритмами факторизации // Программирование — 2004, — Т. 30, № 5, — С. 48–68.

[2] *Sveshnikova N. V., Yurin D. V.* The Factorization Algorithms: Results Reliability and Application for the Epipolar Geometry Recovery // 16-th Int. Conf. on Comp. Graph. and App. GraphiCon'2006, Novosibirsk, 2006.

Рис. 2. *Слева и справа*: изображения сцены «Дом». *В центре*: ТД по точкам, найденным и прослеженным вручную.

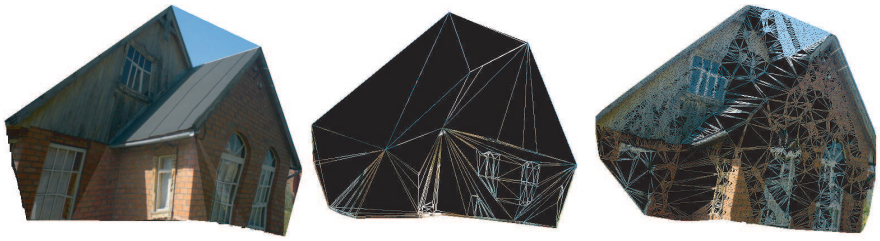


Рис. 3. Восстановленные трехмерные модели. *Слева*: алгоритм ИПП; *в центре*: уточненные ДХ и трекером КЛ; *Справа*: стерео.

[3] *Harris C. G., Stephens M.* A combined corner and edge detector // In Proc. 4th Alvey Vision Conf., Manchester, 1988 — Pp. 147–151.

[4] *Tomasi C., Kanade T.* Shape and Motion from Image Streams: a Factorization Method, Part 3, Detection and Tracking of Point Features // Tech. Rep. CMU-CS-91-132, School of Computer Science, Carnegie Mellon Univ., 1991.

[5] *Kolmogorov V., Zabih R.* Computing visual correspondence with occlusions using graph cuts // Int. Conf. on Computer Vision, Vancouver, 2001.

Об использовании параллельных вычислений в задачах машинного зрения

Семенов А. Б.
semenov@tversu.ru
Тверь, Тверской государственный университет

В настоящей статье исследуется возможность применения параллельных вычислений в задачах машинного зрения с использованием многоядерных процессоров.

Введение

Проблема большого количества вычислений возникает в ряде задач многих направлений науки и техники. Задачи машинного зрения и обработки изображений, решаемые в режиме реального времени, использующие сложный математический аппарат и большой объем обрабатываемых данных, выставляют высокие требования по скорости счета. Поэтому разработка эффективных методов и алгоритмов для решения подобных задач является весьма актуальной проблемой.

Использование программируемых графических процессоров с их параллельной архитектурой позволит повысить производительность довольно ограниченного круга задач, таких как попиксельная обработка изображений. Но для задач, связанных с векторизацией изображений, они не вполне пригодны в силу своей специфичности. Применение же многопроцессорных систем для решения подобного спектра задач является заведомо дорогим способом. В настоящее время начинают очень широко распространяться процессоры с несколькими независимыми вычислительными ядрами. Уже сейчас в настольных компьютерах можно встретить процессоры с 2-мя и 4-мя ядрами. Поэтому интересно оценить возможность использования многоядерных процессоров для ускорения решения задач машинного зрения. Для этого был проведен вычислительный эксперимент на модельной задаче: распознавание (выделение) окружности на растровом изображении с использованием параллельных вычислений. В качестве примера был рассмотрен метод определения достоинства монеты по величичине радиуса найденной окружности.

Привлечение параллельно-вычислительных алгоритмов и методов и использование алгоритмической парадигмы «разделяй и властвуй» позволит рассчитывать на повышение скорости вычислений, пропорциональное числу используемых параллельных процессоров.

Задачи машинного зрения и распознавания образов используют математические методы и алгоритмы, позволяющие получить некую описательную (смысловую) информацию о заданном изображении. Процедура распознавания применяется к изображению и преобразует его в некото-

Алгоритм 1. Классическое преобразование Хафа для окружностей.

1: **для всех** черных точек изображения (x_i, y_i), $i = 1, \ldots, N$
2: **для** X от $X_{\min}$ до $X_{\max}$ с шагом dX
3: **для** Y от $Y_{\min}$ до $Y_{\max}$ с шагом dY
4: $R := \sqrt{(X - x_i)^2 + (Y - y_i)^2}$;
5: $H[X, Y, R] := H[X, Y, R] + 1$;

рое абстрактное описание: набор чисел, цепочку символов, и т. д. Далее изображение относят к одному из классов распознавания.

Описание метода

Исходными данными для исследуемой модельной задачи является цифровое изображение монеты, полученное путем сканирования или фотографирования ее на некотором фоне, Рис. 1. Метод определения достоинства монеты базируется на анализе величины радиуса найденной окружности. Основная идея подхода включает в себя следующие шаги:

— приведение (преобразование) изображения к оттенкам серого цвета;
— выделение контуров (линий с резким перепадом интенсивности) на изображении с помощью цифровых фильтров выделения границы;
— операция бинаризации (пороговое отсечение);
— поиск точек, лежащих на окружности с помощью классического метода Хафа (Hough), см. Алгоритм 1;
— определение радиуса найденной окружности и ее классификация.

Элемент массива $H[X, Y, R]$ соответствует количеству черных точек изображения, лежащих на окружности с центром в точке (X, Y) и радиусом R. Вычислительная сложность классического алгоритма Хафа для окружностей есть $O\big(N(X_{\max} - X_{\min})(Y_{\max} - Y_{\min})\big)$. Если же в задаче априори известен диапазон изменения величины радиуса искомой окружности, то алгоритм Хафа может быть модифицирован так, что это позволит сократить количество арифметических операций, используемых в алгоритме. Кроме того, методы и алгоритмы, основанные на циклах (за исключением итерационных), а также алгоритмической парадигме «разделяй и властвуй», довольно хорошо поддаются распараллеливанию. Таким образом, весь объем вычислений мы можем равномерно распределить между имеющимися в нашем распоряжении параллельными процессорами, так чтобы они одновременно выполняли арифметические операции. Современные операционные системы предоставляют возможность организовать выполнение программных инструкций в параллельном режиме через механизм нитей (threads). Именно благодаря «многонитевому» подходу на компьютерах с многоядерной и многопро-

Рис. 1. Отсканированное изображение монеты

цессорной архитектурой возможно существенное повышение производительности параллельных алгоритмов.

Вычислительные эксперименты

Для проведения вычислительных экспериментов был разработан и реализован параллельный алгоритм поиска центра и величины радиуса окружности на растровом изображении. Входными данными для алгоритма являлись отсканированное изображение монеты, и количество нитей, среди которых и проводилось распределение вычислительной нагрузки. Вычислительный эксперимент проводился на персональном компьютере с двуядерным процессором. Результаты сведены в следующей таблице:

	1 нить	2 нити	4 нити
время обработки (мс)	5600	3320	3390

Таким образом, использование многоядерных и многопроцессорных технологий совместно с параллельно-вычислительным подходом позволит рассчитывать на повышение производительности задач машинного зрения, допускающих распараллеливание. В данной работе удалось достичь порядка 70% увеличения производительности.

Работа выполнена при поддержке РФФИ, проект №05-01-00542, и корпорации INTEL.

Литература

[1] *Воеводин В. В., Воеводин Вл. В.* Параллельные вычисления. — СПб, 2002.

[2] *Прэтт У.* Цифровая обработка изображений. Кн. 1, 2. М.: Мир, 1982.

[3] *Короткий С.* Введение в распознавание образов. Часть 1. Преобразование Хафа. // Монитор. — 1994. — №8.— С. 22–25.

Регуляризация в распознавании изображений: принципы гладкости решающего правила и выбора информативной подобласти

Середин О. С.
oseredin@yandex.ru
Тула, Тульский государственный университет

В докладе приводится отчет о последних совместных работах Лаборатории анализа данных Тульского государственного университета и Вычислительного центра РАН, направленных на улучшение экстраполирующих свойств решающего правила распознавания при обучении на малых выборках в задачах анализа изображений. Рассматриваются две идеи повышения прогнозирующих свойств решающих правил. Первая заключается в наложении специальных априорных ограничений на вариабельность решающих правил в исходном признаковом пространстве; вторая идея заключается в управляемом отборе информативных признаков. Оба идейных подхода являются стандартными [1, 2], мы же предлагаем конкретные операциональные и, по нашему мнению, эффективные решения. Простейшая математическая модель изображения — множество действительных чисел, соответствующих элементам дискретной решетки (обычно пиксельной) и представляющих уровень яркости конкретного пикселя. Естественно рассматривать изображение как вектор в соответствующем евклидовом пространстве. Как правило, размерность такого пространства огромна, поскольку равна числу пикселей изображения. На практике редко удается набрать обучающую выборку, хотя бы сопоставимую по числу объектов с таким размером.

Однако такое евклидово пространство обладает отличительной спецификой, а именно, наличием двух координатных осей, вдоль которых естественным образом упорядочены пиксели. Таким образом, в отличие от обычного евклидова пространства, на элементах вектора признаков установлена дополнительная мера их близости. Две координаты близки друг к другу, если близки на плоскости изображения соответствующие им пиксели. Так можно образно говорить о «гладких» и «негладких» векторах в линейном пространстве. Подобная регуляризация рассматривалась в работе [3, 4], посвященной анализу одномерных сигналов.

В данной работе предлагается способ регуляризации стандартного критерия обучения (метод опорных векторов, SVM [5]) за счет учета расстояния между элементами пиксельной решетки (Рис. 1)

$$d_{ts,t's'} = \sqrt{(t-t')^2 + (s-s')^2} \geqslant 0.$$

Пусть дана обучающая выборка изображений двух классов $(\mathbf{x}_j, g_j)$, $j = 1, \ldots, N$, $\mathbf{x}_j = (x_{ts,j})$, $t = 1, \ldots, T$, $s = 1, \ldots, S$, $g_j \in \{1, -1\}$.

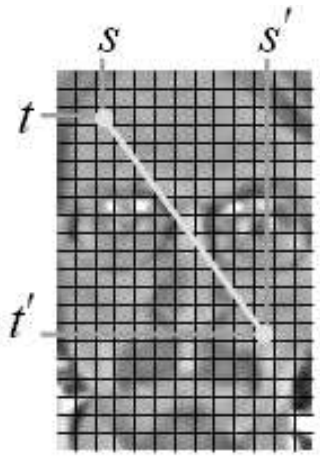


Рис. 1. Расстояние между элементами растровой решетки.

Необходимо найти дискриминантную функцию $\mathbf{a} = (a_{ts}),\ t = 1, \ldots, T,$ $s = 1, \ldots, S \in R^n$.

Регуляризованный критерий обучения:

$$\begin{cases} \mathbf{a}^{\mathrm{T}}(\mathbf{I} + \gamma\mathbf{B})\mathbf{a} + C\sum\limits_{j=1}^{N}\delta_j \to \min\limits_{\mathbf{a},b}; \\ g_j(\mathbf{a}^{\mathrm{T}}\mathbf{x}_j + b) \geqslant 1 - \delta_j, \quad \delta_j \geqslant 0, \quad j = 1, \ldots, N; \end{cases}$$

$$\mathbf{B} = 2\begin{pmatrix} -p_{1,1} + \sum\limits_{j=1}^{TS} p_{1,j} & \cdots & -p_{1,TS} \\ \cdots & \cdots & \cdots \\ -p_{TS,1} & \cdots & -p_{TS,TS} + \sum\limits_{j=1}^{TS} p_{TS,j} \end{pmatrix}.$$

Здесь $p_{ts,t's'}$ — неотрицательная функция похожести, например, вида

$$p_{ts,t's'} = \begin{cases} 1, & d_{ts,t's'} \leqslant \sqrt{2}, \\ 0, & d_{ts,t's'} > \sqrt{2}, \end{cases} \quad \text{или} \quad p_{ts,t's'} = \begin{cases} 1, & d_{ts,t's'} \leqslant 1, \\ 0, & d_{ts,t's'} > 1, \end{cases}$$

где $\gamma \geqslant 0$ — параметр регуляризации. В докладе рассмотрен вопрос подбора параметра регуляризации.

Вторая идея связана с автоматическим отбором информативных подобластей на плоскости изображения. В работе [6] был описан безитерационный чрезвычайно эффективный метод отбора информативных признаков, получивший название Selective Kernel Fusion. Однако, процеду-

ра, будучи применена к задаче анализа изображений, отыскивала слишком локальные особенности конкретной обучающей выборки и отбирала на всей плоскости изображения единичные «выколотые» пиксели-признаки. Внесение в эту процедуру регуляризующей добавки позволило выделять на изображении информативные подобласти:

$$\begin{cases} \mathbf{a}^{\mathrm{T}} \left[\operatorname{diag}\left(1/\sqrt{r_{ts}}\right)\right]^{\mathrm{T}} (\mathbf{I} + \gamma\mathbf{B}) \left[\operatorname{diag}\left(1/\sqrt{r_{ts}}\right)\right] \mathbf{a} + \\ \qquad + \sum\limits_{t=1}^{T} \sum\limits_{s=1}^{S} \log r_{ts} + C \sum\limits_{j=1}^{N} \delta_j \to \min\limits_{\mathbf{a},b,\mathbf{r}}; \\ g_j(\mathbf{a}^{\mathrm{T}}\mathbf{x}_j + b) \geqslant 1 - \delta_j, \quad \delta_j \geqslant 0, \quad j = 1, \ldots, N. \end{cases}$$

Обучение представляет собой процедуру Гаусса-Зайделя по искомым параметрам $\mathbf{a}$ и $\mathbf{r} = (r_{ts})$, $t = 1 \ldots, T$, $s = 1, \ldots, S$, начиная с некоторого заранее заданного начального значения, например, $r_{ts} = 1$. Суть параметров обучения $\mathbf{r}$ — веса при информативных признаках.

В докладе приводятся результаты экспериментальных исследований предложенных критериев как на реальных изображениях людей, так и на модельных изображениях.

Работа выполнена при поддержке грантов РФФИ №05-01-00679, №06-01-08042, №06-01-00412, №06-07-89249.

Литература

[1] *Guyon I., Elisseeff A.* An Introduction to Variable and Feature Selection // Journal of Machine Learning Research. — 2003 — 3. — P. 1157–1182.

[2] *Juwei Lu, Plataniotis K. N., Venetsanopoulos A. N.* Regularization Studies of Linear Discriminant Analysis in Small Sample Size Scenarios with Application to Face Recognition // Pattern Recognition Letter. — 2005. — Vol. 26, issue 2. — P. 181–191.

[3] *Seredin O. S., Dvoenko S. D., Krasotkina O. V., Mottl V. V.* Machine Learning for Signal Recognition by the Criterion of Decision Rule Smoothness // Pattern Recognition and Image Analysis. — 2001. — Vol. 11, №1. — P. 87–90.

[4] *Seredin O., Mottl V.* Regularization in image recognition: the principle of decision rule smoothing // Pattern Recognition and Information Processing: Proc. of the 9th Int. Conf., Minsk, 2007. — Vol. II. — P. 151–155.

[5] *Vapnik V.* Statistical Learning Theory. — New York: Wiley, 1998.

[6] *Mottl V., Krasotkina O., Seredin O., Muchnik I.* Kernel fusion and feature selection in machine learning // Proc. of the 8th IASTED Int. Conf. Intelligent Systems and Control, Cambridge, USA, 2005. — P. 477–482.

Кластерный алгоритм для текстурных изображений

Сидорова В. С.
svs@ooi.sscc.ru
Новосибирск, ИВМиМГ

Рассматривается приложение гистограммного кластерного алгоритма для автоматической классификации аэрокосмических изображений по текстурным признакам. Оценка качества классификаций векторов, представленных с различной детальностью, позволяет выбрать лучшие распределения. Применение подхода к многоканальным спутниковым данным показало, что лучшие распределения соответствуют информационным классам покрытия Земли [1]. Классификация текстур имеет особенности. Алгоритм применен к аэроснимкам леса.

Алгоритм классификации

В основе классификации — быстрый непараметрический алгоритм разделения векторного пространства по унимодальным кластерам, которые соответствуют локальным максимумам гистограммы [2]. Алгоритм используется многократно для различного числа уровней квантования N векторного пространства. Пусть их начальное число $N_0 = 256$, $N < N_0$. Размер ячейки для произвольного уровня квантования $kf = (N_0 - 1)/(N - 1)$, L — число признаков, $f = (f_1, f_2, \ldots, f_L)$ — вектор признаков, а $g = (g_1, g_2, \ldots, g_L)$ — вектор, в который преобразуется f в результате квантования: $g_k = [f_k/kf]$, $k = 1, \ldots, L$, где $[\cdot]$ — целая часть числа. Получается ряд распределений векторов для различных значений N. По предложенной в [1] мере качества определяются лучшие распределения ряда. Мера качества для отдельного унимодального кластера $M^j(N)$ и мера качества распределения в целом $M(N)$:

$$M^j(N) = \frac{1}{H^j(N)} \frac{1}{B^j(N)} \sum_{i=1}^{B^j(N)} h_i^j(N), \quad M(N) = \frac{1}{K(N)} \sum_{j=1}^{K(N)} M^j(N), \quad (1)$$

где $h_i^j(N)$ — значение гистограммы в i–той точке границы кластера j, $B^j(N)$ — число точек границы кластера, $H^j(N)$ — максимальное значение гистограммы, $K(N)$ — число кластеров. Лучшие классификации соответствуют минимумам $M(N)$.

Особенности классификации текстур

Статистические текстурные признаки вычисляются по окрестности точки изображения. Пусть окрестностью будет квадратное окно одного размера для всех точек изображения, его определим автоматически. Начиная с некоторого размера, будем постепенно его увеличивать.

Для каждого найдем лучшую классификацию и соответствующее число кластеров. Предположим, что по достижении определенного размера окна не только признаки стабилизируются для всех объектов, но и перестанет меняться число кластеров. Признаки внутренних точек объектов на изображении перестанут меняться, признаки граничных точек могут измениться, но мало повлияют на образование кластеров, если учесть, что мы выбираем классификации с хорошо изолированными кластерами. Когда число кластеров станет равным для двух последовательных размеров окна, выберем лучшую классификацию для меньшего из них.

Полученные кластеры не могут быть на изображении тоньше размера этого окна по определению. В плоскости изображения на границах возможно появление ложных кластеров. Их можно присоединить к соседним при построении кластерной карты. Для автоматической индикации тонких кластеров посчитаем отношение числа граничных точек на изображении каждого кластера к его площади и, если отношение больше порога, то кластер ложный. Здесь в качестве порога примем это отношение для размеров окна. Найдем два наиболее представительных соседа P_1 и P_2 ложного кластера. Из них для присоединения выберем тот, который менее изолирован от него в пространстве признаков. Для этого используем $M^j(N)$ из (1). Пусть ложный кластер — j, подсчитаем отдельно вклад в $M^j(N)$ точек границы, соседних с P_1 и с P_2.

Эксперименты

На Рис. 1 аэроснимок кедровников Западной Сибири масштаба 1:50000. Три самых светлых пятна соответствуют водным поверхностям. Смешанный лес данного типа представляет собой березово-кедровое сообщество. Возобновляясь на гарях, он закономерно видоизменяется: на ранних стадиях преобладает береза, затем ее вытесняет кедр. Размер электронной версии изображения 600×400, разрешение 5м/пиксель. При таком разрешении текстуру леса составляют чередующиеся светлые группы берез (осенняя съемка) и темные — кедра. Существует несколько качественно отличных фаз развития этого типа леса. Точность наземной таксации и визуального дешифрирования аэроснимков — в пределах фазы. Визуальные характеристики лесных объектов входят в таксационные таблицы. Основные — тон и текстура. Основываясь на физических свойствах, мы выбрали текстурные признаки: средний тон и средний модуль разности тонов в паре соседних пикселей (статистика Харалика [3]) для предварительно эквализованного изображения. Другие исследования показали пригодность этих признаков для различения фаз [4]. Размер окна для расчета признаков получился 14×14 пикселей. Это соответствует участку 70×70 кв. м на земле (размер элементарного участка таксации 100×100 кв. м). Лучшее распределение в смысле меры $M(N)$ (1)

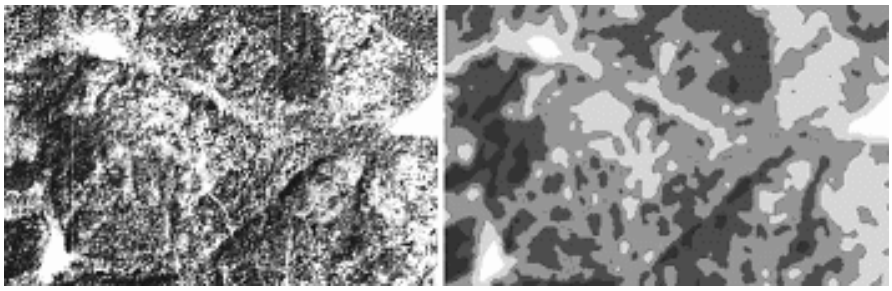


Рис. 1. Изображение лесного ландшафта и кластерная карта.

соответствует двум кластерам, разделяющим область признаков на лес и водную поверхность. Следующее по возрастанию значения минимума меры качества соответствует семи кластерам. Было выявлено два ложных кластера, и они объединены с соседними. На карте (рис. 1 справа) отображено пять полученных кластеров. Самый светлый соответствует воде. Остальные представляют четыре старшие возрастные фазы кедровников (с четвертой по седьмую) в соответствии с классификацией лесоводов.

Выводы

Представленный алгоритм позволил автоматически и быстро получить лучшую в смысле предложенной меры классификацию изображения кедровников на аэроснимке по заданным текстурным признакам. Это распределение оказалось в соответствии с делением кедровников по фазам развития, что соответствует точности наземной таксации.

Работа выполнена при частичной финансовой поддержке РФФИ, проект № 07-07-00085а.

Литература

[1] *Сидорова В. С.* Оценка качества классификации многоспектральных изображений гистограммным методом // Автометрия. — 2007. — № 1 — С. 37–43.

[2] *Narendra P. M., Goldberg M.* A non-parametric clustering scheme for LANDSAT // Pattern Recognition 9. — 1977. — Pp. 207–215.

[3] *Haralick R. M., Shanmugam K., Dinstein I.* Textural Features for Image Classification // IEEE Trans. Syst. Man. Cybern. — 1973. — V. SMS-3 — Pp. 610–621.

[4] *Sidorova. V. S.* Modeling Age Dynamics of the Forest Texture in Aeroimage // Proc. IASTED Int. Conf. ACIT. — Novosibirsk, 2002. — Pp. 441–446.

Эллипсоидальные фильтры для оперативной обработки сигналов в нелинейных стохастических системах

Синицын И. Н., Синицын В. И., Белоусов В. В., Хоанг Тхо Ши

VSinitsin@ipiran.ru

Москва, Институт проблем информатики Российской академии наук, Московский физико-технический институт

Рассматриваются вопросы методического и программного обеспечения для синтеза нелинейных эллипсоидальных фильтров для оперативной обработки информации. Приводятся примеры применения.

Методическое обеспечение

Как известно [1, 2], статистическая информатика обладает обширным арсеналом эффективных статистических методов анализа и оперативной (быстрой) обработки сигналов. В задачах стандартного анализа сигналов в стохастических системах (СтС) обычно ограничиваются спектрально-корреляционными характеристиками, в то время как функционирование СтС в экстремальных условиях требует развития нестандартных методов анализа, основанных на одно- и многомерных распределениях. Для решения задачи анализа распределений в нелинейных СтС применяют следующие три принципиально различных подхода.

Первый подход состоит в использовании прямого численного решения уравнений СтС методом Монте-Карло.

Второй подход состоит в непосредственном составлении и интегрировании эволюционных функциональных уравнений, например, уравнений Фоккера-Планка-Колмогорова, Колмогорова-Феллера и их обобщений, а также уравнений Пугачева для характеристических функций.

Третий подход состоит в применении аналитических методов для приближенного решения уравнений, определяющих параметры одно- и многомерных распределений. К их числу относятся методы нормальной аппроксимации и статистической линеаризации, методы эквивалентной линеаризации, методы моментов, семиинвариантов, квазимоментов и их модификации, методы ортогональных разложений и др.

Радикальным подходом к сокращению числа уравнений для параметров распределения является подход, основанный на параметризации структуры распределения. Так, как обнаружено В. И. Синицыным, радикального сокращения числа уравнений для параметров распределения удается добиться для эллипсоидальной структуры распределения [1, 2, 3].

В докладе приведены новые теоретические результаты в области статистической информатики, среди которых следует выделить следующие:

— получены уравнения методов эллипсоидальной аппроксимации (МЭА) и линеаризации (МЭЛ) в непрерывных (дискретных) негауссовских СтС для анализа сигналов по априорным данным;
— выведены фильтрационные уравнения для эллипсоидальной обработки сигналов в непрерывных (дискретных) гауссовских СтС на основе апостериорных данных.

Практическая ценность работ состоит в том, что они являются основой для создания современных информационных технологий статистического анализа и синтеза сложных информационно-измерительных и информационных систем.

Программное обеспечение

Программное обеспечение (ПО) включает:

— ПО эллипсоидального анализа распределений сигналов по априорным данным (среда «MATLAB — СтС-АНАЛИЗ», шифр «СтС-Э.АНАЛИЗ», 12 модулей);
— ПО эллипсоидального анализа сигналов по апостериорным данным (среда «MATLAB — СтС-ФИЛЬТР», шифр «СтС-Э.ФИЛЬТР», 4 модуля).

Для замыкания уравнений для параметров ЭА разработаны приближенные рекуррентные формулы, связывающие старшие и младшие вероятностные моменты. Для использования МЭЛ созданы таблицы коэффициентов ЭЛ типовых нелинейностей.

Применения

1. Дано решение задачи анализа и фильтрации для нелинейных процессов в интерферометре Фабри-Перо. Построены квазилинейные стохастические модели обработки информации по априорным данным в нелинейном интерферометре Фабри-Перо на основе МСЛ и МЭЛ. Показано, что точность расчётов по МЭЛ, по сравнению с МСЛ, повышается в 1.5–2 раза и составляет 1–2%. Получены аналитические выражения для эффективных собственных частот регулярных колебаний и статистических характеристик флуктуаций. Эти выражения использованы для выбора оптимальных параметров интерферометра. Они позволяют избежать вычисления сложных эллиптических интегралов. Построены гауссовские и эллипсоидальные квазилинейные фильтры для обработки информации по апостериорным данным. Проведена оценка точности эллипсоидальных фильтров с помощью обобщенного фильтра Калмана-Бьюси (ОФКБ) и фильтра второго порядка. Эллипсоидальные фильтры целесообразно использовать только при больших коэффициентах негауссовости (свыше 30%).

2. Разработаны стохастические модели флуктуаций чандлеровских автоколебаний полюса Земли на основе априорных данных для нелинейного обобщенного релеевского механизма диссипации. Изучены основные вопросы статистической динамики автоколебаний полюса Земли. Разработаны квазилинейные стохастические модели флуктуаций полюса Земли на основе апостериорной информации. Оценена точность квазилинейных моделей с помощью моделей ОФКБ и фильтра второго порядка.

Работа выполнена при поддержке РФФИ, проект № 07-07-00031 и Программы ОИТВС РАН «Фундаментальные основы информационных технологий и систем» (проект 1.5).

Литература

[1] *Пугачев В. С., Синицын И. Н.* Теория стохастических систем, — Москва: Логос, 2000. — 1000 с. [пер. на англ. яз. Stochastic Systems. Theory and Applications. World Scientific. Singapore, 2001.], 2004 (2-е изд.).

[2] *Синицын И. Н.* Фильтры Калмана и Пугачева. — Москва: Логос, 2006. — 640 с.

[3] *Синицын И. Н., Синицын В. И.* Эллипсоидальный анализ распределений в стохастических системах и его применение // Наукоемкие технологии, 2006. — № 7–8. — С. 123–130.

К вопросу об инварианте графического изображения

Спиридонов К. Н.
spiridonov@petrsu.ru
Петрозаводск, ПетрГУ

Впервые тема анализа изображений петроглифов была затронута в 1977 г. А. Я. Шером в его статье [3]. В данной статье А. Я. Шер предложил алгоритм классификации петроглифов на основе стилистических инвариантов. На данный момент актуальными направлениями в области численного анализа изображений петроглифов являются:

— разработка методов определения стилистических особенностей петроглифов по технике их выбивания (определение инварианта петроглифа по технике выбивания);
— определение порядка заполнения скалы петроглифами (иногда они наложены друг на друга, и даже опытные специалисты не могут точно установить их границы);
— бинарная сегментация изображений петроглифов.

Для решения этих и других задач необходимо разработать методологию расчета характеристик каменной поверхности на основе фотографических снимков. В качестве метода расчета характеристик поверхности скалы был выбран метод мультифрактальной параметризации струк-

тур [2], первая модификация которого была разработана в 1993 г. в лаборатории прочности металлических материалов ИМЕТ РАН. Одной из основных характеристик в данном методе является спектр фрактальных размерностей Реньи D_q.

Мультифрактал и спектр фрактальных размерностей Реньи

Существуют различные определения мультифракталов, но ни одно из них не описывает все множество мультифрактальных объектов. В данной работе используется определение мультифрактала из [1]. Рассмотрим фрактальный объект, занимающий некую область M размера m в d-мерном евклидовом пространстве. Разобьем всю область M на кубические ячейки со стороной $\varepsilon \ll m$ и объемом ε^d. Пусть на каком-то этапе построения фрактального объекта он представляет собой множество из $N \gg 1$ точек, как-то распределенных в области M. Будем предполагать, что в конце концов $N \to \infty$. Пусть номер занятых ячеек i изменяется в пределах $i = 1, \ldots, N(\varepsilon)$, где $N(\varepsilon)$ — суммарное количество занятых ячеек, зависящее от ε. Пусть $n_i(\varepsilon)$ представляет собой количество точек в ячейке с номером i. Тогда величина $p_i(\varepsilon) = \lim\limits_{N\to\infty} \frac{n_i(\varepsilon)}{N}$ представляет собой оценку вероятности того, что наугад взятая точка из фрактального множества находится в ячейке i. Если значения p_i зависят от i, то фрактальный объект называют *мультифракталом*. Одной из основных характеристик мультифракталов является *спектр фрактальных размерностей Реньи*. Приведем его определение. Введем в рассмотрение обобщенную статистическую сумму $Z(q, \varepsilon)$, характеризуемую показателем степени q, который может принимать любые действительные значения, $Z(q, \varepsilon) = \sum_{i=1}^{N(\varepsilon)} p_i^q(\varepsilon)$. Спектр фрактальных размерностей Реньи D_q, характеризующий распределение точек мультифрактала в занимаемой им области M, определяется с помощью соотношения: $D_q = \frac{\tau(q)}{q-1}$, где функция $\tau(q)$ имеет вид $\tau(q) = \lim\limits_{\varepsilon\to 0} \frac{\ln[Z(q,\varepsilon)]}{\ln[\varepsilon]}$.

Метод расчета спектров D_q для полутоновых изображений

Приведем общее описание метода мультифрактальной параметризации структур (полное описание см. в [2]):

1. Исследуемый объект помещается в евклидово пространство. В частности, так можно рассматривать часть поверхности скалы, какого-либо изделия из металла и т. д.
2. Область, занимаемая объектом, разбивается на ячейки со стороной ε.
3. Каждой i-й ячейке присваивается мера распределения вещества $p_i(\varepsilon)$ в данной ячейке.

4. Используя введенную меру $p_i(\varepsilon)$, рассчитываются различные мультифрактальные характеристики данного объекта, например, спектр фрактальных размерностей Реньи D_q.
5. На основании полученных характеристик делаются те или иные выводы относительно свойств исследуемого объекта.

В нашем случае величина $p_i(\varepsilon)$ (3-й пункт) для i-й ячейки со стороной ε рассчитывается по формуле $p_i(\varepsilon) = C_i/C$, где C — сумма значений цветов пикселей по всему изображению, C_i — сумма значений цветов пикселей по i-й ячейке.

Анализ фотографий петроглифов Карелии с помощью спектров фрактальной размерности Реньи

Рассмотрим четыре свойства признаков изображения, наличие которых необходимо для использования значений спектров фрактальных размерностей Реньи D_q при анализе петроглифов:

1. Близость значений D_q для фрагментов внутри каждой из двух областей — области петроглифа и области скалы на одном фотоснимке.
2. Близость значений D_q для фрагментов из области петроглифа на различных фотоснимках одного и того же петроглифа.
3. Различие значений D_q для фрагментов областей петроглифа и скалы на одном фотоснимке.
4. Различие значений D_q для фрагментов из областей петроглифа различных петроглифов.

На наличие свойств 1–4 было проанализировано 100 петроглифов (4 фотоснимка одного и того же петроглифа, итого 400 фотоснимков). В результате выяснилось, что свойство 1 выполняется для 95 петроглифов, свойство 2 — для 73 петроглифов, свойство 3 — для 89 петроглифов, свойство 4 — для 54 петроглифов. Невысокое значение для 4-го свойства интерпретируется как образование групп петроглифов на основе инвариантности структуры поверхности петроглифов.

Выводы

Результаты проведенных исследований позволяют утверждать, что спектр фрактальных размерностей Реньи можно рассматривать как инвариант петроглифа, определенный по структуре его поверхности. Кроме того, если принять гипотезу о том, что разные люди использовали разную технику выбивания петроглифов, то можно решить задачу о порядке заполнения скалы петроглифами. Используя в качестве характеристики текстуры петроглифа и скалы спектры D_q, можно также решать задачу бинарной сегментации изображений петроглифов [5].

Работа выполнена при поддержке РГНФ, проект № 05-01-12118в.

Литература

[1] *Божокин С. В., Паршин Д. А.* Фракталы и мультифракталы. — Ижевск: НИЦ «Регулярная и хаотическая динамика», 2001. — 128 с.

[2] *Встовский Г. В., Колмаков А. Г. и др.* Введение в мультифрактальную параметризацию структур материалов. — М: R&C Dynamics, 2001. — 115 с.

[3] *Шер А. Я.* Алгоритм распознавания стилистических типов в петроглифах (к теории стиля в первобытном искусстве) // Математические методы в историко-экономических и историко-культурных исследованиях. — Москва, 1977. — С. 138.

[4] *Лобанова Н. А., Саватеев Ю. А., Рогов А. А., Георгиевский И. Ю., Рогова К. А.* Петроглифы Карелии. База данных. / Институт языка, литературы и истории Карельского научного центра РАН. — «Информрегистр» Гос. регистр баз данных РС №. 10542 от 7.09.2006. — ГР №. 0220611248 от 7.09.2006. — Петрозаводск, 2006.

[5] *Спиридонов К. Н.* Применение мультифрактального анализа при изучении петроглифов Карелии // 9 межд. конф. «Интеллектуальные системы и компьютерные науки». — Москва: мех.-мат. МГУ, 2006. — Т. 2, ч. 2. — С. 278–280.

Инвариантное к ориентации и масштабу распознавание визуальных образов с использованием нечеткой нейросети

Станкевич Л. А., Хоа Н. Д.

stankevich_lev@inbox.ru, ndkhoa82@mail.ru

Санкт-Петербург, СПбГПУ

Проблема распознавания образов широко исследовалась в течение двух прошедших десятилетий. Основная цель исследований — разработка обучаемых систем распознавания образов, инвариантных к масштабу, положению и ориентации объектов.

В данной работе предлагается классифицирующая нейронная сеть SFAM (Simplified Fuzzy ARTMAP), которая является упрощенным вариантом сети FAM (Fuzzy ARTMAP) [1]. SFAM определяет принадлежность входных векторов к соответствующим классам, которым она может быть обучена. В качестве специального фильтра, используемого для предварительной обработки распознаваемых объектов, предлагается использовать What and Where (W&W) фильтр [2], с помощью которого можно определить положение, размер и ориентацию объекта в изображении.

Классифицирующая сеть SFAM

Сеть SFAM разработана специально для классификации путем удаления большой избыточности сети FAM. Структура SFAM показана на Рис. 1. Здесь нейрон P_j представляет подкласс, и его весовой вектор

Алгоритм 1. Обучение сети SFAM.

1: Определить дополнительный код из входного вектора a:
$|a| := \sum_{i=1}^{M} a_i$;
для всех $i = 1, \ldots, M$: $a_i := a_i/|a|$; $a_i^c := 1 - a_i$;
$A := (a, a^c)$;
Вход A автоматически нормализуется, поскольку $\sum_{i=1}^{2M} A_i = M$.
2: Вычислить активности нейронов и найти победительный нейрон J:
для всех $j = 1, \ldots, N$: $T_j := |A_j \wedge w_j|/(\alpha + |w_j|)$, где $\alpha > 0$;
$J = \arg\max_j T_j$;
3: **если** J — коммитированный нейрон **то**
4: **если** класс J — обучаемый класс **то**
5: $w_J := \beta(A \wedge w_J) + (1 - \beta)w_J$;
6: **иначе**
7: $T_J := 0$; новый победитель $J = \arg\max_j T_j$; повторить шаг 3;
8: **иначе**
9: $w_J := A$;

является прототипом подкласса. Узел C_k и веса w_{ij} являются меткой класса и весовым вектором нейрона в слое выходных подклассов P_j. Параметр $\rho \in [0; 1]$ — фактор вигильности.

В сети SFAM все непомеченные нейроны с единичными весами $w_{ij} = 1$, $\forall i$, названы некоммитированными. После того, как нейрон назначен в класс, он помечается и называется коммитированным. В режиме работы сети, если некоммитированный нейрон выигрывает конкуренцию со всеми коммитированными нейронами, то входной вектор относится к новому классу, см. Алгоритм 1.

W&W фильтр

В фильтре используются пространственные ориентированные рецептивные поля с различными размерами и ориентациями, которые свертываются с входным изображением. Рецептивное поле имеет ядро с ориентацией φ градусов, размером s пикселей и протяженностью a в точке (x, y), функция ядра от r изображена на Рис. 2.

$$K(x, y, \varphi, s) = (1 - r^6)\exp\bigl(-r^4/(1 + r^2)\bigr),\ r^2 = (x'/a \cdot s)^2 + (y'/s)^2 \quad (1)$$

$$x' = x\cos\varphi + y\sin\varphi; \quad y' = y\cos\varphi - x\sin\varphi \quad (2)$$

Входное изображение свертывается с каждым рецептивным полем. Активность узла A, расположенного в (x, y), рецептивное поле которого имеет ориентацию φ и размер s, определяется дискретной сверткой с вхо-

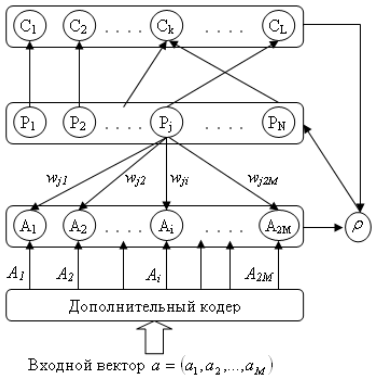


Рис. 1. Структура SFAM

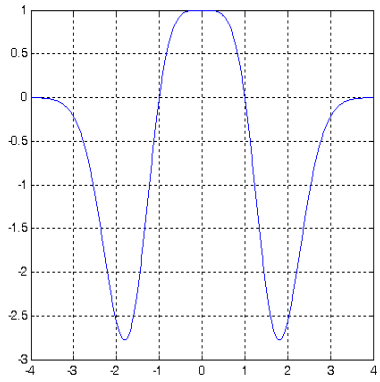


Рис. 2. Функция ядра фильтра в зависимости от r

дом $I(x, y)$. Свертка между входным изображением и фильтрами различных ориентаций и размеров образует четырехмерную матрицу нейронных узлов, которая зависит от x, y, φ и s. Активность каждого из этих узлов дает степень соответствия между фигурой объекта в изображении и рецептивным полем. Позиция, ориентация и масштаб объекта может быть оценена нахождением максимальной активности.

Когда размер объекта увеличивается, ядра фильтров, определенные в (1)–(2), дают неправильные оценки позиции, ориентации и размера объекта. Чтобы устранить эту неправильность, используется нормализация весов фильтра. При этом активность выходного узла нормированного фильтра определяется как

$$A(\varphi, s) = \sum_x \sum_y K_N(x, y, \varphi, s)\, I_C(x, y),$$

где $K_N(x, y, \varphi, s)$ — нормированная по площади рецептивного поля функция ядра фильтра, а $I_C(x, y)$ — центрированное изображение объекта.

Путем конкуренции выбирается узел, который имеет максимальную активность, и определяются параметры φ_I и s_I. В итоге, после преобразования изображения W&W фильтром, получаем изображение объекта с нормированным размером и горизонтальной ориентацией.

Эксперимент

В эксперименте использовались бинарные изображения самолета размером 128×128 пикселей. После W&W фильтра получалось изображение с размером 32×32 пикселей, и подавалось в векторном виде на вход сети SFAM. Распознавалось 10000 изображений 10 типов боевых самолетов: по

1000 изображений каждого типа с ориентацией в диапазоне $[0°; 2°]$ и масштабом в диапазоне $[0,5; 2]$. Обучающая выборка содержала 180 изображений: по 18 изображений для каждого типа самолета с масштабами от $0,5$ до 2 по отношению к их оригинальным изображениям. Обучение заняло $0,6$ секунды. Результаты эксперимента распознавания объектов с нормализованным и ненормализованным фильтрами и с различным количеством грубых фильтров в банках для простых и зашумленных изображений, показана в таблице.

Число фильтров	6	9	12
С нормализацией	$79,10\%$	$91,20\%$	$96,60\%$
Без нормализацией	$25,25\%$	$27,40\%$	$31,00\%$
С нормализацией (с шумом)	$33,70\%$	$42,30\%$	$71,60\%$
Без нормализацией (с шумом)	$14,60\%$	$18,55\%$	$24,95\%$

Эксперименты показали, что сеть SFAM и нормализованный фильтр дают лучшие результаты распознавания, поскольку точнее определяется размер и ориентации объекта.

Литература

[1] *Carpenter G. A., Grossber S.* Fuzzy Artmap : A neural network architecture for incremental supervised learning of analog multidimensional Maps // IEEE Transactions on Neural Network. — 1992. — V. 3. — Pp. 698–712.

[2] *Carpenter G. A., Grossber S. and Leshert G. W.* What-and-Where filter. A partial mapping neural network for object recognition and image understanding // Computervision and image understanding. — 1998. — V. 69, No. 1. — Pp. 1–22.

Анализ и оптимизация процедур псевдоградиентного оценивания геометрических деформаций последовательностей изображений

Ташлинский А. Г.

tag@ulstu.ru

Ульяновск, Ульяновский государственный технический университет

Задачей работы является создание и реализация на базе вычислительных средств оптимальных по различным критериям процедур оценивания параметров межкадровых геометрических деформаций изображений. Эта задача представляет как самостоятельный научный интерес, так и служит составной частью решения многих других задач обработки и анализа изображений.

В результате межкадровых геометрических деформаций изображений (МГДИ), одни и те же элементы сцены на разных кадрах изображений имеют различные координаты. Эту ситуацию можно описать деформацией сетки отсчетов, считая сцену неподвижной. Синтез процедур оценивания МГДИ невозможен без задания модели деформаций. Простейшим подходом является задание смещения элемента сцены в каждом узле $\overline{j} = (j_x, j_y)^{\mathrm{T}}$ сетки $\Omega_{\overline{j}}$ деформированного кадра относительно его положения на сетке опорного кадра, что может быть задано вектором $\overline{h}_{\overline{j}} = (h_{jx}, h_{jy})^{\mathrm{T}}$. Система таких векторов образует векторное случайное поле. Другой подход к описанию МГДИ состоит в том, что каждое положение сетки может рассматриваться как система координат. Тогда МГДИ могут быть представлены как случайное преобразование системы координат деформированного кадра в систему координат опорного кадра. Во многих случаях, когда вид МГДИ известен и описывается неким набором параметров $\overline{\alpha}$, преобразование $\overline{h} = f\left(\overline{j}, \overline{\alpha}\right)$ может быть задано в параметрической форме, что существенно облегчает его описание. Для плоских изображений при ортонормированной системе координат, когда каждой точке изображения $\mathbf{Z}$ ставится в соответствие упорядоченная пара чисел (j_x, j_y) декартовых координат, примерами могут служить евклидова, аффинная и проективная модели МГДИ. Деформированное изображение $\mathbf{S}^{(2)} = \left\{s^{(2)}_{\overline{j}}\right\}$, $\overline{j} \in \Omega_{\overline{j}}$ можно считать полученным из изображения $\mathbf{S}^{(1)} = \left\{s^{(1)}_{\overline{j}}\right\}$ посредством некоторого функционального преобразования $s^{(2)}_{\overline{j}} = f\left(\left\{s^{(1)}_{\overline{j}}\right\}, \overline{j}, \overline{\alpha}\right)$, известного с точностью до параметров МГДИ $\overline{\alpha}$. Следует также заметить, что реально наблюдаемы только кадры изображений, возмущенных помехой.

Получены оптимальные процедуры оценивания МГДИ, в частности с использованием метода максимального правдоподобия [1], однако на практике они не реализуемы, поскольку требуют колоссальных вычислительных затрат. Перспективным направлением при оценивании параметров МГДИ является использование псевдоградиентных процедур (ППП) [3]:

$$\hat{\overline{\alpha}}_t = \hat{\overline{\alpha}}_{t-1} - \mathbf{\Lambda}_t \overline{\beta}_t \left(\mathrm{J}\left(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \hat{\overline{\alpha}}_{t-1}\right)\right), \tag{1}$$

где $\overline{\alpha}$ — вектор оцениваемых параметров преобразования изображения $\mathbf{Z}^{(1)}$ в изображение $\mathbf{Z}^{(2)}$, задающий геометрические деформации изображения $\mathbf{Z}^{(1)}$; $t = 1, \ldots, T$ — номер итерации; $\mathbf{\Lambda}_t$ — матрица усиления; $\overline{\beta}_t\,(\mathrm{J})$ — псевдоградиент целевой функции (ЦФ) $\mathrm{J}\left(\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \hat{\overline{\alpha}}_{t-1}\right)$, характеризующей качество оценивания. В ППП оценивания МГДИ псевдоградиент $\overline{\beta}_t$ находится, как правило, по локальной выборке Z_t объе-

ма μ, представляющей собой отсчеты $z^{(2)}_{\overline{j}t}$ деформированного изображения $\mathbf{Z}^{(2)}$, попавшие в локальную выборку на t-й итерации, и отсчеты $\tilde{z}^{(1)}\left(\overline{j}_t,\, \hat{\overline{\alpha}}_{t-1}\right)$, взятые из некоторого непрерывного изображения $\tilde{\mathrm{Z}}^{(1)}$, полученного из $\mathbf{Z}^{(1)}$ (например, интерполяцией), где $\overline{j}_t \in \Omega_{\overline{j}t} \in \Omega_{\overline{j}}$ — координаты отсчетов $z^{(2)}_{\overline{j}t}$; $\Omega_{\overline{j}t}$ — план локальной выборки на t-й итерации.

На основе упрощения оптимальных процедур можно показать, что в предположении реализуемости ПГП в системах реального времени в качестве ЦФ целесообразно выбирать: если яркостными искажениями можно пренебречь — средний квадрат межкадровой разности; при межкадровых яркостных искажениях, близких к линейным, — выборочный коэффициент межкадровой корреляции. С точностью до постоянного множителя выражения для псевдоградиентов этих ЦФ можно записать как:

$$\overline{\beta}_t = \sum_{\overline{j}_t \in \Omega_t} \frac{\partial \tilde{z}^{(1)}\left(\overline{j}_t,\, \overline{\alpha}\right)}{\partial \overline{\alpha}} \left(\tilde{z}^{(1)}(\overline{j}_t,\, \overline{\alpha}) - z^{(2)}_{\overline{j}t} \right)\Bigg|_{\overline{\alpha}=\hat{\overline{\alpha}}_{t-1}}; \tag{2}$$

$$\overline{\beta}_t = -\sum_{\overline{j}_t \in \Omega_t} \frac{\partial \tilde{z}^{(1)}\left(\overline{j}_t,\, \overline{\alpha}\right)}{\partial \overline{\alpha}}\, z^{(2)}_{\overline{j}t}\Bigg|_{\overline{\alpha}=\hat{\overline{\alpha}}_{t-1}}. \tag{3}$$

С использованием ЦФ, соответствующих (2) и (3), получены различные классы ПГП для ситуаций заданного и неизвестного набора параметров модели МГДИ [2]. Однако вопросы их оптимизации по различным критериям исследованы явно недостаточно.

На погрешность оценок параметров МГДИ влияет большое число факторов. Если разделить их на две группы (к первой отнесем факторы, заданные априорно и не зависящие от вида ПГП: плотности распределения вероятностей (ПРВ) и автокорреляционные функции изображений и мешающего шума, вид ЦФ, а ко второй — факторы, на которые можно воздействовать при реализации ПГП: вид псевдоградиента и матрицы усиления, начальное приближение параметров и число итераций), то используя вероятности: $\rho_i^+(\overline{\varepsilon})$ — изменения оценки i-го параметра в сторону оптимального значения α_i^*, $\rho_i^-(\overline{\varepsilon})$ — от α_i^* и $\rho_i^0(\overline{\varepsilon})$ — того, что оценка останется неизменной, можно найти ПРВ оценок $\hat{\overline{\alpha}}_t$ на каждой итерации, где $\overline{\varepsilon} = \hat{\overline{\alpha}} - \overline{\alpha}^*$ — вектор рассогласования оценок. Вопросы нахождения вероятностей $\overline{\rho}_i(\overline{\varepsilon}) = \left(\rho_i^+(\overline{\varepsilon}),\, \rho_i^0(\overline{\varepsilon}),\, \rho_i^-(\overline{\varepsilon})\right)^{\mathrm{T}}$ и матрицы вероятностей одношаговых переходов вызывают наибольшую сложность. Для нахождения $\overline{\rho}_i(\overline{\varepsilon})$ можно воспользоваться результатами, полученными в работе [4], а для определения вероятностей одношаговых переходов — тем, что последовательность оценок $\hat{\overline{\alpha}}_0, \cdots, \hat{\overline{\alpha}}_t, \cdots, \hat{\overline{\alpha}}_T$, получаемая с помощью ПГП (1), является векторным марковским процессом без

последействия [5]. Заметим также, что $\overline{\rho}_i\,(\overline{\varepsilon})$ зависит не только от модели МГДИ, автокорреляционной функции изображений и параметров помех, но также и от плана локальной выборки Z_t, и может носить сложный характер с несколькими локальными экстремумами.

В работе представлены результаты:

— исследования влияние объема и плана локальной выборки отсчетов изображений, используемых для нахождения псевдоградиента ЦФ, на вероятностные характеристики изменения оценок параметров в процессе их сходимости;
— разработки методики и алгоритма расчета ПРВ погрешностей оценок параметров деформаций, полученных за конечное число итераций, на основе новой математической модели процесса псевдоградиентного оценивания при заданных вероятностных моделях изображений и мешающих шумов.

Анализируются возможности и пути:

— сокращения вычислительных затрат при расчете вероятностных характеристик оценок параметров МГДИ;
— создания методик априорной оптимизации ОЛВ, используемой для нахождения псевдоградиента ЦФ в процессе рекуррентного оценивания МГДИ, при заданных ПРВ яркостей и автокорреляционных функциях изображений;
— синтеза псевдоградиентных процедур, в которых ОЛВ автоматически адаптируется на каждой итерации до условия выполнения итерации, способствующего выходу процедур из локальных экстремумов ЦФ.

Работа выполнена при поддержке РФФИ, проект № 07-01-00138-а.

Литература

[1] *Васильев К. К., Ташлинский А. Г.* Оценивание параметров деформаций многомерных изображений, наблюдаемых на фоне помех // Распознавание образов и анализ изображений: новые информационные технологии: Труды IV Всероссийская конф. в 2 ч. — 1998. — Ч. 1. — С. 261–264.

[2] *Ташлинский А. Г.* Псевдоградиентное оценивание пространственных деформаций последовательности изображений // Наукоемкие технологии. — 2002. — Т. 3, № 1. — С. 32–43.

[3] *Цыпкин Я. З.* Информационная теория идентификации. — Москва: Наука. Физматлит, 1995. — 336 с.

[4] *Minkina G. L., Samojlov M. U., Tashlinskii A. G.* Employment of the Objective Functions in Pseudogradient Estimation of Interframe Geometric Deformations of Image // Patt. Rec. and Image Anal. — 2005. — Vol. 15, No. 1. — Pp. 247–248.

[5] *Ташлинский А. Г., Тихонов В. О.* Методика анализа погрешности псевдоградиентного измерения параметров многомерных процессов // Известия вузов: Радиоэлектроника. — 2001. — Т. 44, №. 9. — С. 75–80.

Способ оптико-электронной диагностики косоглазия

Труфанов М. И.
tmi@pub.sovtest.ru
Курск, Курский государственный технический университет

Представлен способ измерения параметров движения зрачков глаз человека, предназначенный для выявления заболеваний различного характера, связанных с отклонениями бинокулярного зрения. Способ основан на получении изображений глаз человека бинокулярной оптико-электронной системой, обнаружении на изображениях зрачков глаз, определении их трехмерных координат, измерении направлений взгляда каждого глаза и выявлении заболевания на основе анализа изменения направления движения зрачков.

Широко распространенные способы диагностики бинокулярного зрения человека практически не изменились за последние несколько десятилетий и основаны на субъективном анализе признаков заболевания врачом [1], что приводит в случае низкой квалификации врача к несвоевременному и, иногда, неправильному определению диагноза. Применение инструментальных средств диагностики позволяет точно измерять количественные признаки заболевания и объективно ставить диагноз. Для диагностики офтальмологических заболеваний наиболее целесообразным является применение оптико-электронных средств, позволяющих бесконтактно и быстро измерять параметры зрения человека и адекватно и своевременно ставить диагноз.

Недостатками известных способов и оптико-электронных средств диагностики заболеваний, связанных с отклонениями бинокулярного зрения [2, 3], являются сложность их практического применения и невозможность получения результатов исследований в реальном времени.

Предлагаемый способ диагностики отклонений бинокулярного зрения предназначен для выявления косоглазия на ранней стадии и базируется на измерении параметров саккадических движений зрачков глаз [1] при фиксации взгляда человека на заданной врачом точке по изображениям, поступающим с состоящей из двух видеокамер бинокулярной ОЭС. Отличительными особенностями способа являются: измерение трехмерных координат зрачков глаз, позволяющее с большей точностью определять степень косоглазия, а также обнаружение зрачков на изображении

способом, характеризующимся низкой вычислительной сложностью, позволяющим измерять параметры движений зрачков в реальном времени.

Способ заключается в следующем (Рис. 1). Предварительно откалиброванную (блок 1 алгоритма) ОЭС устанавливают напротив лица человека (блок 2), так чтобы каждая из видеокамер находилась примерно напротив каждого из глаз и при этом изображения обоих глаз попадали в поле зрения видеокамер. Освещают лицо человека инфракрасным светом.

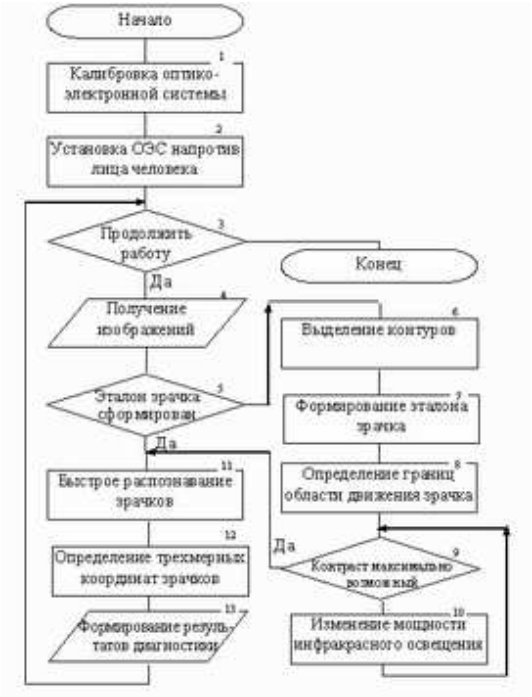


Рис. 1. Алгоритм реализации способа диагностики косоглазия

Производят выделение контуров объектов на изображении (блок 6). Сформированное контурное описание изображения используют для формирования эталона зрачка текущего анализируемого пациента, для чего на основе априорной информации об изображении зрачка и глаза производят распознавание зрачка.

В блоках 9, 10 алгоритма производят анализ контраста зрачка на изображении и установку такой мощности инфракрасного освещения, при которой контраст максимальный.

Следующей операцией (блок 8) является определение границ области движения зрачка, выполняемое для сокращения объема вычислений при обнаружении зрачка.

Операция быстрого распознавания зрачков (блок 11) предназначена для определения координат зрачка по его изображению за достаточно малое время, обеспечивающее возможность отслеживания саккадических движений глаз. Распознавание каждого зрачка производится путем сравнения изображения в области движения зрачка с эталоном, сформированным в блоке 7. Сравнение производят на основе метода, основной операцией которого является поэлементное вычитание изображения эталона из распознаваемого изображения.

Применение двух различных методов распознавания для обнаружения зрачка позволило достоверно обнаруживать зрачок любого человека за счет использования общего описания эталона зрачка и глаза (в блоке 7), а затем, по сформированному эталону зрачка конкретного человека, зрение которого анализируют, на основе метода, характеризующегося низкой вычислительной сложностью, достоверно обнаруживать его зрачки (в блоке 11) в реальном масштабе времени.

В блоке 12 производят определение трехмерных координат зрачков по изображениям, поступающих с видеокамер ОЭС, и определение угла между направления взгляда каждого глаза, характеризующего величину косоглазия. В блоке 13 формируют результаты диагностики, которые могут быть использованы для дальнейшего анализа параметров зрения человека.

Представленный способ позволяет выявлять косоглазие на ранней стадии посредством анализа саккадических движений зрачков, измерять движения зрачков и направления взгляда в реальном масштабе времени. Способ может быть применен при решении других медицинских задач, связанных с анализом движения зрачков глаз, например, для диагностики нистагма и выявлении психического состояния человека.

Литература

[1] *Урмахер Л. С.* Справочник по офтальмологической оптике и приборам. — М.: Медицина, 1971. — 179 с.

[2] Пат. № 2292836 РФ, МКИ A61B 3/08. Устройство для исследования бинокулярного зрения / *В. В. Ковылин.* – №2004137193/14; заявлено 20.12.2004; опубл. 10.02.2007, Бюл. №4. — 7с.

[3] Пат. № 2221475 РФ, МКИ A61B3/113. Способ исследования движения глаз по бинокулярному изображению и устройство для его реализации / *Д. А. Усанов, А. В. Скрипаль, А. В. Абрамов, Т. Б. Усанова, В. Б. Феклистов.* – № 2002116297/14; заявлено 19.06.2002; опубл. 20.01.2004. — 15с.

Статистическая модель деформаций отпечатков пальцев

Ушмаев О. С.
Oushmaev@ipiran.ru
Москва, Институт проблем информатики РАН

В докладе рассмотрена нетипичная для распознавания отпечатков пальцев (ОП) задача учета и моделирования возможных упругих деформаций. Получена статистическая модель деформаций, позволяющая эффективно учитывать нелинейные искажения во многих прикладных задачах распознавания ОП.

К основным искажающим факторам, негативно влияющим на соотношение ошибок ложного совпадения (второго рода, False Acceptance Rate, FAR) и ложного несовпадения (первого рода, False Rejection Rate, FRR) в системах идентификации ОП можно отнести [1]: малый размер зоны пересечения ОП, плохое качество получаемых на вход распознающей системы изображений и искажения отпечатков, вызванных упругими деформациями. Первый фактор скорее субъективный и может быть преодолен еще на этапе ввода как самим человеком в случае, когда он дружественен системе, например, при верификации своей личности, так и оператором в случае автоматизированной дактилоскопической идентификационной системы (АДИС). Малое окно является технологической характеристикой, подавление влияния шумов выполняется на этапах предобработки. Касательно учета деформаций, впервые использовать методы теории упругости [2] для восстановления упругих деформаций ОП было предложено в работе [3], ранее использовали либо модели тонкой пластины [4], слабо отражающих механику деформаций отпечатков, либо эмпирические модели [1]. Получаемые при этом преобразования в общем случае выглядят неестественно, особенно при экстраполяции.

Малые упругие деформации довольно точно описывается решением следующего линейного уравнения механики деформируемого тела [2]

$$\mu\nabla^2\mathbf{u} + (\lambda+\mu)\nabla\mathrm{div}\,\mathbf{u} = -\mathbf{F}, \tag{1}$$

где $\mathbf{u}$ — карта смещений; $\mathbf{F}$ — внешняя сила; λ и μ — коэффициенты упругости Ламе. Очевидной проблемой применения уравнения (1) является принципиальная невозможность измерения внешних сил. В [3] на основе уравнения (1) был предложен метод вычислений деформаций ОП, использующий для приближения правой части уравнения данные о соответствии контрольных точек ОП. В [5] приведены методы быстрого решения (1) и алгоритм внедрения учета деформаций в произвольные алгоритмы распознавания ОП. В [6] показано значительное улучшение

распознавания ОП за счет учета упругих деформаций в терминах ошибок 1-го и 2-го родов.

Однако результаты [3,5,6] не решают значительных проблем применения механических методов — трудоемкость вычислений и большой размер описания деформаций. Далее рассмотрена проблема статистического анализа деформаций ОП с целью сократить размерность описания.

В качестве исходных данных были рассмотрены измерения деформаций методами [3] на открытых базах отпечатков пальцев FVC2002 (DB1, DB2, DB4), всего 8400 примеров. Для сокращения размерности описания использовался метод главных компонент в частотной области. Прямое вычисление главных компонент на прореженной карте смещений размерностью 936 узлов ожидаемо не привело к положительному результату. Причиной этого является довольно сильные корреляции и относительно небольшой объем выборки.

Для повышения эффективности вычислений к картам смещений применялось двумерное преобразование Фурье. В частотной области 92 низкочастотные компоненты составляют 93% энергии карты смещений и 98% энергии во внутренней части изображения (без учета края, где доминируют маргинальные эффекты). В дальнейшем для полученных данных в частотной области вычислялись главные компоненты. Предварительный анализ полученных в спектральной области главных деформаций показал, что дисперсия остатков первых 4 главных деформаций составляет 15%, а первых 20 — только 2%, что позволяет сделать заключение о небольшой «внутренней» размерности деформаций. Визуализация главных компонент в пространственной области позволяет однозначно интерпретировать искажающие факторы: микроповорот (возникающий из-за невозможности точно факторизовать деформации по движениям), продольный сдвиг, поперечный сдвиг и деформационный поворот (поворот пальца со значительным давлением на поверхность сканера).

Полученный массив главных деформаций является ортонормированным, поэтому произвольная деформация D может быть разложена в сумму вычисленных главных компонент деформаций. Проведенный анализ показывает, что для эффективного учета деформаций достаточно использовать первые 10–12 компонент. Такое существенное снижение размерности описания искажений позволяет решить или повысить эффективность существующих решений многих прикладных проблем распознавания ОП.

1. Синтетические базы данных. Исследование синтетической базы FVC 2002 DB4 показало отличие статистики деформаций от естественных приложений. Модель позволяет более эффективно имитировать есте-

ственные деформации ОП при создании искусственных тестовых массивов.

2. Имитация множества приложений на этапе сравнения. В [7] показано, что использование для обучения нескольких приложений ОП значительно улучшает качество распознавания. Приведенная модель позволяет имитировать несколько приложений при регистрации ОП.
3. Улучшение шаблона ОП. В [8] приведена концепция выбора наименее деформированного состояния ОП. Приведенная модель эффективно решает данную задачу, так как дает численную меру деформации.

Работа выполнена при поддержке РФФИ, проект № 07-07-00031.

Литература

[1] *Halici U., Jain L. C., Erol A.* Introduction to Fingerprint Recognition, Intelligent Biometric Techniques in Fingerprint and Face Recognition. — CRC Press, 1999.

[2] *Ландау Л. Д., Лифшиц Е. М.* Теоретическая физика, том VII. Теория упругости. — М.: Физматлит, 2001.

[3] *Ushmaev O., Novikov S.* Registration of Elastic Deformations of Fingerprint Images with Automatic Finding of Correspondences // Proc. MMUA03, Santa Barbara, CA, 2003. — Pp. 196–201.

[4] *Bazen A. M., Gerez S. H.* Thin-Plate Spline Modelling of Elastic Deformation in Fingerprints // 3rd IEEE Benelux Signal Processing Symposium, 2002.

[5] *Novikov S., Ushmaev O.* Registration and Modelling of Elastic Deformations of Fingerprints // Int'l Workshop Biometric Authentication (ECCV8, BioAW-2004), Prague, Chezh Republic, May2004. — Springer, 2004. — Pp. 80–88.

[6] *Ushmaev O., Novikov S.* Efficiency of Elastic Deformation Registration for Fingerprint Identification // 7th Int'l Conf. on Pattern Recognition and Image Analysis (PRIA-2004). — Vol. 3, — St. Petersburg: SPbETU, 2004. — Pp. 833–836.

[7] *Ушмаев О.С., Синицын И.Н.* Опыт проектирования многофакторных биометрических систем // Труды VIII международной конференции «Кибернетика и высокие технологии XXI века». — Т. 1. — С. 17–28.

[8] *A. Ross, S. Dass and A. K. Jain* Estimating Fingerprint Deformation // Int'l Conf. on Biometric Authentication (ICBA), Hong Kong. — Springer Publishers, 2004. — LNCS, Vol. 3072. — Pp. 249–255.

Об одном алгоритме определения местонахождения лица и координат зрачков на изображении

Фазылов Ш. Х., Мирзаев Н. М., Тухтасинов М. Т.
shavkat-faz@mail.ru, mnm2005@rambler.ru, mumtozali@yahoo.com
Ташкент, Институт математики и информационных технологий АН Руз

Одним из перспективных направлений современных информационных технологий является создание систем распознавания личности по изображению лица. Основная проблема в процессе создания систем распознавания личности человека по изображению лица — определение местонахождения лица и его элементов на изображении [1, 2, 3]. В данном докладе рассматривается один из алгоритмов определения местонахождения лица и координат зрачков на изображении.

Алгоритм определения местонахождения лица с использованием корреляционного анализа

Данный алгоритм основан на оценке близости между частью изображения и эталоном [4]. Эталонное изображение формируется из средних значений пикселей изображений лиц, принадлежащих N разным людям (женщинам, мужчинам, молодым, старым и т. д.). Для формирования эталонного изображения используется центральная часть изображений лица. На рис. 1 показано эталонное изображение, в котором $N = 50$. В качестве меры близости используется коэффициент корреляции (КК) между образом в поле маски и эталоном. На основе оценки близости между эталоном и соответствующей частью изображений осуществляется поиск области лица на исходном изображении.

Рис. 1. Маска для поиска лица.

При этом предполагается, что размеры изображения лица на исходном изображении априори неизвестно. В связи с этим размеры эталонного изображения плавно изменяются в процессе вычисления меры близости, обычно в сторону уменьшения. Это зависит от того, как мы выбираем размеры (максимальное или минимальное) эталонного изображения. При каждом изменении размера эталона процесс вычисления КК осуществляется заново и определяется его максимальное значение. Данная процедура повторяется до тех пор, пока размеры эталонного изображе-

ния не будут меньше некоторого размера, от которого зависит конкретная задача.

После этого определяются максимальный КК и соответствующие области лица. Так как в эталонном изображении априори известно местонахождение зрачков, мы можем приблизительно определить их координаты. Однако во многих случаях эти координаты не соответствуют координатам зрачков исходного изображения.

Для решения данной задачи разработан алгоритм, позволяющий уточнить координаты зрачков.

Алгоритм поиска зрачков с помощью преобразования Хафа

Идея данного алгоритма состоит в том, что в рассматриваемой области находятся координаты центра окружности, описанной наиболее тёмными пикселями изображения. Для осуществления поиска уточненных координат зрачков выполняются следующие простые процедуры.

Определение основных параметров проверяемых окружностей. Минимальные и максимальные радиусы $R_{\min}$, R_{max}:

$$R_{\min} = W_{\text{mask}} \cdot c/100;$$
$$R_{\max} = W_{\text{mask}} \cdot d/100;$$

где W_{mask} — ширина маски; c, d — параметры, определяющие минимальные и максимальные радиусы.

Границы для области поиска M_{xy}:

$$LS_x = E_x - kR_{\max};$$
$$RS_x = E_x + kR_{\max};$$
$$TS_y = E_y - R_{\max};$$
$$BS_y = E_y + kR_{\max};$$

где LS_x, RS_x, TS_y, BS_y — левая, правая, верхняя, нижняя границы соответственно; E_x, E_y — координаты зрачков; k — параметр, определяющий область поиска.

Минимальное значение яркости в области поиска:

$$G_{\min} = \min_{\substack{x=LS_x,\ldots,RS_x \\ y=TS_y,\ldots,BS_y}} (M_{xy})$$

Определение и уточнение координат зрачков. В области поиска строятся окружности с радиусом (первоначально $R = R_{\min}$ и

$P_{\max} = 0$) и вычисляется параметр P для каждой окружности:

$$P = \frac{\sum\limits_{x}\sum\limits_{y}[255 - G_R(x,y)]}{(255 - G_{\min}) \cdot \pi \cdot R^2} \cdot 100,$$

где $G_R(x,y)$ — значения яркости точки с координатами (x,y), принадлежащей окружности с радиусом R.

Определяется максимальное значение из (P_i), вычисленных для всех окружностей с радиусом R и соответствующие координатам зрачков:

$$P_R = \max\{P_1, P_2, \ldots, P_N\}$$

Проверка условия $P_R > P_{\max}$. Если условия выполняются, то $P_{\max} = P_R$, и определяются соответствующие координаты (E_x, E_y). В противном случае, вычисляется новый $R = R + 1$. Если $R \leqslant R_{\max}$, то процедура повторяется для нового R.

Здесь необходимо отметить, что для повышения точности результата алгоритма предварительно выполнена медианная фильтрация [5] области поиска.

Разработанные алгоритмы могут быть использованы при создании системы контроля доступа, системы видеонаблюдения, системы поиска в базе данных по фотопортрету человека, системы проверки документов, удостоверяющих личность человека.

Литература

[1] *Кухарев Г. А.* Биометрические системы: Методы и средства идентификации личности человека. — СПб.: Политехника, 2001. — 240 с.

[2] *Самаль Д. И.* Построение систем идентификации личности на основе антропометрических точек лица // Сб. науч. тр. «Цифровая обработка изображений», Минск. Ин-т техн. киберн. НАН Беларуси, 1998. — С. 72–78.

[3] *Фазылов Ш. Х., Тухтасинов М. Т., Старовойтов В. В., Самаль Д. И., Ригол Г.* Локализация фрагментов лица на цветных фотопортретах // Обработка информации и управление в чрезвычайных и экстремальных ситуациях: тез.докл. 4-й Межд. научн. конф., Минск, 2004. — С. 218–223.

[4] *Brunelli R., Poggio T.* Face recognition: features versus templates // IEEE Trans. on Patt. Anal. and Mach. Intel. — 1993. — Vol. 15, № 10. — P. 1042–1052.

[5] *Прэтт У.* Цифровая обработка изображений. — М.: Мир, 1982. — 792 с.

К решению проблемы визуализации и анализа 3D сцен, распознавания пространственных образов методами кватернионного исчисления

Фурман Я. А., Хафизов Д. Г., Рябинин К. Б.
rts@marstu.mari.ru
Йошкар-Ола, ГОУ ВПО Марийский гос. тех. университет

Рассмотрено применение аппарата кватернионных сигналов для обработки пространственных групповых точечных объектов и точечных полей.

Характеристика рассматриваемой проблемы обработки пространственных изображений

Переход от плоских изображений пространственных объектов к их трехмерным изображениям увеличивает количество получаемой информации и повышает потенциальные возможности синтезируемых алгоритмов. При этом для изображений объектов одного класса возрастают вычислительные значения мер их схожести, а для изображений разных классов — степень их ортогональности. Хотя исследования в этой области имеют многолетнюю историю, факторы, вызванные увеличением размерности обрабатываемых сигналов, в первую очередь, рост объема требуемых вычислений и отсутствие эффективных алгоритмов, и в настоящее время стимулируют интенсивные исследования по получению новых подходов к обработке 3D сцен.

Создание исчисления, позволяющего оперировать трехмерными геометрическими образами по правилам алгебры, всегда были актуальной задачей. Первые важные результаты в этом направлении получены У. Гамильтоном, который ценой отказа от свойства коммутативности операции произведения, в поисках объектов, обобщающих комплексные числа, открыл кватернионы. Сотрудниками лаборатории обработки изображений групповых точечных объектов и точечных полей Марийского государственного технического университета на базе алгебры кватернионов и положений теории обработки сигналов получен математический аппарат для решения задач, связанных с пространственными точечными изображениями. По сравнению с другими применяемыми в этой области подходами, он обладает следующими преимуществами:

1) более высокая информативность получаемой меры схожести/различия 3D изображений;
2) формируемые модели сигналов представляют собой одномерные векторы;
3) простая реализация такой массовой операции как вращение 3D вектора.

Рассматриваемый в докладе подход к решению проблемы обработки 3D изображений основан на применении этого аппарата к пространственным точечным изображениям двух типов: тип A — генеральное большой мощности множество точек, расположенных на поверхности объемных тел в пространственной сцене и тип B — пространственный групповой точечный объект (ПГТО) с небольшим количеством точек.

Базовое преобразование точечных множеств

Переход в третье измерение дает возможность установить связи между точечными объектами не только по расстоянию между ними, как в плоском случае, но и по их принадлежности к таким геометрическим образам, как плоскости. Это является основой сегментации 3D точечных изображений и реализуется путем вычисления трехмерного градиента. Более просто подобная сегментация выполняется на базе преобразования кластеризации точек множества (КТМ). Преобразование КТМ задается нормированной гиперкомплексной частью (бивектором) скалярного произведения двух разностных векторных кватернионов

$$\psi(m) = \mathrm{hyp}(\Delta a(m)), \quad \Delta a(0) = r(m), \quad |r(m)| = 1, \quad m = 0, 1, 2, \ldots$$

Здесь $\Delta a(m) = a_m - a_0$, $\Delta a(0) = a_1 - a_0$, где a_m — текущая, a_0 и a_1 — фиксированные точки генерального множества, $r(m)$ — нормаль к локальной плоскости, образованной этими точками. В результате генеральное множество делится на L подмножеств (граней), точки которых характеризуются близкими значениями нормалей

$$r(0), \quad r(1), \ldots, \quad r(m), \ldots, \quad r(L-1).$$

Обработка сцены типа A

Визуализация сцены типа A связана с получением её глобального образа в привычном для восприятия этой сцены человеком виде. После преобразования КТМ действия направлены на замену точечной сцены сценой из изображений исходных объектов набором вложенных аппроксимированных многогранников с плоскими гранями. Для каждого кластеризованного подмножества проводятся следующие операции: захват плоскости грани, формирование и анализ сферических координат точек нормалей, локализация пиков гистограммы, обнаружение точек нормалей к анализируемой грани.

Дальнейшая обработка связана с операциями реконструкции сцены: выделение краевых точек граней, формирование кватернионного описания их контуров, обнаружение смежных граней, учет особых точек, структурный анализ формы граней, фильтрация контуров граней.

Конечным результатом обработки служит аналитическая модель сцены в виде кватернионного сигнала, используемого для решения задач распознавания и оценки параметров.

Обработка сцены типа B

Задачи распознавания и оценки параметров ПГТО могут быть решены при условии упорядочения его отметок. Эта процедура устанавливает закономерность перебора отметок для формирования векторно-кватернионной модели ПГТО. Процедура упорядочения реализуется путем построения семейства вложенных выпуклых многогранников, ассоциированных с ПГТО и выполняется на базе преобразования КТМ. Доказаны теоремы существования и единственности процедуры. Для получения модели ПГТО в виде кватернионного сигнала по критерию максимума расстояния между гранями производится нумерация граней, а затем и точек ПГТО. Пронумерованные точки являются вершинами ориентированного 3D многоугольника, на основании аналитического описания которого проводится распознавание и оценка параметров ПГТО.

Работа выполнена при поддержке РФФИ, проект № 07-01-00058а.

Литература

[1] Комплексные и гиперкомплексные системы в задачах обработки сигналов / под ред. *Я. А. Фурмана.* — Москва: ФИЗМАТЛИТ, 2004.

[2] *Фурман Я.А.* Визуализация изображений в трехмерных сценах. Учебное пособие. — Йошкар-Ола: МарГТУ, 2007.

Оценка количества информации изображения в детерминированном подходе

Харинов М. В.

khar@iias.spb.su

Санкт-Петербург, СПИИ РАН

В современных информационных технологиях вообще, и в области защиты сигналов (изображений) в частности, не достает простых способов формализации понятия информации. При этом:

— понятия «сигнал», «информация» и «смысл информации» недостаточно четко разграничиваются между собой;
— при оценке количества информации обычно не уточняется понятие самой информации;
— недостаточно используются возможности отслеживания распределения информации по амплитудным отсчетам сигнала для управления его обработкой;

— ограниченное внимание уделяется способам экспериментальной проверки рассчетных значений локальной оценки количества информации.

Недостаточная формализация понятия информации ограничивает развитие общих методов анализа сигналов, и для эффективной обработки изображения либо требуется заранее знать, хотя бы приблизительно, что на нем изображено, либо прибегать к трудоемкой процедуре настройки программной системы на обработку изображений требуемого типа. Формализация понятия информации особенно необходима для решения стеганографических и других задач обратимого встраивания одного сигнала в другой при условии, что оба сигнала при приеме и разделении считаются неизвестными.

Перечисленные недостатки преодолеваются в модели изображения, предложенной в СПИИ РАН, которая строится и экспериментально обосновывается в рамках объединения подходов к пониманию информации, выдвинутых ведущими советскими учеными А. Н. Колмогоровым [1], Ф. Е. Темниковым [2, 3] и Н. П. Брусенцовым [4, 5] еще в годы становления информатики.

Модель изображения с виртуальной памятью

В модели [6, 7, 8] цифровое изображение рассматривается как запоминающая среда и средство для передачи кодированной информации, обладающее самостоятельной троичной «виртуальной» памятью. Считается, что виртуальная память состоит из ячеек, подобных ячейкам памяти компьютера, в которых хранятся исходные отсчеты яркости изображения. Каждому пикселу сопоставляется самостоятельная ячейка виртуальной памяти. Ячейки виртуальной памяти образуются из последовательных тритов различного разряда аналогично тому, как байты компьютерной памяти состоят из битов.

Виртуальная память рассчитывается по изображению, и по сравнению с памятью, занимаемой изображением в компьютере, обычно имеет большее число разрядов. Запоминающие элементы (триты) виртуальной памяти подразделяются на переменные (read-write), значения которых допускается изменять при обработке изображения, и фиксированные (read-only), которые при модификации переменных тритов сохраняются и обеспечивают восстановление изображения в некотором упрощенном виде. Под количеством информации, содержащейся в данном отсчете яркости, понимается число переменных тритов, которое определяет объем переменного «сообщения». Помимо модификации переменных тритов в модели предусматривается их преобразование в фиксированные, при котором количество информации в рассматриваемом отсчете ярко-

сти уменьшается, что выражается в уменьшении в изображении числа различных пикселов.

Как показывают эксперименты [6, 8], виртуальная память способна хранить коды информации независимо от предусмотренных линейных и нелинейных преобразований или, например, затухания сигнала в процессе передачи, а также при определенных условиях сохранять информацию устойчиво к возможным искажениям сигнала.

Информация разделяется в виртуальной памяти на явную и неявную компоненты. Объем неявной компоненты обычно составляет не менее 30–50% объема изображения, что позволяет дублировать в неявной компоненте коды явной информации для ее восстановления в случае повреждения за счет повторений в последовательных разрядах виртуальной памяти, которые сопоставляются вычисляемым по изображению вложенным диапазонам шкалы яркости. При этом встраивание в изображение собственной информации имеет самостоятельное значение, поскольку его необходимо выполнять при формировании каждого разряда виртуальной памяти для подавления искажений, возникающих в младших разрядах. В задачах сжатия изображений самовстраивание оказывается полезным для повышения коэффициента сжатия при сохранении качества зрительного восприятия и результатов автоматической обработки.

Перспективы применения

Помимо изображений модель применима к аудиосигналам, а также к сигналам иной природы и, помимо технологии восстановления изображений на основе самовстраивания, оказывается эффективной для решения различных прикладных задач:

— неявной передачи в составе аудиосигнала сопутствующего видеосопровождения без существенного конструктивного изменения передающих и приемных устройств;
— автоматизации распознавания сигналов на основе анализа особенностей распределения информации по амплитудным отсчетам сигнала;
— повышения пропускной способности или надежности передачи по каналам связи за счет максимального использования емкости несущего сигнала;
— защиты электронных и обычных документов (например, денежных знаков) при условии индивидуальной защиты каждого отдельного документа без создания базы данных.

В дальнейшем в рамках модели сигнала с виртуальной памятью [6, 7, 8] предполагается разработать способ выделения объектов в представлении изображения с равномерным распределением количе-

ства информации по градациям яркости, а также развить приложения модели в задачах хранения, передачи и обработки сигналов.

Работа выполнена при поддержке РФФИ, проект №06-07-95007, и в 2007 г. поддерживается грантом международного фонда «Human Capital Foundation» (www.hcfoundation.ru).

Литература

[1] *Колмогоров А. Н.* Три подхода к определению понятия «Количество информации» // Пробл. передачи информации. — 1965. — Вып. 1, Т. 1. — С. 3–8.

[2] *Темников Ф. Е.* Информатика // Известия вузов. Электромеханика. — 1963. — №11. — С. 1277.

[3] *Темников Ф. Е., Афонин В. А., Дмитриев В. И.,* Теоретические основы информационной техники. — М: Энергия, 1979. — 512 с.

[4] *Брусенцов Н. П.* Вычислительная машина «Сетунь» Московского государственного университета. В кн.: Новые разработки в области вычислительной математики и вычислительной техники. — Киев, 1960. С. 226–234.

[5] *Брусенцов Н. П.* Реставрация логики. — М.: Новое тысячелетие, 2005. — 165 с.

[6] *Харинов М. В.* Запоминание и адаптивная обработка информации цифровых изображений. — СПб: Изд-во С.-Петерб. ун-та, 2006. — 138 с.

[7] *Харинов М. В.* Адаптивное встраивание водяных знаков по нескольким каналам. Заявка на патент РФ № 2006119273 (заявители: СПИИРАН–«Самсунг Электроникс Ко., Лтд.»), 2006. — 180 с.

[8] *Kharinov M. V.* Representation of Image Information for Computer Calculations // Pattern Recognition and Image Analysis. — 2007. — Vol. 17, №1. — P. 117–124.

Распознавание 3D изображений групповых точечных объектов по их проволочным моделям на основе кватернионного исчисления

Хафизов Д. Г., Рябинин К. Б.

rts@marstu.mari.ru

Йошкар-Ола, ГОУ ВПО Марийский гос. тех. университет

Предложено решение проблемы распознавания 3D изображений пространственных групповых точечных объектов (ПГТО) с неизвестной нумерацией точек. Разработана программа, позволяющая реализовать алгоритмы упорядочения и распознавания ПГТО по их проволочным моделям при воздействии координатных шумов. Построение проволочных моделей зашумленных ПГТО возможно при объединении граней синтезированного многогранника с близкими нормалями.

Постановка задачи

Широкий спектр задач обработки изображений, в том числе и радиолокационных, связан с обработкой групп точечных объектов. Обработка плоских изображений таких объектов подразумевает решение следующих задач: обнаружение, локализация, измерение групповых признаков, упорядочение точек в групповом объекте, кодирование группового точечного объекта, распознавание и оценка параметров, решению которых посвящено ряд ранее опубликованных работ [1]. Однако при переходе в третье измерение, т. е. при обработке пространственно расположенных групповых точечных объектов применение ранее известных алгоритмов невозможно. При этом одной из важнейшей и наиболее трудных среди перечисленных является задача упорядочения точек в пространственном групповом точечной объекте, так как от успешного решения данной задачи зависит решение таких задач, как распознавание и оценка параметров. Для решения данной задачи применяется теория кватернионного анализа, что обусловлено удобством описания ПГТО векторными кватерниоными сигналами с позиций их последующей обработки и анализа с точки зрения теории сигналов.

Формирование проволочной модели зашумленного пространственного группового точечного объекта

Формирование проволочной модели пространственного группового точечного объекта, необходимой для задания порядка точек в объекте и формирования кватернионного сигнала, основано на том, что любую совокупность точек в пространстве можно ассоциировать с совокупностью вложенных друг в друга выпуклых многогранников. Подход к решению данной проблемы в отсутствии координатных шумов рассмотрен в [3].

При появлении координатного шума проволочная модель ПГТО может разрушиться, что обусловлено разрушением граней многогранника из-за невыполнения условия принадлежности всех точек грани плоскости.

Синтез выпуклого многогранника по заданному множеству точек является многоэтапной процедурой. На каждом этапе выделяется одна из граней многогранника и упорядочиваются точки множества, лежащие в пределах этой грани (плоскости). Полученный многогранник описывается с помощью графа, задающего связи между гранями и является ее математической моделью с уже известным порядком вершин в пределах каждой грани и аналитическим представлением всех граней в виде их контуров. Также известны нормали, площади и периметры граней. Так как в условиях воздействия координатных шумов некоторые из гра-

ней многогранника могут разрушиться, то необходимо ввести условия, по которому эти распавшиеся грани будут объединены в одну общую грань.

В качестве такого условия можно ввести критерий близости нормалей, построенных к граням. Задавшись порогом по расстоянию между векторами нормалей, можно объединить близкие грани и получить многогранник, близкий по своим характеристикам к исходному.

Данный результат позволяет использовать алгоритм построения проволочной модели и упорядочить точки ПГТО в условиях воздействия координатных шумов, что дает возможность получить модель ПГТО в виде одномерного вектора, компонентами которого являются векторные кватернионы.

Процесс дальнейшей обработки — упорядочение граней и точек — совпадает с алгоритмом упорядочения точек ПГТО в отсутствии шумов.

Решение задачи распознавания ПГТО производится по величине меры схожести с эталонным сигналом, введенной на основе скалярного произведения кватернионных сигналов (КТС) [2]. Предполагается, что распознаваемый (сигнальный) КТС принадлежит к одному из M классов. Решение задачи распознавания КТС осуществляется на основе критерия минимума расстояния между сигнальным и эталонным КТС. На основе данных алгоритмов было разработано программное обеспечение, позволяющее проводить исследование влияния координатных шумов на результаты обработки ПГТО. По результатам моделирования данных алгоритмов были получены характеристики правильного распознавания ПГТО.

Заключение

Применение кватернионов для описания моделей пространственных групповых точечных объектов позволяет строить оптимальные, с позиций теории обработки сигналов, алгоритмы распознавания 3D изображений групповых точечных объектов. Рассмотренные в работе алгоритмы были проверены методом машинного моделирования, что позволяет говорить об их работоспособности и практической значимости для решения задач подобного рода.

Работа выполнена при поддержке РФФИ, проект № 07-01-00058а.

Литература

[1] Введение в контурный анализ и его приложение к обработке изображений и сигналов / Под ред. *Я. А. Фурмана.*— М.: Физматлит, 2002.

[2] Комплексные и гиперкомплексные системы в задачах обработки сигналов / Под ред. *Я. А. Фурмана.*— М.: Физматлит, 2004.

[3] *Фурман Я. А.* Визуализация изображений в трехмерных сценах. Учебное пособие. — Йошкар-Ола: МарГТУ, 2007.

Применение методов распознавания образов для сжатия видеоинформации

Хашин С. И.
khash2@mail.ru
Иваново, ИвГУ

На сегодняшний день имеется большое количество алгоритмов сжатия видеоинформации [1, 2, 3, 4]. Но в последнее время появляется возможность добиться существенного прогресса в этой области. Основой этого являются возросшие вычислительные мощности, позволяющие использовать методы распознавания образов на изображении.

Математические методы

В отличие от других областей сжатия информации, при сжатии видео одной из основных проблем является требуемая вычислительная мощность. Используемые алгоритмы должны ориентироваться на процессоры, имеющиеся в наличии у пользователей, разрабатываемые — ориентироваться на перспективные процессоры. Учитывая довольно большой срок разработки новых алгоритмов и стандартов видеосжатия, можно ожидать, что вычислительная мощность не станет препятствием для их реализации. Поэтому для обработки видеоданных можно применить значительно более мощные и сложные методы, по сравнению с используемыми сегодня.

Во-первых, сегментация изображения. Применяемые на сегодняшний день алгориты разбивают изображение на квадраты или прямоугольники, и для каждого из них находят «движение» (точнее говоря, вектор сдвига) по отношнению к другим кадрам. Вместо этого можно разбивать изображение на сегменты произвольного вида.

Во-вторых, переход от междукадрового «движения» в виде сдвига к движениям более общего вида: сдвигам с поворотом и растяжением, произвольным аффинным и проективным преобразованиям плоскости.

В нынешних алгоритмах видеосжатия каждый рассматриваемый сегмент имеет совсем небольшой размер (обычно не более 16×16 точек), поэтому для него рассмотрение лишь простейших движений в виде сдвига вполне достаточно. Если же мы рассматриваем сегменты произвольного размера, то переход к более общим аффинным движениям неизбежен, а к проективным — возможен.

В третьих, для более эффективного нахождения движений сегементов и их кодирования требуется применить методы распознавания образов. Более конкретно, надо сконструировать «объекты» в виде объединения нескольких сегментов на изображении. Каждый объект не обязательно должен двигаться как единое целое, но движения всех его частей должны

быть связаны друг с другом. Целью данного шага является снижение количества запоминаемых движений, ориентировочно с нескольких сотен до нескольких десятков.

Вычислительно наиболее сложные алгоритмы, например, использующие методы искусственного интеллекта, будут применяться лишь на этапе кодирования, то есть когда нет жесткого ограничения по времени. Процесс же декодирования будет значительно более простым и вполне может уложиться в реальное время.

Для демонстрации эффективности предлагаемого метода видеосжатия надо сосредоточиться в основном на математических подходах и показать их работоспособность хотя бы в простейшей ситуации. Более точно, примем следующие ограничения:

— Не рассматриваем кодирование звука.
— Не реализуем полный кодер/декодер. Ограничиваемся минимально возможным вариантом: кодер сжимает заданную последовательность bmp или jpeg-файлов в один файл, декодер — распаковывает сжатый файл в цепочку bmp-файлов.
— Не ставим задачу построения быстрого кода. Для демонстрационной версии будет достаточно, если и кодер/декодер смогут обрабатывать один кадр за несколько минут (или даже десятков минут).
— Для сжатия «дифференциальных кадров», то есть разностей между кадром-прогнозом и точным кадром используем готовый алгоритм, например jpeg2000.
— Для сжатия «дискретных данных» (в основном, это будут вектора движения, коэффициенты аффинных и проективных преобразований) применяем готовый (свободный) алгоритм, например bzip2.

Схема алгоритма

В алгоритмах сжатия видеоданных принято делить все кодируемые кадры на 3 типа:

— I-кадры (Initial), кодируются независимо от остальных;
— P-кадры (Predicted), кодируются на основании прогноза с помощью «движений» от предыщущих I- и P-кадров;
— и, наконец, B-кадры (Bidirectional), кодируются на основании прогноза от соседних (предыщущих и последующих) I- и P-кадров;

I-кадры имеют наибольший объем, служат для начал декодирования видеоданных. При декодировании на основе I-кадра вычисляется цепочка последующих P-кадров, а затем вычисляются промежуточные B-кадры. Типичная последовательность кадров в видеопотоке такова:

$$\mathrm{IBBPBBPBB}\ldots\mathrm{IBBPBB}\ldots$$

P-кадры обычно имеют объем примерно на порядок меньше, чем I-кадры и B-кадры меньше еще в несколько раз. Использование B-кадров не обязательно.

В начальном варианте нашего алгоритма откажемся от использования B-кадров. Для кодирования начальных (I) кадров будем применять стандартный алгоритм Jpeg2000. Таким образом, остается описать лишь алгоритм кодирования P-кадров.

Кодирование P-кадров состоит из следующих шагов:

— разделение кадра на цветовые сегменты;
— поиск движения каждого найденного сегмента на следующих кадрах;
— выделение объектов, состоящих из сегментов, методами распознавания образов;
— построение «прогноза» следующего кадра по имеющимся сегментам и их движениям;
— добавление к кадру-прогнозу разности, сжатой (после некоторой предварительной обработки) алгоритмом jpeg-2000.

Нахождения цветовых сегментов. Задача сегментации изображений достаточно хорошо изучена. Для ее решения предложено несколько различных подходов. В настоящее время сегменты строятся на основе границ, полученных с помощью алгоритма «Canny edge detector». Ожидаемое количество сегментов — от нескольких десятков до нескольких сотен, типичное значение 200–500.

Нахождение движений. Классические методы нахождения движений (например, алгоритм Лукаса-Канады) дают в ответе лишь «вектор движения», т. е. под «движением» подразумевается лишь сдвиг. Такой подход полностью оправдан, если мы находим движение лишь небольших объектов (как квадрат 8×8). В нашем же случае объект может занимать большую часть кадра и для описания его движения требуются аффинные или даже проективные преобразования. Для этого были разработаны модификации метода Лукаса-Канады, позволяющие находить все требуемые типы движений плоскости.

Распознавание образов. Количество сегментов, а значит и их движений, слишком велико для эффективного их хранения в коде. Поэтому на изображении выделяем «объекты», состоящие из сегментов. Каждый объект либо двигается как единое целое, либо движения его частей зависят друг от друга по некоторому правилу. Поскольку все данные лишь приближенные, точного ответа быть не может, приходится применять некоторый эвристический алгоритм, методы распознавания образов.

Ключевым моментом в этой ситуации является то, что мы имеем численную характеристику качества распознавания. В полном виде это раз-

мер полученного файла. На практике, для достижения приемлемой скорости, этот размер приходится заменять некоторой аппроксимацией.

Другая особенность — дискретность задачи: каждый объект представляется в виде объединения небольшого (первые десятки) количества соседних сегментов. Она же дает возможность «наращивать» объекты, присоединяя на каждом шаге по одному сегменту к некоторому объекту.

Для реализации алгоритма используется стандартная двухуровневая нейронная сеть.

Кодирование движений. В текущей версии алгоритма используются аффинные преобразования плоскости:

$$x' = a_0 + a_1 x + a_2 y,$$
$$y' = a_3 + a_4 x + a_5 y.$$

Описание преобразования с помощью коэффициентов a_i не является оптимальным с точки зрения требуемого количества памяти. В частности, практически невозможно сформулировать требования к точности представления коэффициентов a_i. В разрабатываемом алгоритме аффинное преобразование задается с помощью других шести коэффициентов, зависящих еще и от рассматриваемого сегмента. Более точно, пусть A_0 — центр тяжести сегмента, r — среднеквадратичное расстояние его точек от центра, A_1, A_2 — точки, лежащие на расстоянии r над A_0 и вправо от неё. Тогда аффинное преобразование описывается тремя векторами: вектором сдвига для точки A_0; разницей векторов сдвига для точек A_1 и A_0; разницей вектора сдвига для точки A_2 и его прогноза по движению точек A_0 и A_1. При таком подходе требования к точности всех коэффициентов вполне определенные, и можно ограничиться четырьмя двоичными знаками после запятой.

Завершение кодирования. Результатом предыдущих этапов кодирования являются два потока данных: коэффициенты используемых аффинных преобразований и «дифференциальные кадры», содержащие разность между построенным кадром-прогнозом и точным кадром. В распространенных алгоритмах сжатия видео применяются специализированные алгоритмы, разработанные для сжатия именно таких потоков данных. Однако в разрабатываемом алгоритме, по крайнем мере, в начальной версии, для этого применяются стандартные универсальные алгоритмы. А именно, для сжатия дифференциальных кадров применяется Jpeg2000, для сжатия данных о движении — bzip2.

Важнейшим моментом является то, что данные о сегментации не требуется помещать в код видеофайла. Декодер может сам воспроизвести ту же сегментацию, что и кодер. Именно это обстоятельсто и позволяет получить преимущество перед другими методами кодирования видео.

Полученные результаты. Было проведено сравнение результатов сжатия видеоинформации предлагаемым методом с наиболее мощным алгоритмом из распространенных на сегодняшний день — H264 на исскусственных кадрах. Кадры представляют собой движущийся фон, по которому в свою очередь двигаются 4–5 спрайтов, перекрывая друг друга. В этой исскуственной ситуации, наиболее выигрышной для нашего алгоритма, при высоком качестве изображения мы получаем размер дифференциальных кадров вместе с данными о движениях в 2–3 раза (в зависимости от заданного уровня шума) меньше, чем у алгоритма H264 при том же уровне шума.

Разумеется, такое соотношение не может сохраняться для реальных видеоданных. Однако оно показывает предел возможностей рассматриваемого метода сжатия, его потенциал.

Работа выполнена при поддержке РФФИ, проект № 07-07-00178.

Литература

[1] *ITU-T and ISO/IEC JTC 1* Generic coding of moving pictures and associated audio information. Part 2: Video // ITU-T Recommendation H.262 – ISO/IEC 13818-2 (MPEG-2), Nov. 1994.

[2] *ITU-T* Video coding for low bit rate communication // ITU-T Recommendation H.263; version 1, Nov. 1995; version 2, Jan. 1998; version 3, Nov. 2000.

[3] *ITU-T Rec. H.264 / ISO/IEC 11496-10.* Advanced Video Coding // Final Committee Draft, Document JVT-E022, September 2002.

[4] Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification // (ITU-TRec.H.264.ISO/IEC14496-10AVC) Joint Video Team (JVT), Mar. 2003, Doc. JVT-G050.

Автоматическое распознавание контуров зданий на картографических изображениях

Чернов А. В., Титова О. А., Чупшев Н. В.

ache@smr.ru

Самара, Институт систем обработки изображений РАН,
Самарский государственный аэрокосмический университет

Рассматривается задача автоматической геопривязки (калибровки) изображений топографических планов по координатной сетке и задача автоматизированного распознавания на них контуров зданий. Использование разработанных алгоритмов и программных средств позволяет в 3–4 раза сократить время ручных операций при векторизации топопланов.

В рамках создания кадастра недвижимости, инфраструктуры пространственных данных, проводятся работы по переводу бумажных архи-

вов топографических планов городской застройки (планшетов) в цифровой вид. Технология перевода состоит из следующих этапов:

1) сканирование бумажных оригиналов и предварительная обработка;
2) геометрическая коррекция (калибровка) по координатной сетке;
3) ручная или полуавтоматическая векторизация.

Второй этап традиционно выполняется с помощью интерактивных программных средств, позволяющих оператору указать на изображении положение опорных точек (узлов координатной сетки, «крестов»), а затем выполнить геометрическую трансформацию. Затраты времени оператора в среднем составляют от 10 до 20 мин на топоплан (планшет).

Существующая практика использования на третьем этапе автоматических методов распознавания отдельных объектов показывает их малую эффективность из-за наличия в результате недоводов, петель, несовпадающих контуров, что требует ручной постобработки. Высокая трудоемкость ручного труда (от 3 до 5 рабочих дней на планшет) не дает возможности создания полноценных векторных цифровых планов крупных городов масштаба 1:500 за приемлемое время. Поэтому общая тенденция состоит в векторизации только объектов необходимых слоев (здания, уличная сеть).

Разработанные и представленные в докладе алгоритмы и программные средства полностью автоматизируют второй этап геопривязки и позволяют автоматизировать векторизацию зданий на третьем этапе при значительном снижении уровня ошибок и сохранении межобъектных связей в получившемся результате.

Алгоритмическое и программное обеспечение автоматической геопривязки

На топографических планах координатная сетка представлена в виде внутренней рамки размера 50×50 см и пересечений координатных линий (крестов) через каждые 10 см. Реализована технология автоматического поиска крестов и геометрической трансформации изображений, состоящая из следующих этапов:

— нахождение положения внешней рамки планшета и угла ее поворота;
— расчет начального положения крестов и углов внутренней рамки;
— уточнение положения крестов на основе специализированных корреляционных операторов, учитывающих их инвариантность относительно поворота на 90, 180, 270 градусов;
— фильтрация набора опорных точек для неверно найденных или неверно нанесенных на оригинал крестов;
— геометрическая трансформация (при необходимости, с обрезкой по внутренней рамке планшета) и запись выходного файла привязки.

Технология внедрена в Главархитектуре г. Самары и ряде других организаций. Соответствующее программное обеспечение работает в пакетном фоновом режиме, количество отказов в автоматической обработке (обусловленных, чаще всего, ошибками операторов при сканировании) составило менее 2 процентов.

Алгоритмическое и программное обеспечение автоматического распознавания контуров зданий

Объекты распознавания — здания представляют собой замкнутые контуры «почти прямоугольной формы», внутри которых расположена надпись (метка) типа «Ж», «ДЖ», «2КЖ», и пр. параллельно длинной стороне здания. Контуры объектов представлены в виде черных линий одинаковой толщины в несколько пикселей. К результату распознавания предъявляются требования замкнутости, сохранения прямых углов, отсутствия лишних точек «излома» контура, ложных контуров, совпадения границ соседних объектов, а также требования по максимальной погрешности отклонения относительно исходного растрового оригинала.

Реализованы алгоритмы и программные средства предварительного распознавания контуров зданий с последующей возможностью быстрого нахождения и исправления ошибок оператором. Процесс распознавания контуров зданий на топопланах состоит из следующих шагов.

— Построение «остова» исходного бинарного или полутонового изображения.
— Построение по остову линейно-узловой структуры в виде графа с вершинами — точками ветвления и ребрами — ломаными между ветвлениями.
— Предварительная обработка остова (замыкание некоторых линий) и его упрощение (удаление мелких полигонов, петель, спрямление и уменьшение количества точек ломаных, спрямление углов между ломаными), с сохранением допустимого положения узлов и ребер (без выхода за границы исходных растровых линий).
— Построение на основе остова замкнутых полигонов, предварительно классифицируемых как здания — объединение нескольких полигонов, имеющих близкие к прямоугольным углы и ярко выраженное преимущественное направление линий.
— Нахождение внутри потенциальных контуров зданий надписей вида *этажность* (число больше 2 или пропуск), *огнестойкость* («К», «П», «Д» или пропуск), *заселенность* («Н», «Ж»).
— Получение окончательного набора контуров зданий и создание векторных объектов в геоинформационных системах с записью семантических данных из распознанной надписи.

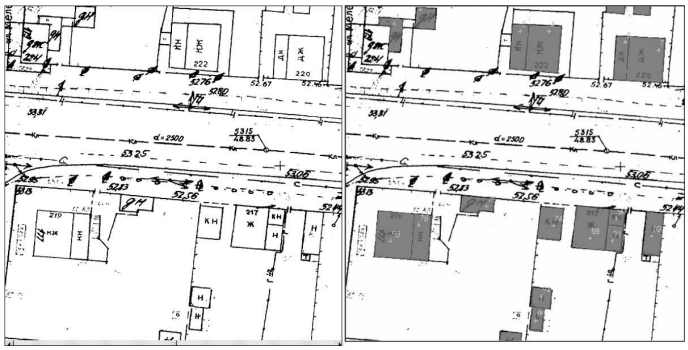


Рис. 1. Фрагмент исходных данных и результат распознавания.

В результате работы программы создается топологически корректная структура распознанных векторных объектов типа «здания» и набор контуров, не идентифицированых однозначно как здания, по которым решение принимает оператор. Технология в настоящее время внедряется в ряде организаций г. Самары. Предварительные исследования показывают снижение трудоемкости ручного ввода с 4–8 до 1–2 часов на планшет.

Работа выполнена при поддержке Министерства образования и науки РФ в рамках программы «Развитие научного потенциала высшей школы (2006-2008 годы)», правительства Самарской области и Американского фонда гражданских исследований и развития (CRDF Project RUX0-014-SA-06), при поддержке грантов РФФИ № 07-07-97603-р-офи и № 07-07-97610-р-офи.

Морфологический анализ изображений, искаженных аддитивным шумом

Чуличков А. И., Мурашев В. Э.

ach@cmp.phys.msu.ru

Москва, МГУ им. М. В. Ломоносова, физический факультет

В морфологическом анализе [1, 2, 3] введена операция сравнения по форме двух изображений (сигналов). Данная работа обобщает это понятие, вводя операцию сравнения по форме изображений, искаженных аддитивным шумом ограниченной яркости. Операция сравнения по форме является основой решения задач узнавания, классификации, оценки параметров формы сигнала.

Сравнение сигналов по форме

Изображением (сигналом) назовем числовую функцию f, заданную в конечном числе узлов сетки $\mathcal{X}$ плоскости $\mathbb{R}^2$ (числовой прямой $\mathbb{R}^1$). Множество узлов $\mathcal{X}$ будем называть полем зрения, а значение $f(x)$ функции f в точке $x \in \mathcal{X}$ — яркостью точки x поля зрения $\mathcal{X}$; множество всех изображений является евклидовым пространством $\mathbb{R}^n$, где n — число узлов сетки.

Пусть $\mathcal{F}$ — класс преобразований $F\colon \mathbb{R}^1 \to \mathbb{R}^1$. Сигнал $g \in \mathbb{R}^n$ по форме не сложнее, чем $f \in \mathbb{R}^n$, если $g(x) = F(f(x))$, $x \in \mathcal{X}$. Например, если $f(x_i) < f(x_j)$ для некоторых узлов $x_i, x_j \in \mathcal{X}$, и $\mathcal{F}$ — класс монотонно неубывающих преобразований, то для сигнала $g(x) = F(f(x))$ выполнено неравенство $g(x_i) \leqslant g(x_j)$. В результате, например, если $f(x_{i-1}) < f(x_i) > f(x_{i+1})$ и $g(x) = F(f(x))$ для монотонной функции F, то может оказаться, что $g(x_{i-1}) = g(x_i) = g(x_{i+1})$, т. е. «локальный пик» сигнала f может пропасть при монотонном преобразовании яркости, «форма» изображения g окажется более простой. Множество всех сигналов, форма которых не сложнее, чем f, в морфологическом анализе носит название формы сигнала f. В случае, когда $\mathcal{F}$ — класс монотонно монотонно неубывающих функций, форма

$$V_f = \bigl\{g \in \mathbb{R}^n\colon g(x) = F(f(x)),\ x \in \mathcal{X},\ f \in \mathcal{F}\bigr\}$$

сигнала $f \in \mathbb{R}^n$ является выпуклым замкнутым конусом в $\mathbb{R}^n$. При морфологическом анализе множеству V_f ставится в соответствие конструктивно определенный оператор проецирования на V_f, называемый формой изображения (сигнала) f.

Пусть теперь сами сигналы f и g ненаблюдаемы, так как процесс их измерения сопровождается аддитивной погрешностью, в результате чего результат измерения имеет вид

$$\xi(x) = f(x) + \nu(x), \quad \eta(x) = g(x) + \mu(x), \tag{1}$$

где погрешности измерения $\nu, \mu \in \mathbb{R}^n$ удовлетворяют условиям $|\nu(x)| \leqslant \varepsilon_1(x)$, $|\mu(x)| \leqslant \varepsilon_2(x)$, $x \in \mathcal{X}$. Ответ на вопрос, будет ли $g \in \mathbb{R}^n$ не сложнее по форме, чем f, должен быть получен на основании наблюдения зашумленных сигналов (1). Ответ на него положителен, если найдутся такие $|\nu(x)| \leqslant \varepsilon_1(x)$, $|\mu(x)| \leqslant \varepsilon_2(x)$, и монотонно неубывающая функция $F \in \mathcal{F}$, что

$$\eta(x) - \mu(x) = F(\xi(x) - \nu(x)) \ \text{ для всех } \ x \in \mathcal{X}.$$

Рассмотрим на плоскости $\mathbb{R}^1 \times \mathbb{R}^1$ множество прямоугольников

$$\bigl\{[f(x) - \varepsilon_1(x), f(x) + \varepsilon_1(x)] \times [g(x) - \varepsilon_2(x), g(x) + \varepsilon_2(x)],\ x \in \mathcal{X}\bigr\}.$$

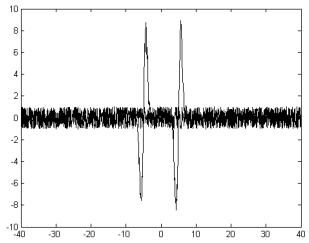


Рис. 1. График результата измерения акустического сигнала с помощью двух разнесенных микрофонов.

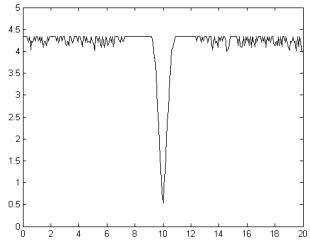


Рис. 2. Зависимость ограничения ε на величину погрешности измерения от значения сдвига сигналов, изображенных на Рис. 1, одного относительно другого.

Теперь $g \in \mathbb{R}^n$ не сложнее по форме, чем $f \in \mathbb{R}^n$ тогда и только тогда, когда в каждом прямоугольнике можно выбрать ровно одну точку так, чтобы через все n точек можно провести ломаную из отрезков прямых с неортицательным наклоном.

Проверка этого условия может быть выполнена методом динамического программирования.

Оценка относительного смещения сигналов

Мощность акустического сигнала измеряется микрофонами, разнесенными в пространстве. Графики результатов измерения изображены на Рис. 1 в условных единицах. Требуется оценить временную задержку сигналов, если известно, что микрофоны могут обладать нелинейным коэффициентом усиления, а измерения сопровождаются аддитивным шумом.

Рассмотрим метод, в основе которого лежит сравнение по форме двух сигналов, искаженных монотонным преобразованием. Будем сдвигать один сигнал относительно другого, и в предположении, что $\varepsilon_1(x) =$

$= \varepsilon_2(x) = \varepsilon$ для всех $x \in \mathcal{X}$ найдем минимальное значение ε, при котором зашумленные сигналы $f, g \in \mathbb{R}^n$ сравнимы по форме.

Если считать, что большие значения погрешности менее возможны, чем малые, то в качестве оценки параметра сдвига следует выбрать то значение сдвига, которое соответствует наименьшему значению ε, при котором сигналы сравнимы по форме [4]. В данном случае это значение оказалось равным 10 усл. ед.

Работа выполнена при поддержке РФФИ, проект № 05-01-00615.

Литература

[1] *Пытьев Ю. П.* Задачи морфологического анализа изображений // Математические методы исследования природных ресурсов Земли из Космоса. — М.: Наука, 1984.

[2] *Пытьев Ю. П.* Морфологический анализ изображений // Докл. АН СССР. — 1983. — Т. 269, № 5. — С. 1061–1064.

[3] *Pyt'ev Yu. P.* Morphological Image Analysis // Pattern Recognition and Image Analysis. — 1993. — V. 3, No. 1. — Pp. 19–28.

[4] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применения. — Москва: Физматлит, 2007.

Детектор границы области на цветных изображениях

Чуличков А. И., Илюшин В. Л.

ach@cmp.phys.msu.ru

Москва, МГУ им. М. В. Ломоносова, физический факультет

В работе построен морфологический метод анализа цветных изображений для определения границы раздела областей одинакового цвета.

Форма цветного изображения

Фиксируем сцену, состоящую из некоторого заданного набора объектов, и рассмотрим множество ее цветных изображений [1], полученных при всевозможных условиях регистрации, заданное в виде

$$V_f = \left\{ \vec{f}\,(x) = \sum_{i=1}^{N} \chi_i(x) \left(\vec{\varphi}_{0,i} + \sum_{j=1}^{k_i} \vec{\varphi}_{j,i} g_{j,i}(x) \right),\ g_{j,i} \in \mathbb{R}^1,\ \vec{f}\,(x) \in K \right\},$$

где $\chi_1, \ldots, \chi_N$ — индикаторы разбиения $A_1, \ldots, A_N$ поля зрения $\mathcal{X}$:

$$A_i \bigcap A_j = \varnothing, \quad i \neq j, \quad i, j = 1, \ldots, N, \quad \bigcup_{i=1}^{N} = \mathcal{X};$$

K — выпуклое множество (конус) векторов, определяемый условиями физической реализуемости (например, неотрицательностью координат).

На каждом подмножестве $A_i \subset \mathcal{X}$ цвет изображения заданной сцены является элементом, принадлежащим гиперплоскости

$$G_i = \left\{ \vec{\varphi}_{0,i} + \sum_{j=1}^{k_i} \vec{\varphi}_{j,i} g_{j,i}(x),\ g_{j,i} \in \mathbb{R}^1 \right\},$$

задаваемой набором векторов цвета $\vec{\varphi}_{0,i}, \ldots, \vec{\varphi}_{k_i,i}$, одним и тем же для фиксированной гиперплоскости G_i, $i = 1, \ldots, N$, при этом векторы $\vec{\varphi}_{1,i}, \ldots, \vec{\varphi}_{k_i,i}$ ортогональны. Гиперплоскости $G_1, \ldots, G_N$ отличаются одна от другой количеством и значениями векторов цвета. Заметим, что для фиксированного изображения представление его в виде

$$\vec{f}(x) = \sum_{i=1}^{N} \chi_i(x) \left(\vec{\varphi}_{0,i} + \sum_{j=1}^{k_i} \vec{\varphi}_{j,i} g_{j,i}(x) \right),$$

вообще говоря, неоднозначно, так как точка x поля зрения, в которой значение $\vec{f}(x)$ принадлежит пересечению гиперплоскостей, может быть отнесена к тому или иному множеству разбиения произвольно. Однако для задач, решаемых в данной работе, эта неоднозначность несущественна.

Будем считать, что при изменении условий регистрации может меняться как набор векторов цвета $\vec{\varphi}_{0,i}, \ldots, \vec{\varphi}_{k_i,i}$, так и коэффициенты $g_{j,i}(x)$, $x \in A_i$, $j = 1, \ldots, k_i$, $i = 1, \ldots, N$. Остаются неизменными лишь числа $k_1, \ldots, k_N$, определяющие размерность гиперплоскости, $k_i \in \{0, 1, \ldots, l-1\}$, $i = 1, \ldots, N$.

Такой моделью описываются изображения сцен, состоящих из наборов окрашенных предметов, освещаемых однородным потоком света. Разбиение $A_1, \ldots, A_N$ в этом случае определяется оптически однородными поверхностями объектов сцены.

Принадлежность регистрируемого изображения $\vec{s}$ форме V_f дает основание считать, что $\vec{s}$ является изображением той же сцены, которая порождает форму V_f.

Отметим, что множество V_f является выпуклым, а потому форма V_f может быть задана оператором проецирования P_f на V_f в $\mathcal{L}_2^l(\mathcal{X})$, который является решением задачи [2, 3, 4]

$$\|P_f \vec{s} - \vec{s}\|^2 = \inf_{\vec{\tilde{f}} \in V_f} \|\vec{\tilde{f}} - \vec{s}\|^2.$$

На практике регистрация изображения $\vec{s}$ сопровождается аддитивной погрешностью, что приводит к тому, что результат регистрации

$$\vec{\xi} = \vec{s} + \vec{\nu}$$

не обязательно принадлежит V_f, даже если $\vec{s} \in V_f$. В таких условиях необходимо иметь характеристику близости $\vec{\xi}$ к V_f. Обобщая критерий близости по форме, используемый для полутоновых изображений, в качестве близости изображений по форме используем функционал

$$\frac{\|P_f\vec{\xi} - \vec{\xi}\|^2}{\|P_f\vec{\xi} - P_0\vec{\xi}\|^2},$$

где P_0 — проектор на множество изображений, имеющих в каждой точке поля зрения один и тот же цвет.

Морфологический детектор границы

Рассмотрим разбиение поля зрения на две области с прямолинейной границей; угол наклона границы является параметром формы. Рассмотрим цветное изображение $\vec{\xi}$, заданное на поле зрения $\mathcal{X}$, и P_ϑ — оператор проецирования на форму изображения, имеющего различные цвета на двух областях поля зрения с границей раздела в виде прямой, наклоненной под углом ϑ, $\vartheta \in [0, \pi)$. Форму такого изображения назовем формой изображения края. Значение функционала

$$\Theta(\vec{\xi}) = \sup_{\vartheta\in[0,\pi)} \frac{\|P_\vartheta\vec{\xi} - \vec{\xi}\|^2}{\|P_\vartheta\vec{\xi} - P_0\vec{\xi}\|^2}, \tag{1}$$

определяет близость формы предъявленного изображения к форме изображения края.

Пусть на поле зрения $\mathcal{X}$ задано некоторое цветное изображение $\vec{\Xi}$. Выделяя на поле зрения $\mathcal{X}$ подмножество $X \in \mathcal{X}$ с центром в точке $x \in \mathcal{X}$, рассматривая фрагмент $\vec{\xi}_x$ изображения $\vec{\Xi}$ на подмножестве X и вычисляя в каждой точке значение функционала, определенного в (1), получим функцию $\varphi(x) = \Theta(\vec{\xi}_x)$, $x \in \mathcal{X}$, значение которой в точке $x \in \mathcal{X}$ определяет близость формы изображения $\vec{\Xi}$ в окрестности точки x поля зрения $\mathcal{X}$ к изображению «края».

Работа выполнена при поддержке РФФИ, проект № 05-01-00615.

Литература

[1] *Pyt'ev Yu. P.* The Morphology of Color (Multispectral) Images // Pattern Recognition and Image Analysis. — 1997. — V. 7, No. 4. — Pp. 467–473.

[2] *Пытьев Ю. П.* Морфологический анализ изображений // Докл. АН СССР. — 1983. — Т. 269, № 5. — С. 1061–1064.

[3] *Pyt'ev Yu. P.* Morphological Image Analysis // Pattern Recognition and Image Analysis. — 1993. — V. 3. No. 1. — Pp. 19–28.

[4] *Pyt'ev Yu. P.* Methods for Morphological Analysis of Color Images // Pattern Recognition and Image Analysis. — 1997. V. 8, No. 4. — Pp. 517–531.

Продолжение меры возможности, определяющее нечеткую форму изображения

Чуличков А. И.

ach@cmp.phys.msu.ru

Москва, МГУ им. М. В. Ломоносова, физический факультет

Морфологические методы анализа изображений [1, 2, 3] предназначены для решения задач узнавания, классификации объектов и сцен, выделения отличий в сценах по их изображениям, оценивания параметров в терминах формы изображения. Под формой понимается множество изображений, полученных при всевозможных условиях наблюдения сцены. Как правило, это множество задается как выпуклое множество евклидова пространства всех изображений, с ним взаимно однозначно связан оператор проецирования. В терминах формы решаются все перечисленные выше задачи.

В докладе рассматривается обобщение морфологических методов, в котором форма изображения заданной сцены строится путем задания распределения $\nu\colon \mathcal{R} \to [0,1]$ возможностей [4] на множестве $\mathcal{R}$ всех изображений. Значение $\nu(f)$ определяет возможность того, что изображение $f \in \mathcal{R}$ порождено заданной сценой. Задание меры возможностей позволяет применять хорошо разработанные методы решения задач нечеткого оценивания и принятия решений [2, 3, 4, 5].

Центральным моментом теоретико-возможностного аналога морфологических методов анализа изображений является задание распределения возможностей $\nu\colon \mathcal{R} \to [0,1]$. В работе предлагается следующий способ: указывается некоторое начальное распределение, в котором мера возможностей определена на фиксированном наборе «четких» множеств $V_1, \ldots, V_n, \ldots$ Каждое из множеств содержит изображения, порожденные данной сценой при некотором классе условий регистрации. Например, если рассматриваются изображения рукописных символов, то множество V_k содержит изображения символа фиксированного начертания, полученные при различных условия освещения, что приводит к различным яркостям подмножеств поля зрения, изображающих фон или знак; множества V_k и V_m при $k \neq m$ отличаются разными вариантами на-

чертания символов. Такой набор можно получить методами «четкого» морфологического анализа из некоторого числа изображений, порожденных заданной сценой при различных условиях наблюдения (образцов). Начальное распределение возможностей $P(V_{f,k})$, $k = 1, 2, \ldots$, на этом наборе множеств строится следующим образом: считается, что для любого изображения $f \in \mathcal{R}$ возможность того, что оно порождено заданной сценой, равна нулю, если $f \notin V_k$ для всех $k = 1, 2, \ldots$, и равна $\sup\{P(V_k) \mid k\colon f \in V_k\}$ в противном случае. Далее используется специфическое продолжение этой меры возможностей, учитывающее сходство по форме предъявленного изображения и изображений из заданного набора форм $V_{f,1}, \ldots, V_{f,n}, \ldots$

Пусть, например, изображение есть числовая функция, заданная на ограниченном подмножестве (поле зрения) $\mathcal{X} \subset \mathcal{R}_2$ с заданной мерой μ, квадрат ее интегрируем на $\mathcal{X}$, тем самым $\mathcal{R} = \mathcal{L}_2^\mu(\mathcal{X})$. Значения формы $V_1, \ldots, V_n, \ldots$ заданы в виде выпуклых замкнутых конусов евклидова пространства $\mathcal{L}_2^\mu(\mathcal{X})$:

$$V_k = \{g \in \mathcal{L}_2^\mu(\mathcal{X})\colon g = F * f_k\},$$

где $g = F * f_k$ означает, что для значений (яркости) $g(x)$ изображения g μ-почти всюду на $\mathcal{X}$ выполнено равенство $g(x) = F(f_k(x))$, а F — произвольная монотонно неубывающая функция, такая, что $F * f_k \in \mathcal{L}_2^\mu(\mathcal{X})$, $k = 1, 2, \ldots$; здесь f_k — изображение сцены, полученное при k-м условии регистрации. Конусам $V_1, \ldots, V_n, \ldots$ сопоставим операторы проецирования $\Pi_1, \ldots, \Pi_n, \ldots$ в $\mathcal{L}_2^\mu(\mathcal{X})$ на $V_1, \ldots, V_n, \ldots$ соответственно. В морфологическом анализе в качестве меры близости некоторого изображения $\xi \in \mathcal{L}_2^\mu(\mathcal{X})$ по форме к f_k используется функционал

$$t_k(\xi) = \frac{\|(I - \Pi_k)\xi\|^2}{\|(E - \Pi_k)\xi\|^2}, \tag{1}$$

где I — единичный оператор, а E — проектор в $\mathcal{L}_2^\mu(\mathcal{X})$ на множество изображений $\{g = \text{const}\}$, равных константе μ-почти всюду на $\mathcal{X}$. Функционал (1) инвариантен относительно любых монотонных преобразований яркости изображения. Для геометрической интерпретации этой близости рассмотрим единичную сферу S в ортогональном дополнении в $\mathcal{L}_2^\mu(\mathcal{X})$ к прямой $\{g = \text{const}\}$, и обозначим ξ_1 пересечение этой сферы с лучом $\{k\xi,\, k > 0\}$ и $\tilde{V}_{f,k}$ — пересечение с конусом $V_{f,k}$. Величина $t_{f,k}(\xi)$ равна тангенсу углового расстояния точки ξ_1 от множества $\tilde{V}_{f,k}$ на единичной сфере S.

Будем считать, что чем дальше на сфере S точка ξ_1 от множества $\tilde{V}_{f,k}$, тем меньше возможность того, что изображение ξ порождено сценой f

при k-м условии регистрации. Продолжение P начальной меры можно получить с помощью распределения возможностей

$$\nu_f(\xi) = \sup_k \min\bigl\{\nu_0(t_{f,k}), P(V_{f,k})\bigr\},$$

где ν_0 — монотонно невозрастающая функция, определенная на неотрицательной полупрямой, $\nu_0(0) = 1$ и $\lim\limits_{z\to\infty} \nu_0(z) = 0$.

Такое продолжение позволяет учесть как различия в возможностях реализаций тех или иных условий наблюдения сцены, так и возможность иных искажений регистрируемого изображения. Формально задача оценивания и принятия решений сводится к анализу событий, изображающихся точками на сфере S с заданной на ней мерой возможности $\tilde{\nu}_f(\tilde{\xi}) = \nu_f(\xi)$.

Работа выполнена при поддержке РФФИ, проект № 05-01-00615.

Литература

[1] *Пытьев Ю. П.* Задачи морфологического анализа изображений // В сб.: Математические методы исследования природных ресурсов Земли из Космоса, М.: Наука, 1984г.

[2] *Чуличков А. И., Морозова И. В.* Классификация размытых изображений и оценка параметров системы регистрации методами морфологического анализа // Интеллектуальные системы. — 2005. — Т. 9, Вып. 1-4. — С. 321–344.

[3] *Чуличков А. И.* Множества, оценивающие параметр формы сигнала // 9 межд. конф. «Интеллектуальные системы и компьютерные науки», Москва: мех.-мат. факультет МГУ, 2006. — Т. 1. Часть 2. — С. 310–313.

[4] *Пытьев Ю. П., Зубюк А. В.* Случайная и нечеткая морфология (эмпирическое восстановление модели, идентификация) // 9 межд. конф. «Интеллектуальные системы и компьютерные науки», Москва: мех.-мат. факультет МГУ, 2006. — Т. 1. Часть 2. — С. 222–225.

[5] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применения. — Москва: Физматлит, 2007.

О выборе весов для подмножеств признаков при распознавании речи

Чучупал В. Я.

chuchu@ccas.ru

Москва, Вычислительный Центр РАН

При распознавании речи с использованием квазифонемных марковских моделей распознанная последовательность моделей $q_1^T = q_1, \ldots, q_T$ для данной последовательности наблюдений $x_1^T = x_1, \ldots, x_T$ обычно

определяется путем максимизации логарифма правдоподобия $\log P(x_1^T|q_1^T)$ на множестве всех допустимых последовательностей моделей [1]:

$$\log P(x_1^T|\Theta) = \arg\max_{q_1^T} \log P(x_1^T|q_1^T).$$

При этом

$$\log P(x_1^T|q_1^T) = \sum_{i=1}^{T} \log P(x_i|q_i). \tag{1}$$

Наблюдение x_i обычно содержит несколько подмножеств параметров, $x_i = (x_i^1, x_i^2, \ldots, x_i^K)$. Здесь K — число подмножеств, x_i^j — набор параметров j подмножества в момент i. Величина локального, в момент i, правдоподобия вычисляется как:

$$\log P(x_i|\Theta) = \sum_{j=1}^{K} \lambda_j \log P(x_i^j|\Theta) = \sum_{j=1}^{K} \mu_j(x_i^j, \Theta), \tag{2}$$

где через Θ обозначены параметры модели состояния q_i, а λ_j — весовой коэффициент j подмножества параметров.

Величина коэффициентов λ_j устанавливается из эвристических соображений и, как правило, эти коэффициенты выбираются равными, не зависящими от модели Θ или параметров [2]. Возникает вопрос: можно ли систематически и более общим образом выбрать λ или μ, например, как функции от моделей и текущих параметров?

Оценка вероятностей на основе максимума энтропии

Пусть x — случайная величина, и нас интересует распределение, либо плотность ее вероятностей. В соответствии с принципом максимальной энтропии [3], если для функций $f_i(x)$, $i = 1, \ldots, K$ известны математические ожидания, то существуют такие константы $\lambda_0, \ldots, \lambda_K$ и такое распределение вероятностей $P(x)$, что оно обладает максимальной энтропией на множестве всех вероятностных распределений, которые удовлетворяют заданным ограничениям и может быть выражено как:

$$P(x) = \frac{\exp\bigl(-\lambda_0 - \lambda_1 f_1(x) - \cdots - \lambda_K f_K(x)\bigr)}{\sum\limits_x \exp\bigl(-\lambda_1 f_1(x) - \cdots - \lambda_K f_K(x)\bigr)}. \tag{3}$$

Распределение $P(x)$ оптимально в том смысле, что кроме заданных ограничений никаких других зависимостей не предполагает.

Если обозначить

$$\mu_0(x) = \frac{\exp(-\lambda_0)}{\sum\limits_x \exp\bigl(-\lambda_1 f_1(x) - \cdots - \lambda_K f_K(x)\bigr)},$$
$$\mu_1(x) = e^{-\lambda_1 f_1(x)}, \ldots, \mu_K(x) = e^{-\lambda_K f_K(x)},$$

то уравнение (3) можно переписать в виде

$$\log P(x) = \sum_{j=0}^{K} \log \mu_j(x). \tag{4}$$

Алгоритм оценки весов подмножеств

Рассмотрим случай дискретных марковских моделей: когда параметры сигнала кодируются элементами кодовой книги: $x_t^i \to c_i^t$, или $x_t^i \in c_i^t$. Здесь c_i — элемент кодовой книги C^i для описания i-го подпространства параметров.

Оценка условной вероятности $P(\Theta|c_{i0})$ модели Θ при параметрах, принадлежащих коду c_{i0}, есть среднее значение:

$$P(\Theta|c_{i0}) = K_{\Theta,c_{i0}} = \mathop{\mathsf{E}}_{\substack{(c_1,\ldots,c_K)\\ c_i=c_{i0}}} \bigl[P(\Theta|c_1, c_2, \ldots, c_K)\bigr]. \tag{5}$$

Равенства (5), при всех Θ, c_i, можно рассматривать как набор ограничений.

Введем (аналогично [4]) бинарные селекторные функции

$$f_{\Theta,c_i}(x) = \begin{cases} 1, & \text{если } x \in \Theta \text{ и } x \in c_i; \\ 0, & \text{в противном случае.} \end{cases} \tag{6}$$

Тогда ограничения (5) можно записать в виде:

$$\sum_{c_1,\ldots,c_K} P(c_1, \ldots, c_K) \sum_{\Theta} P(\Theta|c_1, \ldots, c_K) f_{\Theta,c_i}(x) = K_{\Theta,c_i} P(H_{c_i}), \tag{7}$$

где $P(H_{c_i})$ обозначает вероятность того, что наблюдаемое значение кода для i-го подпространства параметров есть c_i.

Алгоритм 1 вычисления $\mu(\Theta, x^j)$ по обучающей выборке основан на алгоритме GIS (Generalized Iterative Scaling) [5].

Алгоритм 1. Оценка оптимальных весов подмножеств для дискретного случая.

Вход: параметры модели Θ, $j = 0$;

1: Оценим по частотам обучающей выборки ограничения (7) и начальные значения $\log \mu$ для всех i, c_{i0}:
$\log K^{(0)}_{\Theta,c_{i0}} := \log P(\Theta|c_i)$;
$\log \mu^{(0)}(\Theta, c_{i0}) := \log K_{\Theta,c_{i0}} + \log P(H_{c_{i0}})$;

2: **повторять**

3: Оценим вероятности (4):
$$\log P(\Theta|c_1,\ldots,c_K) := \sum_{i=1}^{K} \log \mu^{(j)}(\Theta, c_j) - \log\Big(\sum_{\Theta} \mu^{(j)}(\Theta, c_j)\Big);$$

4: Вычислим новые значения ограничений:
$$\log K_{\Theta,c_{i0}} := \log\Big(\sum_{\substack{(c_1,\ldots,c_K)\\ c_i=c_{i0}}} P(c_1,\ldots,c_K)P(\Theta|c_1,\ldots,c_K)\Big);$$

5: Переоценим $\log \mu$:
$\log \mu^{(j+1)}(\Theta, c_i) := \log \mu^{(j)}(\Theta, c_i) + \log K^{(0)}_{\Theta,c_i} - K^{(j)}_{\Theta,c_i}$;

6: $j := j + 1$;

7: **пока** $\log \mu^{(j+1)}(\Theta, c_i) - \log \mu^{(j)}(\Theta, c_i) \geqslant \varepsilon$;

Чтобы полученные оценки можно было использовать обычным для распознавания речи образом, перепишем (1), используя формулу Байеса:

$$\sum_{i=1}^{T} \log P(c_i^1,\ldots,c_i^K|q_i) =$$
$$= \sum_{i=1}^{T} \log P(q_i|c_i^1,\ldots,c_i^K) + \sum_{i=1}^{T} \log P(c_i^1,\ldots,c_i^K) - \sum_{i=1}^{T} \log P(q_i). \quad (8)$$

Априорные вероятности $P(q_i)$ оцениваются по обучающей выборке, далее, поскольку последовательность наблюдений известна, то сумма вероятностей $\sum_{i=1}^{T} P(c_i^1,\ldots,c_i^K)$ постоянна, и при поиске максимума её можно игнорировать.

Работа выполнена при поддержке РФФИ, проект № 07-01-00657а.

Литература

[1] *Jelinek F.* Statistical methods for speech recognition. MIT Press, 1998.

[2] *Young, S.* The HTK BOOK. Ver. 2.1. Cambridge University, 1997.

[3] *Janes, E.T.* Probability theory: the logic of science. Cambridge University Press, 2006.

[4] *Rosenfeld R.* A maximum entropy approach to adaptive statistical language modeling // Computer Language and Speech. — Vol. 10, No. 3. — Pp. 187–228.

[5] *Darroch J. N., Ratcliff D.* Generalized iterative scaling for log-linear models // The annals of Mathematical Statistics, 1972. — No. 43. — Pp. 1470–1480.

Обнаружение незнакомых слов при распознавании речи

Чучупал В. Я., Маковкин К. А., Чичагов А. В.

chuchu@ccas.ru

Москва, Вычислительный центр РАН

Речевой сигнал, подлежащий распознаванию, как правило, содержит шумы, речь посторонних лиц, нарушения речевого потока или слова, которые не входят в словарь системы распознавания речи.

Целью настоящего исследования являлась разработка и исследование эффективности нового алгоритма обнаружения незнакомых слов на основе оценок правдоподобия для наблюдаемого речевого сигнала при заданном множестве акустико-фонетических моделей.

Акустические счета

Основные методы выявления незнакомых слов основаны на использовании величин локальных мер сходства, оценок правдоподобия или акустических счетов. Наиболее часто используются: средний счет

$$S(w) = \frac{1}{N_w} \sum_{j=1}^{N_w} s_j, \tag{1}$$

где N_w — число наблюдений, s_j — счет наблюдения j, и центрированный средний счет:

$$\tilde{S}(w) = \frac{1}{N_w} \sum_{j=1}^{N_w} (s_j - \tilde{s}), \tag{2}$$

где $\tilde{s}$ — среднее счета для длительного промежутка времени.

Обе оценки широко используются, однако несколько лучшие результаты были получены при использовании дважды нормированного счета:

$$S_1(w) = \frac{1}{N_w} \sum_{s=1}^{N_w} \frac{1}{N_s} \sum_{j=1}^{N_s} s_j, \tag{3}$$

где N_w — число состояний модели слова w, N_s — длина состояния s.

По аналогии с (2) определим дважды нормированный центрированный счет:

$$\tilde{S}_1(w) = \frac{1}{N_w} \sum_{j=1}^{N_w} \frac{1}{N_s} (s_j - \tilde{s}). \tag{4}$$

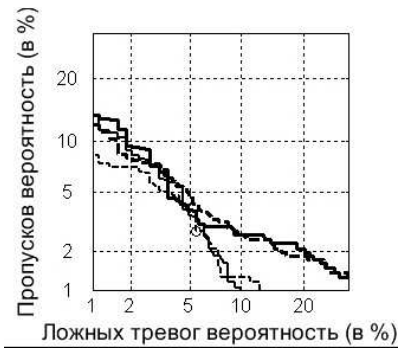


Рис. 1. DET-характеристики акустических счетов (1)—(4).

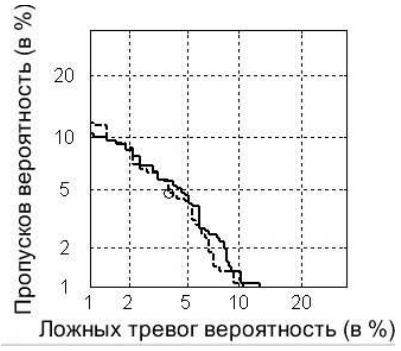


Рис. 2. DET-кривая счетов (5)–(7).

На Рис. 1 приведены DET-кривые (Detection Error Tradeoff или соотношение ошибок обнаружения [1]) для счетов (1)–(4), вычисленные на речевом корпусе данных FaVoR [2]. Жирная сплошная линия соответствует счету (1), жирная штриховая — счету (2), сплошная — (3) и штриховая — (4). Очевидно, что счет (4) — самый эффективный.

Оценка счетов по параметрам моделей звуков

Существенным недостатком рассмотренных акустических счетов является то, что оценка порогового значения требует наличия настроечной выборки, т. е. какого-то количества произнесений слова в заданной акустико-фоновой обстановке.

Рассмотрим метод формирования оценки счета на основе параметров акустических моделей звуков.

Для счета (4) определим среднее, отклонение и порог принятия решения:

$$\tilde{m}(w) = \frac{1}{N_w}\sum_{s=1}^{N_w}\frac{1}{N_s}\sum_{j=1}^{N_s} m_j; \tag{5}$$

$$\sigma(w) = \sqrt{\frac{1}{N_w}\sum_{s=1}^{N_w}\frac{1}{N_s}\sum_{j=1}^{N_s} \sigma_j^2}; \tag{6}$$

$$th(w) = s(w) - m(w) - \sigma(w). \tag{7}$$

Оцениваем текущее значение счета как:

$$\tilde{m}(w, t+1) = \tilde{m}(w, t) + \alpha\big(\tilde{m}(w, t) - S(w, t)\big), \tag{8}$$

где $S(w, t)$ — наблюдаемое значение счета, α — параметр.

На Рис. 2 приведены DET-кривые для оценок (5)–(7). Сплошная линия соответствует счету (5), штриховая счету (7).

Заключение

Предложен метод оценки априорных дисперсий счетов, выбора априорного порога, а также процедура адаптации средних значений и порогов в соответствии с наблюдаемым речевым сигналом. Показано, что эффективность предложенных априорных оценок счетов и метода адаптации порога практически соответствует использованию апостериорных оценок пороговых счетов на настроечной выборке.

Работа выполнена при поддержке РФФИ, проект № 07-01-00657а.

Литература

[1] Программа DETware. Национальный институт стандартов США, NIST. — http://www.nist.gov.

[2] *Десятчиков А. А и др.* Комплекс алгоритмов для устойчивого распознавания человека // Известия РАН. Теория и системы управления. — 2006. — Т. 45, № 6. — С. 73-85.

Алгоритмы эмпирического восстановления случайной и нечеткой формы изображения

Шишаков В. В., Пытьев Ю. П.

shift@cmpd2.phys.msu.ru, pytyev@phys.msu.ru

Москва, МГУ им. М. В. Ломоносова, физический факультет

Пусть регистрируемое изображение объекта является элементом евклидова пространства $\mathcal{R}_N$. Формой изображения объекта будет множество $V \in \mathcal{R}_N$ его изображений, получаемых при всех возможных изменениях условий регистрации [1]. Например, в задаче анализа изображений рукописных цифр, формой изображения цифры «1» будет множество любых изображений единицы, в том числе и различной яркости, контраста и с различными искажениями формы единицы.

Методы морфологического анализа изображений разрабатывались для анализа и интерпретации изображений объектов, полученных при различных и неизвестных условиях регистрации (таких как освещение, размытие), при которых «геометрическая» форма объекта остается неизменной. Очевидно, что решая задачу узнавания рукописных цифр, мы ожидаем увидеть их на предъявленном изображении более-менее «прямыми» и «правильными», Однако, на реальных изображениях рукописных цифр искажения по форме будут всегда, например, из-за разницы почерков разных людей. Подобные искажения формы, будут носить случайный характер, и разумно будет предположить, что более вероят-

но можно будет увидеть менее искаженное изображение цифры. Чтобы учесть эти соображения, воспользуемся понятием случайной формы [3].

Пусть дано разбиение $\bigcup\limits_{\omega\in\Omega}\omega=\mathcal{R}_N$ пространства $\mathcal{R}_N$ на непересекающиеся множества $\omega\subset\mathcal{R}_N$, совокупность которых обозначим Ω, минимальная σ-алгебра $\mathcal{A}$ подмножеств множества Ω, содержащая все множества $\{\omega\}$, $\omega\in\Omega$, и вероятность $\Pr\colon\mathcal{A}\to[0,1]$. Вероятностное пространство $(\Omega,\mathcal{A},\Pr)$, в котором элементарными событиями являются множества $\omega\in\Omega$ в пространстве $\mathcal{R}_N$ является математической моделью случайной формы.

Каждому событию $A\in\mathcal{A}$ соответствует форма $V=\bigcup\limits_{\omega\in A}\omega\subset\mathcal{R}_N$, вероятность которой есть $\Pr(A)$. Множество всех форм, соответствующих σ-алгебре $\mathcal{A}$, обозначим $\mathbb{V}_{\mathcal{A}}$, $\quad\mathbb{V}_{\mathcal{A}}=\bigl\{V\colon V=\bigcup_{\omega\in A}\omega,\,A\in\mathcal{A}\bigr\}$.

Например, если множество Ω состоит из всех лучей в $\mathcal{R}_N$, исходящих из начала координат. Каждый луч — форма любого изображения, как элемента $\mathcal{R}_N$, принадлежащего лучу. S_N — единичная сфера в $\mathcal{R}_N$ с центром в начале координат, $\mathcal{A}'$ — борелевская σ-алгебра подмножеств сферы S_N, а $\mathcal{A}$ — взаимно однозначно связанная с ней σ-алгебра подмножеств множества Ω. Задав на $\mathcal{A}'$ вероятность Pr, зададим соответствующую вероятность и на $\mathcal{A}$. При этом соответствующее $\mathcal{A}$ множество форм $\mathbb{V}_{\mathcal{A}}$ содержит все линейные подпространства в $\mathcal{R}_N$, все конусы с вершиной в начале координат, и т. п.

Таким образом, задачу идентификации случайной формы изображения можно сформулировать следующим образом. Пусть заданы случайные формы: $F_i=(\Omega,\mathcal{A},\Pr_i)$, где вероятности $\Pr_i$ заданы плотностями $\mathrm{pr}^{(i)}(\cdot)$, где $i=1,2,\ldots$ соответственно, и предъявляемое для идентификации изображение ξ формируется по схеме

$$\xi=f+\nu,\tag{1}$$

где $f\in\mathcal{R}_N$ — произвольный элемент F_i, а $\nu\in R_N$ — случайный элемент с плотностью распределения $\mathrm{pr}_\nu(\cdot)$, независимый от f. Требуется по предъявленному изображению ξ определить, какую случайную форму F_1, $F_2,\ldots$ имеет элемент f, т. е. определить ν.

Заметим, что если f имеет случайную форму F_t, $t=1,2,\ldots$, то ξ является случайным элементом пространства $\mathcal{R}_N$ с одной из следующих плотностей распределения: $\mathrm{pr}_\xi^{(i,\,\varphi)}(x)=\int_\Omega\mathrm{pr}^{(i)}(\omega)\,\mathrm{pr}_\nu(x-\varphi_\omega)\,\mathrm{d}\omega$, где символом φ обозначена произвольная совокупность элементов $\mathcal{R}_N$ вида $\varphi=\{\varphi_\omega\in\omega,\,\omega\in\Omega\}$.

Введём следующие множества распределений: $\mathbb{P}\mathrm{r}_i\stackrel{\mathrm{def}}{=}\bigl\{\mathrm{pr}_\xi^{(i,\,\varphi)}\bigr\}$, $\mathcal{H}_0\stackrel{\mathrm{def}}{=}\stackrel{\mathrm{def}}{=}\Cap\mathbb{P}\mathrm{r}_i$, $\mathcal{H}_i\stackrel{\mathrm{def}}{=}\mathbb{P}\mathrm{r}_i\setminus\mathcal{H}_0$, $i=1,2,\ldots$ Для идентификации предъяв-

ленного изображения можно воспользоваться минимаксным критерием $\bigl(\widetilde{\pi}_0(x), \widetilde{\pi}_1(x), \ldots\bigr),\ x \in \mathcal{R}_N$ проверки гипотез $\mathcal{H}_0, \mathcal{H}_1, \ldots,$ который минимизирует вероятность ошибки и находится как решение задачи на минимакс [2]

$$\begin{cases} \max\limits_{i=0,1,\ldots} \alpha_i \sim \min\limits_{\pi_i,\ i=0,1,\ldots}; \\ \sum\limits_{i=0,1,\ldots} \pi_i(x) = 1, \quad x \in \mathcal{R}_N; \\ \pi_i(x) \geqslant 0, \quad x \in \mathcal{R}_N, \quad i = 0, 1, \ldots; \end{cases} \tag{2}$$

$$\text{где } \alpha_i \stackrel{\text{def}}{=} \max_{\mathrm{pr}_\xi^{(t,\,\varphi)}(\cdot) \notin \mathcal{H}_i} \int_{\mathcal{R}_N} \mathrm{pr}_\xi^{(t,\,\varphi)}(x)\, \pi_i(x)\, dx; \quad i = 0, 1, \ldots$$

При этом гипотеза $\mathcal{H}_1$ свидетельствует в пользу случайной формы F_1, гипотеза $\mathcal{H}_2$ — в пользу F_2 и т. д., а гипотеза $\mathcal{H}_0$ означает, что в рамках постановки (3) по предъявленному изображению невозможно определить, изображением какой случайной формы является f.

По аналогии с понятием случайной формы изображения, определим понятие нечёткой формы как возможностное пространство $(\Omega, \mathcal{A}, \mathrm{P})$, где Ω — множество непересекающихся форм, образующих разбиение $\mathcal{R}_N$, $\mathcal{A}$ — σ-алгебра подмножеств Ω, а P — заданная на ней возможность [2].

Тогда задача идентификации нечеткой формы предъявленного изображения будет формулироваться так: введём следующие множества распределений: $\mathbb{P}_i \stackrel{\text{def}}{=} \bigl\{\mathrm{p}_\xi^{(i,\,\varphi)}\bigr\},\ \mathcal{H}_0 \stackrel{\text{def}}{=} \mathbin{\mathbb{\cap}}\mathbb{P}_i,\ \mathcal{H}_i \stackrel{\text{def}}{=} \mathbb{P}_i \backslash \mathcal{H}_0$. Для идентификации предъявленного изображения можно воспользоваться минимаксным критерием $(\widetilde{\pi}_0(x), \widetilde{\pi}_1(x), \ldots),\ x \in \mathcal{R}_N,$ проверки гипотез $\mathcal{H}_0, \mathcal{H}_1, \ldots,$ который минимизирует возможность ошибки и находится как решение задачи на минимакс [2]

$$\begin{cases} \max \alpha_i \sim \min\limits_{\pi_i}, \\ \max \pi_i(x) = 1,\ x \in \mathcal{R}_N \\ \pi_i(x) \geqslant 0,\ x \in \mathcal{R}_N,\ i = 0, 1, \ldots, \end{cases} \tag{3}$$

$$\text{где } \alpha_i \stackrel{\text{def}}{=} \max_{\mathrm{p}_\xi^{(i,\,\varphi)}(\cdot) \notin \mathcal{H}_i} \sup_{x \in \mathcal{R}_N} \min \left(\mathrm{p}_\xi^{(i,\,\varphi)}(x),\ \pi_i(x) \right),\ i = 0, 1, \ldots$$

При этом принятие той или иной гипотезы интерпретируется аналогично рассмотренной выше статистической гипотезе [3].

В докладе рассматриваются алгоритмы эмпирического восстановления [4] случайных и нечетких форм изображения по данным обучающей выборки на примере задачи распознавания рукописных цифр, а также

будет проведено сравнение статистического и возможностного подходов к решению этой задачи.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00532-а, № 05-01-00615-а.

Литература

[1] *Пытьев Ю. П.* Морфологический анализ изображений // Докл. АН СССР. — 1983. — Т. 269, № 5. — С. 1061–1064.

[2] *Пытьев Ю. П.* Возможность как альтернатива вероятности. Математические и эмпирические основы, применения. — М.: Физматлит, 2006.

[3] *Зубюк А. В., Пытьев Ю. П.* Случайная и нечёткая морфология: эмпирическое восстановление модели, идентификация // 9-я межд. конф. «Интеллектуальные системы и компьютерные науки». — 2006.

[4] *Пытьев Ю. П.* Математические методы и адаптивные алгоритмы эмпирического построения теоретико-возможностной модели стохастического объекта // ММРО-13 (в настоящем сборнике). — 2007. — С. 54–56.

О едином подходе к программной реализации фильтрации изображений по локальной окрестности

Юрин Д. В.

yurin_d@inbox.ru

Москва, ВМиК МГУ им. М. В. Ломоносова

Любая практически полезная система обработки изображений включает в себя, наряду с другими компонентами, разнообразные алгоритмы фильтрации изображений. Это могут быть алгоритмы подавления шумов, улучшения изображений, выделения характеристических особенностей, таких как края, уголки, дифференциальные инварианты, текстурные признаки. Часто фильтры используются последовательно, образуя «сложный» фильтр (СФ). Более того, многие широко распространенные фильтры, такие как детекторы краев Канни, DiZenzo или детектор уголков Харриса, по сути, тоже являются СФ, так как их эффективная реализация основана на свойстве сепарабельности функции Гаусса. Такие СФ можно представить в виде направленного ациклического графа, узлами которого являются элементарные фильтры (ЭФ), а ребрами (стрелками) представляются изображения, получаемые в результате обработки ЭФ, и передаваемые на вход следующего фильтра. На Рис. 1 представлен такой граф для детектора Харриса. Прямолинейная программная реализация таких фильтров обычно состоит в выделении памяти под каждое промежуточное изображение (ребро графа на Рис. 1), что приводит к чрезмерным затратам памяти, учитывая, что типичный размер изображения для любительских фотоаппаратов достигает 10 Мпикс., а в аэро-

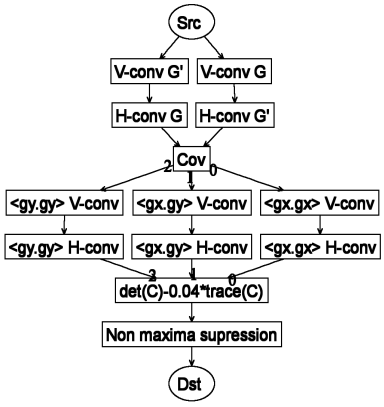


Рис. 1. Детектор характеристических точек Харриса.

космической отрасли перевалил за гигабайт. В то же время, многие из используемых фильтров работают по локальной окрестности точки, и для обработки отдельной строки изображения требуется знание только некоторой полосы вокруг этой строки на исходном изображении.

Постановка задачи

Целью работы являлась разработка программных библиотек для реализации разнообразных фильтраций изображений по локальной окрестности, обеспечивающих невысокие требования к объему оперативной памяти, возможно, ценой незначительного снижения быстродействия, простоту реализации новых фильтров и легкость адаптации к различным способам представления изображения в памяти, включая построчное чтение с жесткого диска.

Структура системы

Требования быстродействия предопределили выбор шаблонного проектирования на C++. Полиморфные методы могут вызываться только на начальной и конечной стадии обработки, и не более нескольких раз на строку изображения. Каждый ЭФ реализуется в виде класса, производного от `FilterElementary`, фильтр имеет N входов (`class InPort`) и M выходов (`class OutPort`). Буфер с полосой изображения содержится в классе `OutPort`; для подключенного к `OutPort` класса `InPort` буфер доступен только для чтения, что обеспечивает существенную экономию в том случае, когда к выходу фильтра подсоединено много входов других фильтров. Задача дополнения изображения за его края (на размер требуемой окрестности) возложена на класс буфера. Для составления СФ

создаются объекты — ЭФ, и выходы одних из них соединяются со входами других, типичный код выглядит как

```
filter1.out3.ConnectTo(&filter2.in);
```

Часто используемые СФ могут быть представлены в виде объекта `FilterComplex`, имеющего общую базу с `FilterElementary`. Источником данных является фильтр с $N = 0$, а приемником — с $M = 0$. Для приведения разработанного фильтра в действие создается объект `FiltersSystem`, через функцию `Assign` ему задаются источники и приёмники, после чего вызывается функция `Run()`. Система сама решает, когда передать управление (`processLine()`) каждому ЭФ на основе готовности входных данных и готовности подключенных фильтров принять данные. Реализация ЭФ состоит в переопределении функции `processLine()`, которой показываются необходимые фрагменты входных изображений в виде массивов, результат обработки полосы следует записать в выходной(ые) порт(ы) через предоставленный(ые) итератор(ы).

Важным компонентом системы является автоматическая настройка параметров и средства отладки. По вызову функции `Assign()` путем прослеживания связей от источника к приемнику с помощью поиска в глубину по графу находятся все задействованные ЭФ, отслеживаются возможные петли, если указано — результат немедленно сохраняется в формате `.dot`, по которому с помощью пакета `GraphViz` может быть визуализирован фактический граф системы как на Рис. 1. Вместе с диагностическими сообщениями это позволяет легко локализовать ошибку соединения фильтров. Повторным поиском от приемников к источникам достигается проверка, что нет неподключенных входов. Мультиграф фильтра сводится к графу и топологически сортируется (ТС). В порядке ТС для каждого фильтра инициализуются значения размеров полных входных изображений и предоставляется возможность вычислить необходимый размер локальных окрестностей для обработки. Теперь можно определить высоту необходимых буферов. Однако, проблемой является то, что размер запрашиваемой локальной окрестности по высоте одновременно является величиной задержки, с которой фильтр будет обрабатывать строку с данным номером. Чтобы система функционировала, должно выполняться условие равенства задержек по всем ветвям входящим в данный узел (фильтр). Такая задача широко известна в области конструирования микросхем [1] как задача балансировки графа информационных потоков, и, после некоторой адаптации, эти методы могут быть с успехом здесь применены, как для балансировки, так и для минимизации суммарного объема буферов в системе в целом. Применяемые алгоритмы также основаны на ТС графа.

Заключение

Тестирование разработанной системы показало, что накладные расходы по времени чрезвычайно малы, структура программы становится прозрачной, а затраты памяти — приемлемыми. Испытания, проведенные на студентах показали, что использование разработанной библиотеки не представляет проблем и ведет к повышению качества кода и ускорению разработки как ЭФ, так и СФ.

Работа выполнена при поддержке РФФИ, проект № 06-01-00789-а.

Литература

[1] *Chatterjee M., Banerjee S., Pradhan D. K.* Buffer Assignment Algorithms on Data Driven ASICs // IEEE Transactions on computers. — January 2000. — V. 49, No. 1. — P. 16–32.

Методы и алгоритмы совмещения изображений и их применение в задачах восстановления трехмерных сцен и панорам, анализе медицинских изображений

Юрин Д. В., Крылов А. С., Волегов Д. Б., Насонов А. В., Свешникова Н. В.

yurin_d@inbox.ru

Москва, ВМиК МГУ им. М. В. Ломоносова

Задача совмещения изображений имеет широкие приложения, особенно если ее рассматривать в расширенном варианте, включающем грубое совмещение изображений, которые не могут быть совмещены точно из-за эффекта перспективы или временной изменчивости. Необходимость какого-либо (хотя бы грубого) совмещения изображений возникает в задачах восстановления трехмерных сцен и панорам для установления взаимнооднозначного соответствия между изображениями характеристических особенностей реальной трехмерной сцены, а при анализе медицинских изображений — между неизменившимися частями органов. Таким образом, подсистема совмещения изображений является обязательной частью задач, упомянутых в названии.

Постановка задачи

Основной целью проводимых работ является построение системы восстановления трехмерных сцен по набору изображений, получаемых с помощью цифрового фотоаппарата, движущегося достаточно произвольно. Сцена предполагается замороженной на время съемки. Идея предлагаемого подхода состоит в том, чтобы предварительно грубо совместить имеющийся набор изображений попарно, после чего задача установления взаимнооднозначных соответствий может решаться методами, близ-

кими к традиционным. Учитывая разнообразие трехмерных сцен, имеет смысл использовать параллельно различные методы предварительного совмещения изображений с выбором наилучшего результата. Часть этих методов может быть с равным успехом использована в задачах построения панорам и обработки медицинских изображений.

Структура системы

Фильтрация изображений. Частью любой системы обработки изображений является подсистема фильтрации — улучшения, выделения характеристических особенностей, и т. п. Разработаны алгоритм предобработки изображений, приводящий к подавлению эффекта Гиббса [1], и единый подход к программной реализации фильтрации по локальной окрестности, названный *потоковой фильтрацией*, на основе которого реализованы детекторы локальных особенностей.

Предварительное совмещение изображений. Изображения, получаемые с помощью фотоаппарата, могут значительно отличаться даже при малых его перемещениях за счет неконтролируемого изменения направления оптической оси. При более существенных пространственных перемещениях начинают сказываться эффекты перспективы. Задача совмещения решается в два этапа: на первом выполняется грубое совмещение изображений как целого в рамках аффинной или проективной модели в зависимости от контекста, на втором — путем построения карты смещений по пирамиде детальности [2]. Задачи первого этапа решаются в общем случае несколькими методами, в частности, по прямым линиям [3] и методом [4]. Так как первый метод основан на вычислении преобразования Хартли, а второй может быть к нему адаптирован, то преобразование вычисляется однократно.

Поиск характеристических точек. Характеристические точки ищутся детектором Харриса, модифицированным для учета цветовых компонент. Соответствие между точками на разных кадрах осуществляется на основе известной грубой карты смещений, правильность соответствия проверяется по дифференциальным инвариантам или путем вычисления корреляции по локальной окрестности.

Первичное восстановление трехмерной сеточной модели. Восстановление трехмерных координат прослеженных точек, позы камер и эпиполярной геометрии и погрешностей этих величин выполняется методом факторизации [5, 6].

Поиск особенностей при ограничениях, уточнение сетки. Полученная сеточная модель не адекватно отображает трехмерную структуру сцены из-за небольшого количества точек и автоматической процедуры триангуляции. Тем не менее, построенная сеточная модель дает

оценку формы сцены и диапазон ожидаемых расстояний до других точек сцены. Это позволяет произвести поиск дополнительных соответствующих точек при найденных ограничениях на паре кадров [7]. Учитывая, что матрица движения уже известна, путем репроецирования можно получить оценку координат изображений точки на всех кадрах и проверить сходство по локальной окрестности. Грубая сеточная модель позволяет также строить гипотезы о том, чем вызвано расхождение — ошибочно найденным соответствием или закрытием более близким объектом. Точка, прослеженная по всей последовательности изображений, может быть использована для уточнения модели путем перерасчета факторизации. В противном случае она просто уточняет сетку.

Следующий шаг — добавление виртуальных точек, образованных пересечением обнаруженных на изображении линий детектором границ и эпиполярных линий, что приводит к существенному улучшению сеточной модели за счет добавления большого количества точек, особенно на скачках глубины сцены. На этом шаге контролируются два параметра: угол пересечения линий не должен быть слишком острым, а погрешность эпиполярной линии — слишком большой.

Поиск закономерностей формы и проверка гипотез. Имея трехмерные координаты восстановленных точек с погрешностями, можно путем иерархического преобразования Хафа выявить точки, возможно принадлежащие плоским поверхностям сцены. На основе выделенных подмножеств точек вычисляется уравнение плоскости, осуществляется проективное преобразование изображений в единую проекцию и проверка этой гипотезы путем аффинного совмещения изображений алгоритмом Ши-Томаси [2].

Работа выполнена при поддержке РФФИ, проект № 06-01-00789-а.

Литература

[1] *Nasonov A., Krylov A., Lukin A.* Post-processing by Total Variation Quasi-solution Method for Image Interpolation // Conf. Graphicon, Moscow, 2007.

[2] *Волегов Д. Б., Юрин Д. В.* Поиск карты смещений по пирамиде детальности // Конф. Графикон, Москва, 2007.

[3] *Волегов Д. Б., Юрин Д. В.* Грубое совмещение изображений по найденным на них прямым линиям // Конф. Графикон, Новосиб., 2006. — С. 463–466.

[4] *Reddy B. S., Chatterji B. N.* An FFT-based technique for translation, rotation, and scale-invariant image registration // IEEE PAMI. — 1996. — V. 5, No. 8. — P. 1266–1271.

[5] *Свешникова Н. В., Юрин Д. В.* Алгоритмы факторизации: достоверность результата и применение для восстановления эпиполярной геометрии // Конф. Графикон, Новосибирск, 2006. — С. 158–165.

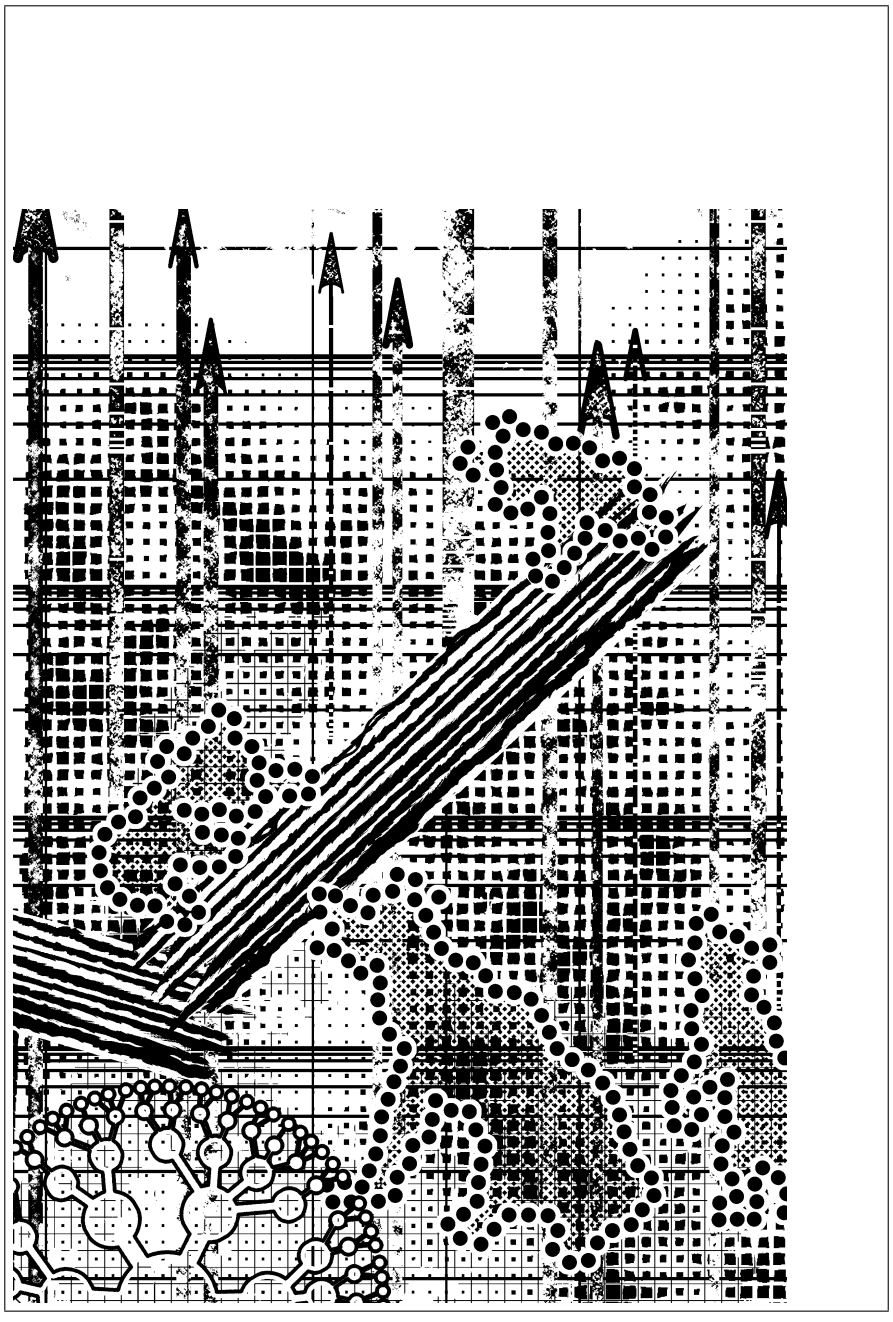
[6] *Калиниченко А. В., Свешникова Н. В., Юрин Д. В.* Эпиполярная геометрия и оценка ее достоверности по результатам восстановления трехмерной сцены алгоритмами факторизации // Конф. Графикон, Новосиб., 2006. — С. 343–346.

[7] *Свешникова Н. В.* Уточнение сеточной модели трехмерной сцены, предварительно восстановленной по малому количеству характеристических точек // Конф. Графикон, Москва, 2007.

Прикладные задачи интеллектуального анализа данных

Код раздела: AP (Applied Problems)

- Прикладные задачи распознавания и прогнозирования в биоинформатике, медицине, технических науках, химии, социологии, экономике, лингвистике.
- Анализ и понимание текста (text mining).
- Анализ данных о содержимом, структуре и посещаемости документов в сети Интернет (web mining).



Развитие методов искусственного интеллекта и обработки данных на примере анализа патологий сетчатки

Анисимов Д. Н., Астахова Ю. Ю., Вершинин Д. В., Зуева М. В., Колосов О. С., Мамакаева И. Р., Резвых С. В., Титов Д. А., Хрипков А. В., Цапенко И. В., Шевченко М. В.
anisimovdn@mpei.ru

Научная проблема, которой посвящена данная работа, заключается в дальнейшем исследовании и развитии теории нечетких множеств, а также временных рядов при диагностировании сложных плохо определенных (нечетких, неточных) проблемных ситуаций на основе экспертных знаний. В качестве примера рассматриваются способы диагностирования сложных патологий сетчатки.

Отличительной особенностью исследуемой проблемной области является то, что, помимо имеющихся объективных знаний о проблемной ситуации и возможном диагнозе, часто имеющих статистический характер, существенную роль играют также субъективные, эмпирические знания специалистов-экспертов (физиологов), отражающие накопленный ими опыт. В этой связи представляется актуальной организация процесса сбора нечетких экспертных знаний, их анализа и последующей обработки с использованием методов искусственного интеллекта таким образом, чтобы в итоге получить достаточно формализованное описание проблемной ситуации, позволяющее с приемлемой степенью правдоподобия или даже достоверно её диагностировать с целью последующего принятия решения о наиболее предпочтительном лечении.

Для решения поставленной проблемы необходимо исследование и определение набора специфических черт исходной информации для возможности идентификации и диагностики патологий сетчатки. Процесс анализа данных происходит на основе массивов показателей клинических и электроретинографических исследований больных с заболеваниями сетчатки различного генеза, а также знаний экспертов — специалистов-физиологов. Электроретинограмма (ЭРГ) представляет собой графическое отображение изменений биоэлектрической активности нейронов сетчатки в ответ на световое раздражение и зависит от количества здоровых функционирующих нейронов. Каждый из компонентов ЭРГ генерируется различными структурами сетчатки. На основании вариации формы ЭРГ имеется электрофизиологическая классификация ретинальных патологий.

В работе предлагаются два подхода к решению поставленной задачи.

1. Использование оригинальных алгоритмов логического вывода на основе нечетких ситуаций и нечетких соответствий для формализа-

ции экспертных знаний, позволяющих учитывать субъективные знания экспертов-физиологов и наиболее существенные зависимости между наблюдаемыми симптомами.

2. Анализ динамических характеристик сетчатки. При этом используются: оригинальный метод экспоненциальной модуляции для идентификации динамических объектов; известные методы поиска экстремумов функций; методы анализа частотных характеристик ЭРГ.

На данный момент сформирована значительная часть базы данных и базы знаний, которые позволяют ставить в соответствие область предпосылок (уровень и время экстремумов ЭРГ) и область заключений (предположительных диагнозов).

На основе эмпирических данных составлена структурная модель сетчатки как динамического объекта. Показано, что с высокой степенью достоверности она может быть представлена параллельным соединением звеньев третьего порядка и второго порядка с запаздыванием.

При параметрической идентификации динамических характеристик используются метод экспоненциальной модуляции и метод покоординатного спуска.

Кроме того, при подаче на зрачок периодического сигнала в виде кратковременных вспышек, получены спектры выходного сигнала. Эти данные могут быть весьма информативными. Затруднением в данном случае является неизвестность спектра входного сигнала. Однако, на основании результатов, полученных с помощью вышеизложенных методов, вполне возможно отделить динамические характеристики входного сигнала от характеристик объекта.

Работа выполнена при поддержке РФФИ, проект № 07-01-00-762.

Литература

[1] *Анисимов Д. Н.* Использование нечеткой логики в системах автоматического управления // Приборы и системы. Управление, контроль, диагностика. — 2001. — № 8. — С. 39–42.

[2] *Баларев Д. А., Вершинин Д. В., Зуева М. В., Колосов О. С., Цапенко И. В.* Динамическая модель сетчатки глаза для целей диагностики патологий методами искусственного интеллекта // Тр. XVI Международного семинара «Современные технологии в задачах управления, автоматики и обработки информации», Алушта: 2007. — С. 5.

Многомасштабный динамический анализ корреляционного типа в исследовании ЭЭГ записей эпилептических разрядов

Анциперов В. Е., Морозов В. А., Обухов Ю. В.
antciperov@cplire.ru
Москва, Институт радиотехники и электроники РАН

В докладе излагаются результаты применения разработанного авторами метода многомасштабного динамического анализа нестационарных процессов к задачам локализации и определения динамических характеристик эпилептических разрядов. Обсуждаются основы время-временного графического представления, приведена интерпретация типичных фрагментов представления; для типичных фрагментов приведены иллюстрации на основе реальных ЭЭГ записей.

Многомасштабный динамический корреляционный анализ

На протяжении ряда лет авторы доклада занимались проблемой исследования нестационарных сигналов медико-биологического происхождения, используя аппарат коротких кросс-корреляционных функций [1–3]. В результате удалось выявить ряд значимых характеристик коротких корреляционных функций, а именно — значения и положения боковых пиков, и интерпретировать эти параметры в терминах степени квазипериодичности, значений периода основного колебания, его динамики и т. д.

Изначально короткие корреляционные функции формировались на некоторых характерных для сигнала, но фиксированных временных окнах. В дальнейшем было обнаружено, что гораздо лучшие и полные результаты получаются, если подобный анализ проводить на нескольких временных масштабах, т. е. если положить многомасштабность в основу разрабатываемого подхода. Однако, при этом возникает проблема большого числа переменных параметров метода — текущее время, размер окна, временной сдвиг — что приводит к потере наглядности и обозримости представления. Решением проблемы послужила одна из идей вейвлетного частотно-временного анализа. Преимущество вейвлетных преобразований состоит в использовании переменных временных окон: малых для высоких частот и больших для низких. Во время-временном анализе, где период обратно пропорционален частоте, это означает, что для подчеркивания квазиколебаний с малым периодом необходимо использовать малые окна, для обнаружения больших периодов — большие.

В итоге, для обнаружения квазипериодических фрагментов (в частности, эпилептического разряда) была выбрана следующая мера квазипериодичности:

$$r(t,\vartheta) = \frac{\int G^2(2t'/\vartheta)x(t'+t-\vartheta/2)x(t'+t+\vartheta/2)dt'}{\sqrt{\int G^2(2t'/\vartheta)x^2(t'+t-\vartheta/2)dt'}\sqrt{\int G^2(2t'/\vartheta)x^2(t'+t+\vartheta/2)dt'}},$$

где t — текущее время, θ — временная шкала метода, $\mathrm{x}(t')$ — анализируемый сигнал, $\mathrm{G}(t')$ — масштабирующее окно. Для представления динамических характеристик сигнала используется плоскость (t,θ) с раскраской псевдоцветом величины меры $r(t,\theta)$.

Исследование ЭЭГ записей эпилептических разрядов

Наблюдаемые на электродах спонтанные колебания электрических потенциалов мозга на интервалах эпилептического разряда имеют вид пакетов квазипериодических колебаний со средней частотой 3–5 Гц («тэта-диапазон»). Длительность и форма элементарных сигналов в таком пакете варьируют от «периода» к «периоду», но в целом приближённо периодическая структура пакета сохраняется достаточно долго (вполне реально в течение 5–10 секунд).

Поскольку изучаемый процесс явно обладает временной многомасштабностью, предложенный метод должен позволить судить в рамках единого образа о динамике процесса, как на масштабе порядка основного периода ($T = T_0 + \delta_i$), так и на значительно более длительных интервалах при априори неизвестной и, часто, достаточно сложной форме элементарных сигналов.

В наших исследованиях анализируемая метрика $r(t,\theta)$ представляет собой коэффициент корреляции исследуемой реализации $x(t)$ для двух соседних интервалов («полуокон») равной длительности ($T_1 = T_2 = \theta$), сдвинутых на θ. Специфика такого построения метрики в том, что временной масштаб фигурирует одновременно в длительности и сдвиге полуокон, по которым формируется коэффициент корреляции $r(t,\theta)$. Многомасштабность процесса определяет отображение используемой метрики на плоскости $(t-\theta)$. (Предполагаем, что средние значения выборок удалены на этапе предварительной обработки.)

Рельеф $r(t,\theta)$ на плоскости $(t-\theta)$ передаётся условным цветом и формируемая картина даёт хорошее зрительное представление о периодичности исходного процесса и её нарушениях в виде вариаций основной частоты и фазы основного «ритма». На Рис. 1 представлены наиболее часто встречающиеся фрагменты представления ЭЭГ записей эпилептических разрядов.

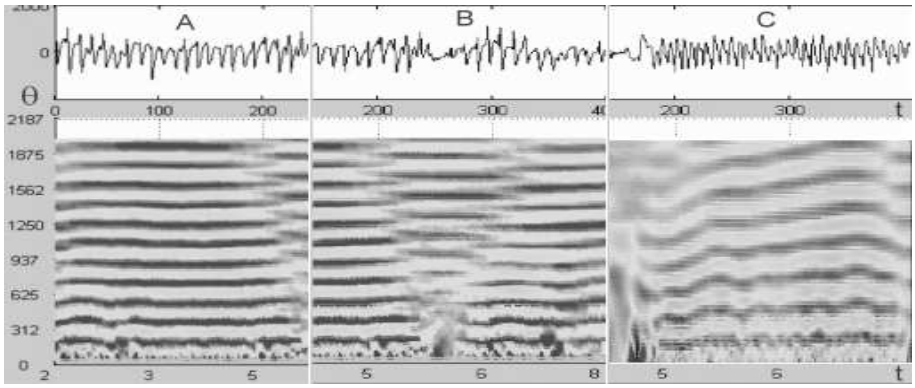


Рис. 1. Типичные фрагменты представления ЭЭГ записей: A: Стационарный участок разряда; B: Скачок фазы внутри разряда; C: Тренд основного периода разряда.

Заключение

Разработанное представление для ЭЭГ записей эпилептических разрядов оказывается в высокой степени информативным и легко интерпретируемым. С его помощью можно «на взгляд» определить границы разряда, особенности его динамики, такие как стационарность, уход частоты колебаний, наличие фазовых скачков.

Авторы надеются, что предложенный анализ и его дальнейшее развитие будут столь же плодотворны, как и ставшие общепринятыми спектральный или вейвлетный подходы.

Работа выполнена при поддержке РФФИ, проекты № 06-01-00754-а, № 06-07-89302-а.

Литература

[1] *Antciperov W. E.* Vowels Detection / Recognition on the Base of Short Cross-correlation Function Side Peak Parameters // Proc. of 11-th Int. Conference Speech and Computer (SPECOM'2006), S-Peterburg: 2006. — P. 400.

[2] *Antciperov V. E., Morozov V. A.* The Dynamics of Characteristics of Short Autocorrelation Functions of Speech Signals // J. Communications Technology and Electronics. — 2004. — V. 49, No. 12. — Pp. 1333–1341.

[3] *Antciperov V. E., Morozov V. A., Nikitov S. A.* Isolated-Word Segmentation Based on the Dynamics of the Parameters of Short Correlation Functions // J. Communic. Technology and Electronics. — 2006. — V. 51, No. 12. — P. 1356.

Современный подход к традиционной балистокардиографии: измерения, обработка данных и диагностика

Анциперов В. Е., Морозов В. А., Сударев А. М.
antciperov@cplire.ru
Москва, Институт радиотехники и электроники РАН

Содержанием доклада являются последние результаты, полученные в ходе исследований в области торсионной балистокардиографии, выполненных в рамках проекта РФФИ № 06-01-00754-а — «Применение метода динамической сегментации квазипериодических сигналов в прикладных задачах контроля сердечного ритма при бесконтактной балистокардиографии».

Описана принципиально новая техника (основанная на высокоточных сейсмических датчиках вращения) измерения механической активности сердца и сердечно-сосудистой системы. Обсуждаются особенности балисто-кардиографических сигналов и новый (основанный на адаптивно-корреляционной обработке) метод сегментации и выделения значимых параметров регистрируемых данных. Приведены экспериментальные результаты, полученные применительно к прикладной задаче мониторинга сердечного ритма, в частности, обсуждаются вопросы борьбы с мешающими факторами, обусловленными другими физиологическими процессами.

Безконтактное измерение сердцечной механической активности

Физическим источником балисто-кардиографического сигнала является гидромеханическая активность сердечно-сосудистой системы: сердце, являясь своеобразным насосом прокачивает через кровеносные сосуды около 70–80 мл крови за цикл. Побочным эффектом сердечного выброса крови является импульс отдачи (Рис. 1), который передается телу пациента и затем, соответственно, ложементу. В классической балистокардиографии именно этот импульс измеряется датчиком (чаще всего пьезоэлектрического типа), прикрепленным к ложементу.

Очевидно, что движение крови по замкнутым кровеносным сетям приводит также к появлению и механического вращательного момента. В соответствии с законом сохранения полного механического момента в замкнутой системе, тело пациента (и ложемент) при этом приобретает некоторый угловой момент вращения, противоположный моменту массы крови, динамика которого может быть измерена (Рис. 1).

Для измерения углового момента была создана (Constel Ltd) специальная установка, включающая трехкомпонентный торсионный сейсмический датчик (динамический диапазон более 120 дБ) METR–003.

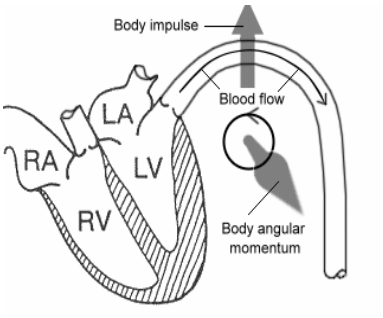


Рис. 1. Механический импульс отдачи и момент импульса, обусловленные выбросом крови из левого желудочка сердца.

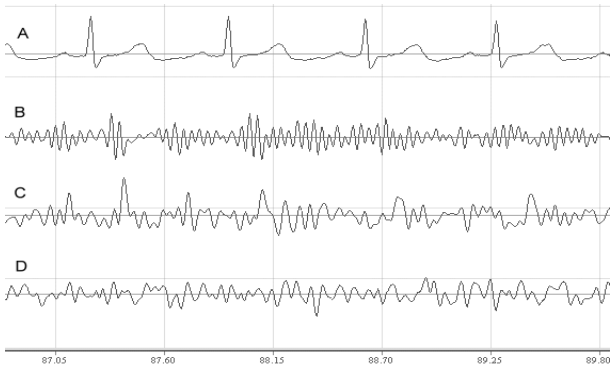


Рис. 2. Типичные данные, полученные на ТБКГ установке в процессе мониторинга сердечной активности. A — ЭКГ-запись сердечного ритма; B, C, D — Сигнал БКГ по осям X, Y, Z.

Он имеет следующие технические характеристики: частотный диапазон: не менее 0.03–100 Гц, коэффициент преобразования: не менее 50 В/рад-1. Совместно с сейсмобаллистодатчиком используется регистратор ЭКГ-сигналов. Типичные измерения, полученные на данной установке, представлены на Рис. 2.

Обработка ТБКГ сигнала с целью выделения сердечного ритма

Существо адаптивного подхода при обработке фрагментов TBCG сигнала заключалось в следующем (Рис. 3). Фрагмент сигнала обрабатывался двумя окнами длительностью нескольких ожидаемых периодов (длительность окна является параметром процедуры и может устанавливаться оператором). Одно из окон является основным, оно выделяет анализируемый сегмент рассматриваемого фрагмента. Вдоль основ-

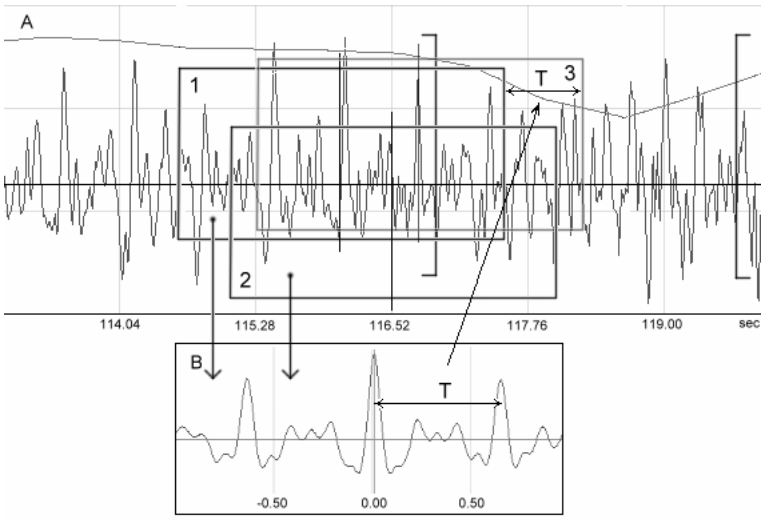


Рис. 3. Графическое представление схемы адаптивно-корреляционной обработки. A — выделенный квази-периодический участок сигнала обрабатывается двумя окнами — основным 1 и подвижным 2; данные из обоих окон используются для расчета корреляционной функции B. По корреляционной функции находится смещение бокового максимума T, который рассматривается как текущий ритм сердца и используется в алгоритме для определения нового положения основного окна 3.

ного окна перемещается вспомогательное, которое выделяет смещенный относительно анализируемой части сегмент. На основе обоих сегментов формируется корреляционная функция, определяется смещение бокового максимума, которое и берется в качестве оценки текущего периода сердечного ритма. Далее основное окно смещается на величину полученной оценки, процедура повторяется и так, пока не будет обработан весь фрагмент. Схематически адаптивная процедура оценки ритма показана на Рис. 3.

Экспериментальные результаты

Для решения поставленной задачи (бесконтактной регистрации длительности кардиоциклов) был проведен ряд экспериментов на бодрствующих и спящих испытуемых, во время которых одновременно регистрировались ЭКГ и трёхосевая ТБКГ. По данным измерений проводилась обработка результатов с целью сравнения длительности кардиоциклов полученных с помощью ЭКГ и с помощью адаптивной автокорреляционной обработки ТБКГ. Практически у всех испытуемых во время отно-

сительного покоя (в отсутствие помех от «макродвижений», таких как перевороты, смены позы, и т. п.) результаты совпали по практически каждому кардиоциклу с точностью не хуже 5–10 мсек.

Работа выполнена при поддержке РФФИ, проект № 06-01-00754-а.

Литература

[1] *Baevsky R. M., Bogomolov V. V., Funtova I. I.* Less Contact Ballistogram Recording during Sleep as a Perspective Technology for the Medical Monitoring System in a Mission to Mars // Proc. of 34th COSPAR Scientific Assembly, The Second World Space Congress, 2002. — C. G-P–05.

[2] *Velichko A. D., Sudarev A. M., Kadin I. L., Isaev I. A.* Soft hardware complex for fuctional diagnostics of cardiovascular system //Proc. of 3-d International Scientific-Practical Conference «Noninvasive monitoring of cardiovascular system in clinical practice», Moscow, 2001. — Pp. 160–163.

[3] *Анциперов В. Е.* Метод коротких корреляционных функций в задачах структурирования сигналов сложной природы // докл. конф. Математические методы распознавания образов (ММРО-12), Москва, 2005. — С. 10–13.

[4] *Eblen-Zajjur A.* A simple ballistocardiographic system for a medical cardiovascular physiology course. — Advan. Physiol. Edu., 2003. — С. 224–229.

[5] *Antciperov W. E., Morozov V. A., Nikitov S. A.* Vowel Detection and Recognition on the base of Short Correlation Function Parameters Dynamics //Proc. of 10-th International Conference Speech and Computer (SPECOM'2005), Patras, Greece, 2005. — С. 535–538.

Методы выделения признаков двумерных спектров нестационарных биомедицинских сигналов

Боснякова Д. Ю., Морозов А. А., Кузнецова Г. Д., Обухов Ю. В.

obukhov@cplire.ru

Москва, ИРЭ РАН, ИВНД РАН

В биологических исследованиях и медицинской диагностике широко применяются многоканальные приборы, измеряющие различные физические величины на поверхности и в глубине тела. Характерными примерами таких обследований являются электроэнцефалография (ЭЭГ, измерения электрических потенциалов на скальпе и внутри мозга) и магнитная энцефалография (пассивные измерения магнитного поля вокруг головы). Число каналов в таких исследованиях составляет от нескольких десятков до нескольких сотен. Целями анализа результатов таких исследований являются выяснение временно́й динамики сигналов в различных каналах, установление связи между пространственно разнесёнными участками мозга, источников генерации сигналов, изменения сиг-

налов при физиологических и фармакологических воздействиях на объект, и т. п.

Особенностями такого рода сигналов являются существенная их нестационарность, зашумлённость и присутствие артефактов. Обработке и анализу таких сигналов посвящено большое количество работ, в основе которых лежат спектрально-корреляционные методы и методы нелинейных динамических систем [1]. Основным недостатком этих подходов является их слабая приспособленность к анализу нестационарности сигналов; как правило, в них сигнал разбивается на квазистационарные участки, внутри которых эти методы применимы. Исключением являются оконные преобразования и вейвлет-анализ, в котором частотно-временная плоскость разбивается на предельно малые участки, размеры которых ограничиваются фундаментальным соотношением неопределённости [2]. Используя такие представления сигналов, можно попытаться выделить именно нестационарные характеристики исследуемого объекта. Это и является предметом наших исследований.

Основная идея нашего подхода заключается в следующем. С помощью вейвлет-преобразований и скользящих оконных корреляций мы получаем двумерное изображение каждого сигнала, заданное на плоскости частота-время. В этом изображении присутствуют объекты и структуры, порождённые различными физиологическими процессами. Выделив эти объекты и структуры, мы можем вычислить характеристики нестационарных процессов.

В настоящей работе представлены разработанные нами методы и результаты выделения хребтов изображений двумерных спектрограмм ЭЭГ, которые отображают доминирующие процессы. Они экспериментально проверялись на примере анализа ЭЭГ животных и людей с абсансной эпилепсией. Выбор данного заболевания обусловлен как его социальной значимостью, так и тем, что при абсансной эпилепсии происходит существенная синхронизация электрической активности мозга, и выявление пространственно-временных связей участков мозга чрезвычайно актуально.

Выделение хребтов вейвлет-спектров

Вейвлет-спектры Морле ЭЭГ эпилептического разряда пик-волна (называемые SWD-разрядами) содержат систему ярко выраженных пиков. Точки хребта приближённо соответствуют точкам, где $\partial\varphi/\partial t = \mathrm{const}$, где φ — комплексная фаза сигнала [2]. Однако, дифференцирование фазы слабого зашумлённого сигнала является некорректно поставленной задачей. Для её решения мы разработали вейвлет, парный к вейвлету Морле, который зануляет амплитуды спектрограмм в точках хребта, т. е. проецирует хребет на плоскость частота-время.

Трассируя спектрограмму с помощью круга с задаваемым радиусом, можно получить зависимость частоты от времени в точках хребта (траверс хребта). Траверсы хребтов являются характеристиками нестационарности SWD-разряда при абсансной эпилепсии.

Обработка ЭЭГ-сигналов при абсансной эпилепсии

Обработка ЭЭГ-сигналов модельных животных и людей с абсансной эпилепсией показала, что траверс хребта отличается в разных участках мозга, что даёт информацию о месте возникновения эпилептического разряда. Сравнение траверсов сигналов в разных каналах позволяет оценить корреляцию между сигналами различных участков мозга. Он чувствителен к фармакологическим препаратам [3]. Оказалось также, что форма траверса у модельных животных и у людей с абсансной эпилепсией имеют одинаковую форму, но траверсы разнесены по частоте. Таким образом, разработанный подход позволяет выделить нестационарные пространственно-временные признаки сигналов многоканальных ЭЭГ при абсансной эпилепсии.

Недостатком разработанного метода является необходимость вручную задавать начальную и конечную точки траверса хребта. В случае фазовых сбоев, например, при смене типа разряда, это может представлять определённые неудобства. Для нахождения таких точек и участков спектрограмм, а также для анализа корреляционных свойств сигналов ЭЭГ при абсансной эпилепсии нами также разрабатывается подход, основанный на многомасштабном динамическом анализе корреляционного типа [4]. Другим недостатком является его невысокая эффективность при анализе вызванных различными стимулами электрических потенциалов мозга. Для таких коротких событий мы разрабатываем методы непараметрического многофакторного анализа [5, 6], в частности для анализа реакции мозга на зрительные стимулы [7]. Эти подходы описаны в отдельных публикациях данной конференции.

Работа выполнена при поддержке РФФИ, проекты № 05-01-00651 и № 06-07-89302, а также программы Президиума РАН «Фундаментальные науки — медицине».

Литература

[1] *Bin He (ed.)* Neural Engineering. — New York: Kluwer, 2005.

[2] *Малла С.* Вейвлеты в обработке сигналов. — Москва: Мир, 2005.

[3] *Bosnyakova D., Gabova A., Kuznetsova G., Obukhov Yu, et al.* Time-frequency analysis of spike-wave discharges using a modified wavelet transform // J. Neuroscience Methods. — 2006. — Vol. 154, June, № 1-2. — pp. 80–88.

[4] *Анциперов В. Е., Морозов В. А., Морозов А. А., Обухов Ю. В.* Адаптивный метод на основе коротких кросскорреляционных функций для обработки

сигналов биологической природы // 7 межд. научно-техн. конф. «Искусственный интеллект. Интеллектуальные и многопроцессорные системы» (ИИ-ИМС'2006), Таганрог: Изд-во ТРТУ, 2006. — Т. 3. — С. 162–165.

[5] *Морозов А. А., Морозов В. А., Обухов Ю. В., Строганова Т. А.* Метод многофакторного анализа электроэнцефалограмм человека на основе вейвлет-спектрографии и непараметрической статистики // Доклады VII межд. научно-техн. конф. «Физика и радиоэлектроника в медицине и экологии» (ФРЭМЭ'2006), Владимир: Изд-во «Собор», 2006. — Книга 1. — С. 145–147.

[6] *Морозов А. А., Морозов В. А., Обухов Ю. В., Строганова Т. А.* Непараметрический метод многомерного многофакторного анализа электроэнцефалограмм человека // Искусственный интеллект. — 2006. — № 3. — pp. 603–612.

[7] *Stroganova T. A., Orekhova E. V., Prokofyev A. O., Posikera I. N., Morozov A. A., Obukhov Y. V., Morozov V. A.* Atypical event-related potentials response to illusory contour in boys with autism // NeuroReport. — 2007. — Vol. 18, № 9. — pp. 931–935.

Эволюционный подход к задаче кластеризации на концептуальных графах и его применение в системах поддержки электронных библиотек

Богатырев М. Ю., Латов В. Е., Столбовская И. А., Тюхтин В. В.

okkambo@mail.ru

Тульский государственный университет

Одним из направлений развития современных электронных библиотек является расширение их функциональных возможностей. Такое расширение было бы значительным, если бы в библиотеках хранились не только тексты, но и их смысловое содержание. Решить эту задачу можно, исследуя и реализуя семантические модели текстов.

В данной работе предлагается применить одну из семантических моделей текста — *концептуальный граф* [1] — в качестве объекта хранения электронной библиотеки.

Применение концептуальных графов позволяет развивать технологии поддержки электронных библиотек, по крайней мере, в двух направлениях:

1) автоматизация построения каталогов библиотек, модификация и коррекция существующих каталогов на основе анализа потока входных текстов;
2) извлечение знаний из электронных библиотек в виде *концепций* и *онтологий*.

Оба указанных направления в настоящее время представлены множеством методов и технологий [2, 3]. Эффективное применение здесь концептуальных графов связано с решением задач *агрегирования* и *кластеризации* на графах.

В работе содержится обзор современных подходов к использованию концептуальных графов и некоторые результаты авторов, касающиеся решения задач кластеризации на концептуальных графах.

Концептуальные графы и их применение

Вместе с *концептуальными решетками* концептуальные графы относятся к *концептуальным структурам*, которые являются одним из формальных представлений знаний [4].

Концептуальный граф — это двудольный направленный граф, состоящий из двух типов узлов: *концептов* и *концептуальных отношений*.

Разработан стандарт представления концептуальных графов [1] и языки их описания, среди которых наиболее популярны CGIF (Conceptual Graph Interchange Form) и XML-представление концептуальных графов.

В системах поддержки электронных библиотек важными задачами являются задачи автоматической классификации и автоматической фильтрации входных документов при известном множестве тематических интересов. Данное множество может быть представлено как система концепций (концептов), выражаемая через концептуальные графы.

В результате, задачи классификации сводятся к решению фундаментальной задачи кластеризации на графах.

Кластеризация на концептуальных графах

В любой задаче кластеризации важной проблемой является построение меры близости кластеризуемых объектов. Под мерой близости концептуальных графов понимается количественная характеристика, призванная отразить семантическую близость порождающих графы предложений, что сделать, очевидно, полностью невозможно. Поэтому проблема близости остается центральной в анализе концептуальных графов.

Для двух графов G_1 и G_2 мера близости зависит от двух значений: концептуальной близости s_c и относительной близости s_r.

$$s_c = \frac{2n(G_c)}{n(G_1) + n(G_2)}, \tag{1}$$

где $G_c = G_1 \cap G_2$, $n(G)$ — число концепций — концептуальных узлов графа G.

$$s_r = \frac{2m(G_c)}{m_{G_c}(G_1) + m_{G_c}(G_2)}, \tag{2}$$

где m_{G_c} — число отношений — относительных узлов концептуального графа G_c, $m_{G_c}(G)$ — число отношений — относительных узлов концептуального графа G, для которых хотя бы одна из вершин принадлежит графу G_c.

Анализ мер близости (1), (2), выполненный в работе, демонстрирует их несовершенство. В результате применения мер близости (1), (2) близкими могут оказаться графы, имеющие просто большое число одинаковых концептов или отношений, но по смыслу совершенно далекие. Поэтому вместо мер близости (1), (2) предлагается использовать их модификации:

$$s_c = \frac{2n(G_c)l}{n(G_1) + n(G_2)}, \tag{3}$$

где

$$l = \begin{cases} k\frac{n(G_1)}{n(G_2)}, & \text{если } n(G_1) \geqslant n(G_2); \\ k\frac{n(G_2)}{n(G_1)}, & \text{если } n(G_1) < n(G_2); \end{cases}$$

k — масштабирующий коэффициент.

Для формулы (2):

$$m_{G_c}(G) = m_{\text{both}} + b_1 + b_2 + \cdots + b_i, \quad i = 1, \ldots, m - m_{\text{both}},$$

$b_i \in \{0, 1\}$, где m — число всех отношений графа G, m_{both} — число всех отношений графа G, для которых обе вершины принадлежит графу G_c, b_i — коэффициент общезначимости — принимает значения от 0 до 1, в зависимости от типа отношения.

Под мерой близости двух концептуальных графов, принимающей значение от 0 до 1, будем понимать значение

$$s = d_1 s_c + d_2 s_r, \tag{4}$$

где d_i — масштабирующие коэффициенты, определяемые экспериментально.

Как следует из экспериментов, данная мера близости повышает качество кластеризации, но не решает проблему в целом.

Эволюционный подход к кластеризации концептуальных графов

Изменение коэффициентов k и b_i, введенных выше, влечет появление множественных вариантов кластеризации даже в рамках одной меры близости концептуальных графов.

В работе предполагается исследование мер близости на концептуальных графах в рамках эволюционного подхода к решению задачи кластеризации, который состоит в применении эволюционных алгоритмов

оптимизации при поиске экстремума целевой функции, отражающей меру близости графов друг другу.

Коэффициенты k и b_i в этом случае входят в целевую функцию в качестве параметров.

Эволюционные алгоритмы основаны на генетическом алгоритме.

1. Генетический алгоритм работает с множеством (популяцией) приближений решения задачи. В результате оптимальное решение может быть найдено как множество различных объектов, что характерно для задач со сложными целевыми функциями.
2. Генетический алгоритм эффективен при поиске экстремума целевых функций, имеющих конечные разрывы, что имеет место для функций, являющихся мерами близости концептуальных графов.

Известны несколько подходов к решению задач кластеризации с применением генетических алгоритмов [6, 7]. Реализация эволюционного подхода требует настройки параметров алгоритма: выбора системы кодирования, способа рекомбинации хромосом, введения механизма мутации [8].

В работе исследованы варианты настроек параметров генетического алгоритма классификации. В частности, построена специфическая кодировка решений, дающая высокое качество кластеризации. Кодировка использует хромосомы вида $a_1 a_2 \dots a_n$, где $a_i \in \{1, \dots, n\}$, n — количество объектов кластеризации.

В докладе представлены результаты вычислительных экспериментов на конкретном материале текстов аннотаций научных статей. Обсуждаются вопросы реализации подхода в информационной системе — пилотном проекте электронной библиотеки научных статей.

Работа выполнена при поддержке РФФИ, проект № 07-07-00276-а.

Литература

[1] *Sowa R.* Conceptual Graphs: Draft Proposed American National Standard // International Conference on Conceptual Structures ICCS-99, Lecture Notes in Artificial Intelligence 1640, Springer 1999.

[2] *Городецкий В. И., Самойлов В. В., Малов А. О.* Современное состояние технологии извлечения знаний из баз и хранилищ данных // Журнал Российской ассоц. искусственного интеллекта. — 2002. — № 3. — С. 3–31.

[3] *Hirst G.* Ontology and the Lexicon // Handbook on Ontologies in Information Systems, Berlin: Springer, 2003.

[4] *Sarbo J.* Formal conceptual structure in language. In Dubois, D. M., editor, Proceedings of Computing Anticipatory Systems (CASYS98).— Woodbury, New York, 1999.— Pp. 289–300.

[5] *Montes-y-Gomez, Gelbukh, Lopez-Lopez, Baeza-Yates* Flexible Comparison of Conceptual Graphs // Lecture Notes in Computer Science 2113, Springer-Verlag, 2001.

[6] *Maulik U., Bandyopadhyay S.* Genetic algorithm-based clustering technique // Pattern Recognition. — 2000. — vol. 33. — Pp. 1455–1465.

[7] *Kivijarvi Juha, Lehtinen Joonas , Nevalainen Olli* A Parallel Genetic Algorithm for Clustering // Turku Centre for Computer Science, Tech. report № 469, August 2002.

[8] *Богатырёв М. Ю., Латов В. Е.* Исследование генетических алгоритмов кластеризации // Изв. ТулГУ. Сер. Математика. Механика. Информатика., Тула, 2002. — Т. 8, вып. 3. Информатика.— С. 101–107.

Агрегированное равновесие лабораторных сетевых рынков

Голубцов А. А.

business2002@bk.ru

Москва, Вычислительный центр РАН

В основе экспериментальной экономики лежит проверка экономических гипотез в контролируемых условиях лаборатории.

Среди всех типов экспериментов можно выделить специальный класс со случайным составом группы. Такой эксперимент состоит из нескольких периодов. В каждом из периодов участники случайным образом разбиваются на группы, причем никто не знает, в какую группу он попал. В каждой из групп реализуется игра с одновременным и независимым выбором действий. После проведенной игры каждый участник знает действия только тех игроков, с которыми он оказался в одной группе в данном периоде. Этот подход, в частности, заложен в программную оболочку Z-Tree (Университет Цюриха, Швейцария) [2]. Смысл такой структуры эксперимента состоит в том, чтобы уменьшить эффекты повторяющейся игры, связанные с использованием стратегий угроз и сговора [2, 3].

В работе предлагается рассматривать такие эксперименты как расширение игры в нормальной форме. Для этого вводится понятие k-расширения игры. Игра происходит в два этапа. На первом этапе случайным образом формируется k групп по n игроков, и игрокам в каждой группе приписываются номера классов от 1 до n. На втором этапе каждая группа играет в исходную игру G. Полученная таким образом игра называется k-расширением игры G.

Формально, данная структура эксперимента задает довольно сложную динамическую игру с неполной информацией. Для такой игры обычно используется понятие совершенного байесовского равновесия [3] или

аналогичные понятия [4], которые основаны на параллельном рассмотрении стратегий игрока и его представлений о стратегиях остальных. Полное исследование равновесий данной динамической игры не представляется возможным.

Вводится понятие агрегированного равновесия игры в нормальной и байесовской форме, которое является расширением понятий равновесий Нэша и Байеса-Нэша. При определении агрегированного равновесия сравниваются наилучший ответ и средняя стратегия игроков одного класса. Идея заключается в том, чтобы рассматривать рациональность на групповом уровне. Вводится понятие размаха, который служит естественной мерой разнообразия индивидуальных действий внутри агрегированного равновесия. Если все размахи равны нулю, то агрегированное равновесие совпадает с равновесием Нэша.

В лаборатории экспериментальной экономики МФТИ и ВЦ РАН была проведена серия лабораторных экспериментов с однозвенными сетевыми рынками STB с торговым механизмом, предложенным Верноном Смитом [5]. В этих экспериментах сетевой рынок состоит из одного продавца, одного покупателя и одного транспортировщика.

Эксперименты проводились с использованием специального программного обеспечения, разработанного на основе системы Z-Tree. Участниками экспериментов являлись студенты и сотрудники МФТИ, МГИМО, ВЦ РАН.

Проведенный анализ экспериментальных данных показывает, что агрегированное поведение участников гораздо лучше соответствует теоретико-игровым принципам, чем индивидуальное поведение. Это обосновывает целесообразность использования агрегированного равновесия и равновесной траектории для анализа лабораторных сетевых энергетических рынков.

Результаты проведенных экспериментов по лабораторным сетевым рынкам типа STB хорошо согласуются с понятием агрегированного равновесия при положительной величине размаха. Принцип наилучшего ответа на действия остальных проявляется не на индивидуальном, а на групповом уровне.

Планируется более детально изучить теоретические аспекты введенного понятия агрегированного равновесия, а также провести лабораторный анализ сложных сетевых рынков. При этом потребуется преодолеть вычислительные трудности, связанные с поиском наилучших ответов для равновесной траектории.

Работа выполнена при поддержке РФФИ, проекты № 07-01-00605а, № 06-01-08057-офи; гранта Президента РФ по поддержке ведущих научных школ, проект НШ-5379.2006.1; программы «Развитие научного по-

тенциала высшей школы (2006–2008 годы)» Федерального агентства по образованию, проект РНП.2.2.1.1.2467.

Литература

[1] *Меньшиков И. С.* Анализ влияния психофизиологических параметров участников на агрегированное поведение рынка методами экспериментальной экономики // ММРО-13 (в настоящем сборнике). — 2007. — С. 497–499.

[2] *Fischbacher U.* Z-Tree — Zurich Toolbox for Readymade Economic Experiments — Experimenter's Manual // Working Paper № 21, Institute for Empirical Research in Economics, University of Zurich, 1999.

[3] *Меньшиков И. С.* Лекции по теории игр и экономическому моделированию. — М.: МЗ-ПРЕСС, 2006.

[4] *Myerson R.* Game Theory: Analysis of Conflict. — Harvard Univ., Press, 1991.

[5] *McCabe K. A., Rassenti S. J., Smith V. L.* Designing "Smart" Computer-Assisted Markets — An Experimental Auction for Gas Networks // European Journal of Political Economy. — 1989. — Vol. 5, issues 2–3 — Pp. 259–283.

[6] *Голубцов А. А.* Формализация равновесия для случая лабораторных сетевых рынков // Магистерская диссертация, ФУПМ МФТИ, 2006.

Применение метода главных компонент при построении кластерной структуры обучающей выборки молекул.

Григорьева С. С., Кумсков М. И., Захаров А. М.

qsar_msu@mail.ru

Москва, Кафедра Вычислительной математики механико-математического факультета МГУ им.Ломоносова

В работе предложен метод выбора метрик на молекулярных графах на основе главных компонент матрицы «молекула-дескриптор», формирующийся полным перечислением двоек (троек) особых точек на молекулярной поверхности.

Постановка задачи

Задача «структура-свойство» — это задача распознавания образов [1], где объектами являются молекулы, векторное описание которых заранее не задано. Решение этой задачи разбивается на два этапа:

1) этап построения описания обучающей выборки: формируется матрица «молекула-дескриптор» (МД);
2) этап поиска по матрице МД функциональной зависимости.

При представлении пространственной структуры молекул в виде МД матрицы она может получиться очень «широкой», например, в методе CoMFA (Comparative Molecular Field Analysis) [2] или при использовании

структурного символьного спектра молекулярных графов [3]. Для преодоления этой проблемы в методе CoMFA линейная регрессия строится на главных компонентах МД матрицы (метод PLS) [4]. В работе используется аналогичный подход при построении функциональной зависимости в виде деревьев решений. Главные компоненты используются при выборе метрики в алгоритме кластер-анализа, а также при построении собственно линейной регрессии на каждом найденном кластере.

Метод решения

Признаками молекулярного графа являются инварианты различных типов. В основе их построения лежит понятие алфавита примитивов описания графов. Мощность алфавита, формируемого по данной обучающей выборке, зависит от определения отношения эквивалентности на элементах примитивов — особых точках (ОТ). Инвариантом первого уровня является число повторений примитивов в графах обучающей выборки; второго уровня — число повторений пар примитивов, находящихся в графе на данном интервале расстояния; третьего (четвертого) уровня — число повторений троек (четверок) примитивов, расположенных в графе на заданных интервалах расстояний.

МД матрица, построенная описанным выше способом, получается очень «широкой», то есть $M \gg N$, где N — количество молекул обучающей выборки, M — мощность алфавита. Метод главных компонент позволяет формировать существенные для прогноза столбцы-факторы матрицы. За основу базовой модели взята линейная множественная регрессия, которая строится на главных компонентах. В силу неоднородности обрабатываемой выборки, ищем зависимость значения активности от значений дескрипторов в виде дерева решений. Мы разбиваем обучающую выборку на кластеры-классы, внутри которых строится функциональная зависимость. Для выделения классов применяем метод кластерного анализа [5].

Перед применением кластерного анализа выделяем главные компоненты, на которых задаем евклидову метрику и формируем матрицу расстояний между молекулами. На полученной матрице расстояний запускаем кластерный анализ и на каждом из «содержательных» кластеров строим линейную регрессию на «кластерных» главных компонентах. Далее, вычисляем коэффициент корреляции скользящего контроля, позволяющий оценить качество прогностической устойчивости полученного дерева решений. Алгоритм (без построения кластеров) был применен к выборкам амбровых одорантов (низкомолекулярных соединений, обладающих амбровым запахом), состоящих из 50 и 129 молекул. Для каждого соединения выборки формировалось 3D-описание соответствующего молекулярного графа — перечислены вершины графа (атомы) с допол-

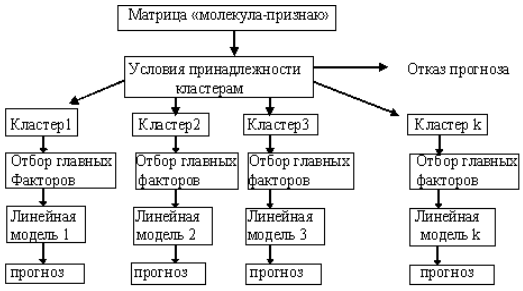


Рис. 1. Схема построения дерева решений (k — число кластеров).

нительными атрибутами: символом химического элемента, трехмерными координатами в ангстремах и электрическим зарядом. Матрица МД содержала 703 дескриптора.

Результаты вычислений следующие: на «большой» выборке на 3, 4 и 5 факторах Q^2 (квадрат коэффициента множественной регрессии на скользящем контроле) получился равным соответственно 0.705, 0.74 и 0.66; на «маленькой» выборке на 3, 4 и 5 факторах Q^2 получился равным соответственно 0.74, 0.76 и 0.74. Таким образом, оптимальным является использование всего четырех факторов.

Работа выполнена при поддержке РФФИ, проект № 07-07-00282.

Литература

[1] *Стьюпер Э., Брюгер У., Джурс П.* Машинный анализ связи химической структуры и биологической активности // М.: Мир, 1982.

[2] *Cramer III R. D., Patterson D. E., Bunce J. D.* Comparative molecular fields analysis (CoMFA) // Effect of shape on binding of steroids to carrier proteins J. Am. Chem. Soc. 110 (1988) 5959-5967. — С. 109–112.

[3] *Кумсков М. И., Смоленский Е. А., Пономарева Л. А., Митюшев Д. Ф., Зефиров Н. С.* Системы структурных дескрипторов для решения задач «структура-свойство» // М.: Доклады Академии Наук, 1994. — С. 336.

[4] *Clark M., Cramer III R. D.,Jones D. M., Patterson D. E., Simeroth P. E.* Comparative Molecular Field Analysis(CoMFA) Toward Its Use with 3D-Structural Databases // Tetrahedron Comput. Methodol. , 1990. — С. 3, 47–59.

[5] *5. Сошникова Л. А., Тамашевич В. Н., Уебе Г., Шефер М.* Многомерный статистический анализ в экономике // Учеб. Пособие для ВУЗов под ред. проф. Тамашевича, М.: ЮНИТИ-ДАНА, 1999.

Поиск комбинированных структур в ДНК-последовательностях

Гусев В. Д., Мирошниченко Л. А.
luba@math.nsc.ru
Новосибирск, Институт математики СО РАН

Элементарными структурами в ДНК-последовательностях будем считать повторы следующих трех типов: *прямые* (... *gaactc* ... *gaactc* ...), *симметричные* (... *gaactc* ... *ctcaag* ...) и *комплементарные симметричные* (... *gaactc* ... *gagttc* ...). В последнем случае речь идет о симметричных повторах с точностью до переименования элементов алфавита в соответствии с известным в молекулярной биологии отношением комплементарности: $a \leftrightarrow t$, $c \leftrightarrow g$. *Комбинированными* назовем структуры, состоящие из двух разнотипных повторов (прямой плюс симметричный, прямой плюс комплементарный симметричный, и т. п.) с ограничениями снизу на длины повторяющихся цепочек и сверху — на те же длины и расстояния между соседними элементами структуры. Ограничения снизу нужны для отсеивания случайных («шумовых») структур, а сверху — для обеспечения компактности структуры. Порядок чередования цепочек, образующих повторы разных типов — произвольный, возможны наложения и совпадения цепочек, относящихся к разным повторам.

Целью работы является реализация и исследование алгоритма выявления комбинированных структур в ДНК-последовательностях. Алгоритм может быть использован для обнаружения возможных регуляторных областей и «горячих точек» генома, выявления потенциально многофункциональных фрагментов с наложением структур, поиска нестандартных (из-за учета симметрии и переименования элементов алфавита) образцов с двумя переменными [1].

Описание алгоритма

Для поиска комбинированных структур мы используем разработанный нами ранее аппарат построения сложностного профиля символьной последовательности, опирающийся на факторизацию Лемпеля и Зива [2], но с расширенным спектром операций копирования. Применительно к нашему случаю мы добавляем операции симметричного копирования, как с переименованием элементов алфавита, так и без [3]. Поясним схему алгоритма на примере выявления комбинированной структуры, составленной из прямого и симметричного повторов, не совпадающих тождественно друг с другом.

Шаг 1. Задаем нижнюю и верхнюю границу длин повторяющихся фрагментов (соответственно, r и R), а также максимально допустимое

расстояние между соседними компонентами структуры. Эти параметры определяют максимально возможный размер структуры $W = 4R + 3d$.

Шаг 2. Используя операцию симметричного копирования, вычисляем сложностной профиль текста с помощью скользящего окна размера W [3]. Выделяем окна, содержащие хотя бы один *нерасширяемый* компонент (фактор) с длиной $r \leqslant l \leqslant R$ и указателем копирования, равным 1 (первый символ окна). Тем самым фиксируется новая (не рассматривавшаяся ранее) симметрия, которая при наличии прямых повторов в ближайшей ее окрестности может образовать комбинированную структуру. Нетрудно показать, что прямые повторы следует искать в выделенном окне, расширенном влево на $(2R + 2d)$ символов.

Шаг 3. Построив L-граммное дерево (trie-структура) для расширенного окна [3] , фиксируем листья, соответствующие двум (или большему числу) одинаковых, но позиционно разнесенных цепочек длины L. Пары, которые могут быть расширены по тексту до максимально возможной длины l $(r \leqslant l \leqslant R)$ при сохранении свойства идентичности, являются искомыми прямыми повторами.

Шаг 4. Совмещаем (позиционно) симметрию с каждым из найденных прямых повторов. Если в образовавшейся 4-компонентной структуре расстояния между соседними компонентами не превышают d, фиксируем наличие комбинированной структуры.

Если параметр r логарифмическим образом зависит от величины W, т. е. соответствует средней длине максимального случайного повтора для окна анализа, трудоемкость алгоритма в среднем имеет порядок NL, где N — длина текста, L — длина цепочек, представленных в поисковом (L-граммном) дереве (в наших экспериментах L выбиралось близким к r).

Апробация алгоритма

Алгоритм проверялся на хорошо изученном геноме фага λ (длина $N = 48502$) и подборке кодирующих участков различных генов человека (всего 1469 последовательностей). При значениях параметров $r = 7$, $R = 20$, $d = 13$ в λ выделено 16 комбинированных структур типа «прямой повтор + симметричный комплементарный», в подборке генов — 1785 структур. Приведем примеры наиболее интересных структур.

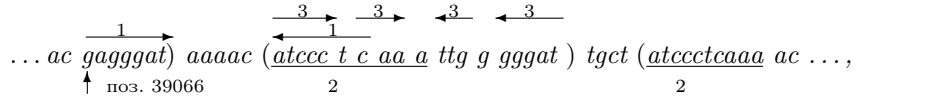
Пример 1 (фаг λ). В некодирующей области выделена структура

$$\ldots at\ \underline{gacAaaaa}\ \overrightarrow{attagcgcaag}\ aa\ \underline{gacaaaaa}\ tcac\ \overleftarrow{cttgcgctaat}\ gc \ldots,$$

точка окончания транскрипции (поз. 27538)

содержащая прямой повтор (подчеркнут) и симметричный комплементарный повтор (указан стрелками сверху). Эта структура содержит точку окончания транскрипции по комплементарной цепи (А) и терминатор транскрипции (справа от А).

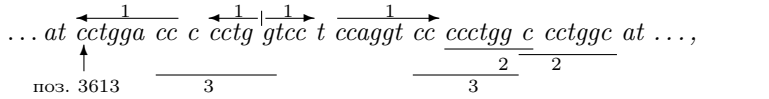
Пример 2 (фаг λ). В кодирующей области выделена структура



построенная на несовершенных тандемных повторах длины 19 (выделены скобками) с точно совпадающими ядрами длины 10 (подчеркнуты) и симметричными комплементарными повторами на стыке (см. 1) и внутри периода (см. 3). Выявленная структура лежит в области начала репликации фага λ. Представляет интерес наложение разных элементарных структур $(1, 2, 3)$ в центральной части фрагмента.

Структур типа «прямой повтор + симметричный» для обоих типов данных выявлено больше.

Пример 3 (ген «Collagen II α 1»). Выделена структура



содержащая несовершенную симметрию (1) с одной заменой и два прямых повтора (см. 2, 3). Структура возникает на периодичностях вида $(ccXggXccX)^n$, где $n = 4$, а X — произвольный нуклеотид. На аминокислотном уровне ей также соответствует симметричная структура $(PGP)^4$.

Заключение

Предложен эффективный алгоритм выявления в ДНК-последовательностях компактных комбинированных структур, состоящих из неслучайных повторов разного типа. Эксперименты на текстах с известной разметкой демонстрируют наличие таких структур в регуляторных областях и «горячих точках» генома. Близкими в идейном плане являются работы по отысканию фрагментов ДНК с аномально низкой сложностью [3] и образцов с двумя переменными [1]. Наше продвижение относительно [3] состоит в выявлении альтернативных накладывающихся друг на друга структур в зоне аномальной сложности. Продвижение по отношению к [1] состоит в расширении трактовки повтора и варьировании понятия образца.

Работа выполнена при поддержке РФФИ, проект № 06-06-80467.

Литература

[1] *Neraud J.* Algorithms for detecting morphic images of a word // Information and Computation. — 1995. — V. 120. — Pp. 126–148.

[2] *Lempel A., Ziv J.* On the complexity of finite sequences // IEEE Trans. Inform. Theory. — 1976. — V. IT–22, № 1. — Pp. 75–81.

[3] *Гусев В. Д., Немытикова Л. А.* Учет проявлений повторности, симметрии и изоморфизма в символьных последовательностях // Вычислительные системы. — Вып. 167. — Новосибирск, 2001. С. 11–33.

Анализ кластерных конфигураций в одной проблеме фильтрации спама

Дьяконов А. Г.

djakonov@mail.ru

Москва, ВМиК, МГУ им. М. В. Ломоносова

В докладе представлен метод настройки спам-фильтра, который был разработан для участия в соревновании «ECML/PKDD 2006 Challenge» [1] и занял там четвертое место. Конкурсная задача имела две особенности:

1. Для анализа предоставлена только частотная информация (какое слово сколько раз встречается в письме), нет структурной информации (в какой последовательности слова входят), нет контекстной информации (описание заголовка письма, входные данные и т.д.).
2. Обучение происходит на данных из спамовых ловушек (spam traps), а контроль на электронных письмах, которые приходят реальным пользователям. Таким образом, нет гарантии сходства распределений писем в обучении и контроле.

Постановка задачи

Заданы матрицы $S = \|s_{ij}\|_{N_s \times T}$, $M = \|m_{ij}\|_{N_m \times T}$, $U = \|u_{ij}\|_{N_u \times T}$, N_s — число спамовых писем в ловушке, N_m — число нормальных писем в ловушке, N_u — число всех писем в ящике пользователя (эти значения достигают нескольких тысяч), T — число знакомых слов (достигает нескольких сотен тысяч), $s_{ij} = p$ тогда и только тогда, когда j-е слово входит в i-е спамовое письмо p раз (аналогично, элементы m_{ij} описывают вхождения в нормальные письма, u_{ij} — в контрольные). Задача состоит в построении алгоритма, который классифицирует контрольные письма, описанные в U («спам» или «норма»).

Стандартные алгоритмы показывают достаточно плохое качество классификации, и это не связано с эффектом переобучения [2]. Проблема заключается в выборе хорошего пространства признаков. Напри-

мер, в простейшем пространстве ⟨длина письма, число различных слов⟩ при переходе к контролю классы «меняются местами»: спамовые письма в ящиках пользователя имеют значения этих признаков такими же, как нормальные письма в спам-ловушках, и наоборот.

Формирование пространства признаков

Пусть, для простоты, письмо (строка матрицы U) содержит слова с идентификаторами $1,\dots,r$ в количестве $c_1,\dots,c_r$ (соответственно). Рассмотрим матрицу S (для матрицы M все делается аналогично). Рассмотрим преобразование $F(H(G(S)))$, где

$$G(\|s_{ij}\|_{N_s\times r}) = \|g(s_{ij}, s_{i1},\dots,s_{iT})\|,$$

$$H\big(\|g_{ij}\|\big) = \big(h(g_{11},\dots,g_{1N_s}),\dots,h(g_{r1},\dots,g_{rN_s})\big).$$

Функция G осуществляет построчные преобразования матрицы и ее «обрезание» (оставляет столбцы, соответствующие словам письма). Примеры построчных преобразований: нормировка (деление числа вхождения на число слов в письме), «обезличивание» (замена ненулевых чисел единицами). Функция H «схлопывает» матрицу, получая вектор (например, суммирует элементы по вертикали или находит максимальный элемент в каждом столбце). Функция h должна быть монотонной. Функция F — монотонная функция от строки, например сумма элементов или скалярное произведение на вектор $(c_1,\dots,c_r)$.

Признаки ищем в виде (где f — значение на рассматриваемом письме)

$$f = c_1F_1^s(H_1^s(G_1^s(S))) - c_2F_2^m(H_2^m(G_2^m(M))),\ c_1,c_2\in\mathbb{R}^+$$

с помощью генетического алгоритма. Определив наиболее удачные классы преобразований F, H, G, часто удается провести полный перебор в этих классах, или даже найти аналитические выражения для параметров.

Качество пространства признаков

Качество пространства $[f_1,\dots,f_n]$ оценивается следующим образом. На множестве писем в ловушке осуществляется кластеризация каким-то фиксированным методом. Этим же методом осуществляем кластеризацию на контроле (для писем пользователя). Предполагаем, что в хороших признаковых пространствах

1) в кластеры входят объекты только одного класса («почти» одного);
2) кластерные конфигурации ловушки и ящика (ящиков) совмещаются друг с другом несложным преобразованием (параллельный перенос, поворот);

3) это преобразование легко определяется по помеченной кластерной конфигурации ловушки (известна классификация) и непомеченной кластерной конфигурации ящика.

На практике чаще всего требуется устойчивость относительно параллельных переносов [3]. Хотя, например, в некоторых задачах из области BCI [4] при применении аналогичной техники наблюдается поворот кластерной конфигурации относительно центра координат.

Формализация условий 1–3 и является «ядром метода». Для решения задачи предлагается оценивать устойчивость кластерных концигураций, и выбирать пространства, в котором они наиболее устойчивы.

Работа выполнена при поддержке РФФИ, проект №05-01-00332, Минобрнауки РФ, гранта Президента РФ, МК-533.2007.9.

Литература

[1] `www.ecmlpkdd2006.org/challenge.html`.

[2] *Воронцов К. В.* Обзор современных исследований по проблеме качества обучения алгоритмов // Таврический вестник информатики и математики. — Симф.: 2004. — №1.— С. 5–24.

[3] *Дьяконов А. Г.* Об одном подходе к решению задач из области BCI // Докл. XII всеросс. конф. ММРО-12, М.: МАКС Пресс, 2005. — С. 95–97.

[4] *Jose del R. Millan* Brain-computer interfaces // M. A. Arbib (ed.), Handbook of Brain Theory and Neural Networks, 2nd ed., Cambridge: MIT Press, 2002.

О методах промежуточного контроля в сложной системе обнаружения и распознавания лиц

Ковальчук А. В., Беллюстин Н. С., Тельных А. А., Яхно В. Г.

`iandx@mail.ru, nbell@awp.nnov.ru`

Нижний Новгород, Институт прикладной физики РАН,
ФГНУ Научно-исследовательский радиофизический институт

В работе показан пример организации внутреннего промежуточного телеметрического контроля в системе обнаружения и распознавания лиц, необходимый для точной настройки системы и анализа совершаемой ею ошибок распознавания. Совместно с другими мерами система промежуточного контроля должна способствовать повышению надежности работы системы автоматического распознавания лиц.

Автоматическое распознавание лица человека на электронных видеоизображениях является ключевым вопросом для многочисленных приложений, оно необходимо для борьбы с терроризмом и преступностью, для общего контроля перемещения людей, для идентификации личности при банковских операциях в электронных сетях, и для целого ряда аналогич-

ных задач. Многие из этих приложений связаны с высокой ответственностью, и цена ошибочной идентификации может оказаться чрезвычайно высока.

Жесткие требования к системам распознавания заставляют все более тщательно проводить их настройку, включающую анализ ошибочных решений, принятых системой. При этом важно использовать уже наработанный опыт создания сложных технических систем с высокой ответственностью — авиационных и других транспортных систем, ядерных объектов, и т. д. В таких системах повышенной надежности катастрофические ошибки обычно возникают в результате непредвиденного взаимодействия различных блоков системы. Для разбора ошибок необходима запись «телеметрической» информации, анализ которой позволяет выявить особенности «патологических» режимов системы и затем внести в систему изменения, повышающие ее надежность.

Ключевым элементом системы автоматического распознавания лиц является алгоритм обнаружения лица, который сканирует анализируемое изображение, осуществляя анализ очень большого числа его фрагментов и формируя по каждому из них свое заключение — является ли этот фрагмент лицом человека, или нет. При этом анализируемый фрагмент масштабированием приводится к стандартному небольшому размеру — 32×32 пиксела, например. На Рис. 1 показан пример такого изображения лица качества невысокого, однако вполне достаточного для уверенного обнаружения. Принимающий решение детектор лиц предварительно обучается на достаточном количестве примеров лиц и не лиц, при этом обучающий алгоритм находит наиболее важные участки этих изображений, по которым лица и не лица наиболее уверенно разделяются друг от друга. Процедура нахождения областей-признаков организована таким образом, чтобы из сотен тысяч возможных признаков (рецептивные поля 14-ти типов с разными масштабами) можно было отобрать всего несколько сотен наиболее значимых областей-признаков [1,2]. Отобранные признаки сначала соединятся параллельно в «каскады», фильтрующие поток изображений, а затем несколько таких фильтров-каскадов включаются последовательно и формируют «детектор лиц», эффективно выделяющий лица из фоновых фрагментов на большом изображении. Для реализованного детектора на базе данных, содержащей около 3000 изображений 200 человек и около 2 000 000 примеров не лиц, были получены ошибки FAR $\approx 0.002\%$ и FRR $\approx 5\%$.

На Рис. 2 и Рис. 3 эллипсами различного положения, эксцентриситета и яркости показаны примеры контрольных изображений, визуализирующие местоположения признаковых областей на стандартизованном по размеру изображении, которые оказались наиболее эффективными в

Рис. 1. Пример изображения стандартизованного размера 32×32.

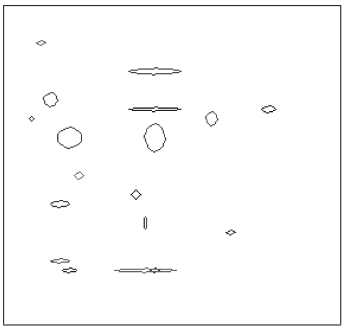


Рис. 2. Геометрическое изображение ключевых разделяющих признаков I типа по 5 первым каскадам детектора лиц.

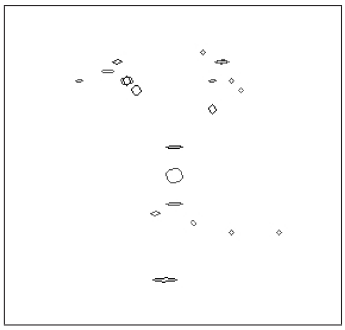


Рис. 3. Геометрическое изображение ключевых разделяющих признаков I типа по 6–12-му каскадам детектора лиц.

разделении на лица и не лица на первых пяти каскадах — Рис. 2, и на каскадах с 6-го по 12-ый — Рис. 3. Яркость эллипса показывает сумму рейтинговых весов данного признака при его голосовании в каскадах.

Визуализация расположения наиболее важных признаковых рецептивных полей позволяет анализировать структуру модельного описания для данного детектора. Видно, что признаки достаточно хорошо соответствуют тем ключевым областям изображения лица, на которые обычно обращает внимание человек при анализе и распознавании лица на изображениях.

Литература

[1] *Яхно В. Г., Нуйдель И. В., Иванов А. Е., Беллюстин Н. С., и др.* Исследование динамических режимов нейроноподобных систем // Информационные технологии и вычислительные системы. — 2004. — № 1. — С. 126–148

[2] *Беллюстин Н. С., Разумов В. А., Тельных А. А.* Настройка и тестирование фрагментов системы анализа лица человека на видеоизображениях // VIII Всероссийская научно-техническая конференция «Нейроинформатика-2006», Сб. трудов, часть 1. — МИФИ, Москва, 2006, — С. 157–164.

Разработка математических методов формализации профессионального знания врача

Котов Ю. Б.
kotsem@voxnet.ru
Москва, ИПМ РАН

Статья посвящена методам прикладного анализа данных, используемых врачом для принятия решений. Распространен традиционный подход: специалист дает полный набор критериев для исчерпывающего решения задачи, математики вырабатывают процедуру принятия решений по этому набору. В задачах постановки диагноза конкретному больному даже опытный врач обычно использует некоторое количество неотрефлексированных (не осознанных) критериев. Поэтому первая цель математика в таких задачах состоит в выявлении и приведении в строгую логическую форму (формализации) этих критериев. Вторая цель: обеспечить запись и обработку суждений врача в форме, нечувствительной к наличию пробелов в сведениях при сохранении логической корректности на всем протяжении преобразований. Основные требования к критериям (непротиворечивость, эффективность, универсальность и отсутствие синонимов) врач проверить не может. Работа предполагает совместную работу врача и математика.

В процессе профессиональной работы врачу приходится в ряде случаев добавлять новые критерии к рабочему набору (для особо трудных больных), что предполагает использование открытых наборов критериев. Эта проблема носит фундаментальный характер для многих видов познавательной деятельности человека.

Врач в процессе лечения больного часто принимает решения на основе недостаточной информации, когда статистические методы малоэффективны. Математик, работающий с медицинскими данными, вынужден создавать новые, нестатистические методы описания и анализа структур, включающих разнородные данные для моделирования суждений врача в условиях неполной информации.

Логические симптомы. Сведения, используемые врачом в диагностической практике, часто включают тексты, числа и изображения, несопоставимые друг с другом. Врач делает собственные предположения относительно их роли в диагнозе данному больному, которые можно мо-

делировать логическими переменными. Обычная булева логика порождает в задачах диагностики излишне сложные конструкции из-за недоступности некоторых сведений.

Удобной оказалась трехзначная логика [1], разработанная Яном Лукасевичем. Например, одна формула трехзначной логики, использующая 10 логических переменных со значениями (`ДА`, `НЕТ`, `НЕИЗВЕСТНО`) при допустимом количестве неизвестных значений не более 3 заменяет около 800 обычных булевых формул с условиями применимости. Трехзначные логические переменные со значениями (`ДА`, `НЕТ`, `НЕИЗВЕСТНО`) будем называть *логическими симптомами*. К классическому набору операций трехзначной логики была добавлена операция наследования, сохраняющая совпадающие значения. Разработаны программы вычислений с трехзначными логическими переменными в диалоговом режиме.

Объект (больной или отдельное обследование больного) соответствует *симптомному вектору*, каждая координата которого есть логический симптом. Для любых двух симптомных векторов определим меры их близости: *сходство* и *неотличимость* [2], основанные на подсчете совпадающих и неизвестных значений координат. Обе величины достигают 1 у пары полностью совпадающих векторов и равны 0 у полностью несовпадающих. Численные значения этих величин различны при частичном совпадении.

Классы объектов зададим их *масками*, т. е. симптомными векторами с максимальной неотличимостью от объектов «своего» класса и с минимальной — от объектов всех остальных классов. Разработаны алгоритмы генерации масок классов по обучающим выборкам [2].

Задачу классификации объектов удобно решать в несколько этапов. На первом методом *диагностических игр*, предложенным И. М. Гельфандом [3], определяем набор логических симптомов, используемых врачом. Затем, используя данные об известных больных (класс обучения) в пространстве этих симптомов, строим маски классов. И, наконец, используя маски классов, вычисляем неотличимость каждого нового больного от всех масок. Максимальная величина ее указывает на возможную принадлежность больного определенному классу. Простейший метод классификации — пороговый — уже дает удовлетворительные результаты. Ошибки и двусмысленности классификации отдельных больных прямо указывают на недостающие данные или неустойчивые суждения врача, т. е. на содержательные «болевые точки» процесса диагностики. Окончательная классификация проводится по результату вычисления значения критерия.

Проиллюстрируем возможности этого подхода примерами решенных задач диагностики, результаты которых внедрены в практику.

1. После кесарева сечения на теле матки остаются рубцы от хирургических разрезов. Повторные роды несут опасность разрыва матки в случае несостоятельного рубца. Механические свойства миометрия у конкретной пациентки неизвестны и недоступны для исследования. Требовалось дать прогноз опасности для конкретной беременности. Мы прогнозировали не механические процессы, а решение опытного врача на трех последовательных этапах (предварительное обследование, решение накануне плановых родов, экстренное решение в родах). Использование технологии логических симптомов позволило резко сократить объем собираемой информации (8 симптомов вместо 450 переменных, названных врачом).

2. Дети, родившиеся живыми от матерей с сахарным диабетом, поразному адаптируются к самостоятельной жизни. Особенно велика роль раннего периода жизни (первые 7 дней). Для своевременного начала лечения нужен диагноз уже в первые часы жизни. На первом этапе несколько врачей единогласно и независимо указали крайних (тяжелых и благополучных) новорожденных. Неотличимость больных от маски благополучного класса может служить оценкой тяжести новорожденного.

3. При поликлиническом осмотре женщин пожилого возраста для выявления ранних опасных изменений шейки матки испольуют три метода: ультразвуковое исследование с измерением скорости кровотока методом Доплера, расширенную кольпоскопию и цитологию мазков. Наиболее эффективен (и дорог) первый из них. Задача состояла в оценке эффективности более простых методов. Логические симптомы, связанные с методом кольпоскопии, позволяют решить задачу диагностики. Сравнение индивидуальных наборов логических симптомов каждой больной с двумя масками позволило выделить всех пациентов с неблагоприятным прогнозом.

Итог. Разработан формальный язык логических симптомов, позволяющий моделировать суждения врача, выдвигаемые в процессе постановки диагноза конкретному больному. Диагностическое правило допускает формальную запись (например, в виде ДНФ) достаточно сложных решений. Разработан набор диалоговых программ для обслуживания операций над данными.

Работа выполнена при поддержке РФФИ, проект № 04-01-00434.

Литература

[1] *Карпенко А. С.* Логики Лукасевича и простые числа.— М: Наука, 2000. — 319 с.

[2] *Котов Ю. Б.* Метод логических симптомов в моделировании профессиональных суждений врача // Информационные технологии и вычислительные системы. — 2005. — № 2. — С. 29-42.

[3] *Гельфанд И. М., Розенфельд Б. И., Шифрин М. А.* Очерки о совместной работе математиков и врачей— М: Наука, 1989. — 272 с.

Применение логических алгоритмов классификации в задачах кредитного скоринга и управления риском кредитного портфеля банка

Кочедыков Д. А., Ивахненко А. А., Воронцов К. В.
kochedykov@forecsys.ru, andrej_iv@mail.ru, vokov@forecsys.ru
Москва, ЗАО «Форексис»

Принятие решений о выдаче кредитов физическим лицам — это стандартная задача классификации, в которой объектами являются клиенты, а признаки соответствуют полям анкеты, заполняемой клиентом при подаче заявки на выдачу кредита. В простейшем случае клиенты разделяются на два класса — «хорошие» (good) и «плохие» (bad).

Задача имеет следующие особенности: признаки разнотипны; в данных могут присутствовать пропуски и ошибки; объём обучающей выборки может быть крайне мал, в частности, при создании новых кредитных продуктов, при выходе на новые рынки или изменении экономической ситуации. Недоверие кредитных экспертов к «чёрным ящикам» влечёт ещё и требование интерпретируемости: алгоритм классификации должен быть прост и понятен, допускать осмысленную модификацию «вручную», а выдаваемые им решения — иметь логичные объяснения [1]. Этим требованиям удовлетворяют логические классификаторы, в частности, решающий список и взвешенное голосование конъюнкций [2].

Ещё одна важная особенность задачи заключается в том, что наряду с классификацией клиента алгоритм должен выдавать оценку вероятности дефолта (probability of default, PD), т. е. вероятности того, что клиент окажется «плохим» и не вернёт кредит или его часть. Оценки PD заёмщиков необходимы для анализа риска кредитного портфеля банка.

В мировой практике *кредитного скоринга* широко применяется логистическая регрессия, в которой оценки PD получаются естественным образом. Однако, по сравнению с логическими алгоритмами, она хуже интерпретируема, требует заметно бóльших объёмов обучающей выборки и основана на труднопроверяемых вероятностных предположениях.

В данной работе предлагается метод оценивания PD клиента для логических алгоритмов, основанный на понятии переобученности.

Переобученность логических закономерностей

Пусть X — множество объектов (допустимых описаний клиентов), $Y = \{-1, +1\}$ — классы «плохой», «хороший». Будем говорить, что предикат $\varphi\colon X \to \{0, 1\}$ *выделяет* объект x, если $\varphi(x) = 1$. Обозначим через

$b(\varphi, U)$ и $g(\varphi, U)$ число объектов классов -1 и $+1$ соответственно, выделяемых предикатом φ из выборки $U \subset X$. Долю объектов класса -1, выделяемых предикатом φ из U, обозначим $\beta(\varphi, U) = \frac{b(\varphi,U)}{b(\varphi,U)+g(\varphi,U)}$.

Закономерность класса y — это предикат $\varphi(x)$, выделяющий достаточно много объектов класса y и достаточно мало объектов класса $-y$. Будем рассматривать логические алгоритмы классификации вида

$$a(x) = \arg\max_{y \in Y} \sum_{t=1}^{T_y} w_y^t \varphi_y^t(x), \tag{1}$$

где φ_y^t — закономерности, w_y^t — веса закономерностей, T_y — число закономерностей класса y. В данном исследовании закономерности φ_y^t строились в виде конъюнкций элементарных (однопризнаковых) пороговых предикатов [2]. Такие закономерности называются *правилами* (rules).

В общем случае для обучения алгоритма $a(x)$ по выборке U применяется некоторый метод поиска закономерностей $\mu\colon U \mapsto \{\varphi_y^t\}_{y\in Y}^{t=1,T_y}$. Закономерности отбираются по критериям $\min_\varphi \beta(\varphi, U)$ для класса $+1$ и $\max_\varphi \beta(\varphi, U)$ для класса -1. В результате оптимизации величина $\beta(\varphi, U)$ оказывается смещённой оценкой PD — заниженной для закономерностей класса $+1$ и завышенной для закономерностей класса -1. Несмещённую оценку $\mathrm{PD} = \beta(\varphi, V)$ можно было бы получить по независимой контрольной выборке V. Однако для надёжного оценивания требуются сотни контрольных объектов, а в условиях малого объёма данных всю выборку хотелось бы использовать как обучающую.

Возникает задача: спрогнозировать $\beta(\varphi, V)$, имея только информацию о закономерности φ, полученную на этапе обучения по выборке U.

Переобученностью закономерности $\varphi \in \mu(U)$ класса y будем называть величину $\delta(\varphi, U, V) = y\beta(\varphi, V) - y\beta(\varphi, U)$.

Эмпирическая методика оценивания переобученности

Заданная выборка $X^L \subset X$ длины L разбивается N способами на обучающую подвыборку длины ℓ и контрольную длины k, $X^L = X_n^\ell \cup X_n^k$, $L = \ell + k$, где индекс разбиения n пробегает значения от 1 до N.

В данной работе использовалась стандартная методика разбиения $t\times q$-fold cross-validation [3]: выборка X^L разбивалась t раз случайным образом на q блоков примерно равной длины и с равными долями классов, и каждый блок поочерёдно становился контрольной выборкой. Таким образом, $N = tq$ и $k \approx \frac{L}{q}$ с точностью до округления.

Пусть $F_n(\varphi) = F(\varphi, X_n^\ell, X_n^k)$ и $Z_n(\varphi) = Z(\varphi, X_n^\ell)$ — две числовые характеристики закономерностей $\varphi \in \mu(X_n^\ell)$. Зависимость F от Z будем

оценивать *среднесглаженным* значением характеристики F в точке z:

$$F(z) = \frac{\sum_{n=1}^{N} \sum_{\varphi \in \mu(X_n^\ell)} \big[|Z_n(\varphi) - z| \leqslant \varepsilon\big] F_n(\varphi)}{\sum_{n=1}^{N} \sum_{\varphi \in \mu(X_n^\ell)} \big[|Z_n(\varphi) - z| \leqslant \varepsilon\big]},$$

где параметр ε играет роль ширины окна сглаживания. В эксперименте значение ε полагалось равным 2% от размаха распределения Z.

В роли характеристики Z будем рассматривать:

E: число ошибок на обучении, равное $b(\varphi, X_n^\ell)$ для закономерностей класса $+1$ и $g(\varphi, X_n^\ell)$ для закономерностей класса -1;

C: число объектов, выделяемых на обучении, равное $b(\varphi, X_n^\ell) + g(\varphi, X_n^\ell)$;

K: число элементарных предикатов в конъюнкции φ;

I: информативность (энтропию) закономерности φ на выборке X_n^ℓ;

R: рейтинг закономерности φ (номер в порядке убывания информативности) в списке всех предикатов, найденных и оцененных в процессе поиска закономерностей методом μ.

Для каждой характеристики Z из $\{E, C, K, I, R\}$ строится график зависимости среднесглаженной переобученности $F_n(\varphi) = \delta(\varphi, X_n^\ell, X_n^k)$ от Z (на Рис. 1 кривая с заштрихованной областью под ней). На графиках также выводятся среднесглаженные: 90%-й доверительный интервал переобученности (тонкие кривые), частота ошибок на обучении (нижняя жирная кривая) и на контроле (верхняя жирная кривая), а также число правил с данным значением характеристики Z (нижний график).

Эксперименты на 7 задачах из репозитория UCI (две из которых, `crx` и `german` — задачи оценки кредитоспособности), показали, что переобученность правил зависит от характеристик $\{E, C, K, I, R\}$ довольно сложным образом, и в разных задачах по-разному. Анализ графиков переобученности позволяет настраивать поиск закономерностей под конкретную задачу. Некоторые характерные примеры показаны на Рис. 1.

(C): закономерности с малым C переобучены сильнее.

(I): при некоторых «особых» значениях информативности переобученность падает до нуля, что может сопровождаться всплеском числа найденных закономерностей.

(R): оптимизация улучшает качество закономерностей как на обучении, так и на контроле. В таких случаях можно усиливать оптимизацию, расширяя пространство поиска. Но в некоторых случаях зависимость оказывается противоположной, что свидетельствует о переобучении, и тогда пространство поиска следует сокращать.

Было также обнаружено, что число K элементарных предикатов в конъюнкциях практически не влияет на переобученность.

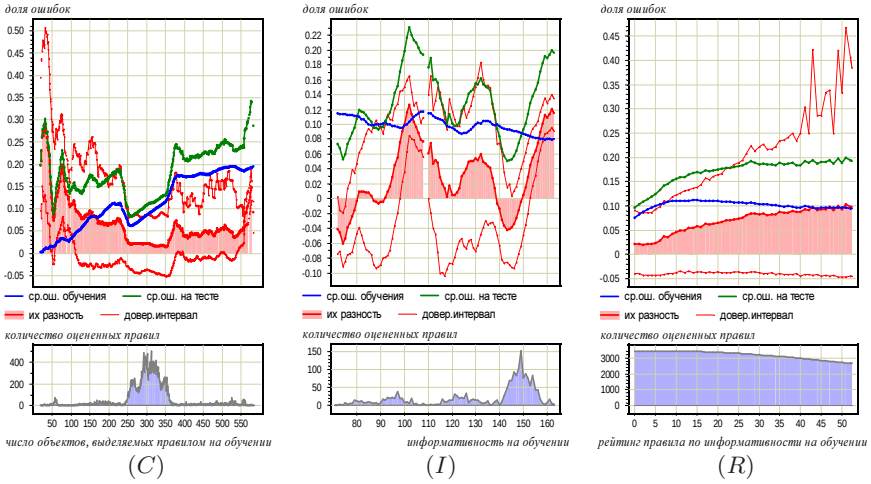


Рис. 1. Зависимость среднесглаженной переобученности правил от: (C) числа выделяемых объектов, задача `german`; (I) информативности (энтропии) правил, задача `crx`; (R) рейтинга правил, задача `german`.

Методика оценивания $\mathrm{PD}(x)$ для произвольного заёмщика x

1. По данным скользящего контроля методами непараметрической регрессии оценивается зависимость переобученности $\delta(E, C, K, I)$ от характеристик E, C, K, I, вычисленных по обучающей выборке.
2. На этапе обучения алгоритм $a(x)$ строится по всей выборке X^L.
3. Для каждой закономерности φ_y^t из (1) вычисляются характеристики E, C, K, I и по ним — оценка переобученности $\delta(\varphi_y^t) = \delta(E, C, K, I)$. При этом характеристики E, C, I приводятся к той длине обучающей выборки ℓ, при которой оценивалась регрессионная зависимость δ.
4. На этапе классификации объекта x оценка PD усредняется по всем закономерностям, выделяющим данный объект:

$$\mathrm{PD}(x) = \frac{\sum_{y,t} w_y^t \cdot \left[\varphi_y^t(x) = 1\right] \cdot \left(\beta(\varphi_y^t, X^L) + \delta(\varphi_y^t)\right)}{\sum_{y,t} w_y^t \cdot \left[\varphi_y^t(x) = 1\right]}.$$

Описанные методы реализованы в системе анализа кредитных рисков и поддержки принятия кредитных решений Forecsys Scoring Solution [4].

Литература

[1] *Соложенцев Е. Д., Степанова Н. В., Карасев В. В.* Прозрачность методик оценки кредитных рисков и рейтингов. — С.-Пб. ун-т, 2005. — 195 с.

[2] *Кочедыков Д. А., Ивахненко А. А., Воронцов К. В.* Система кредитного скоринга на основе логических алгоритмов классификации // ММРО-12, Москва: Макс Пресс, 2005. — С. 349–353.

[3] *Webb G. I.* MultiBoosting: A technique for combining boosting and wagging // Machine Learning. — 2000. — Vol. 40, No. 2. — Pp. 159–196.

[4] `http://www.forecsys.ru/creditr.php` — Скоринг, анализ кредитных рисков и поддержка принятия кредитных решений. — 2005–2007.

Анализ клиентских сред: выявление скрытых профилей и оценивание сходства клиентов и ресурсов

Лексин В. А., Воронцов К. В.
`vleksin@mail.ru, voron@ccas.ru`
Москва, ЗАО «Форексис», МФТИ

Клиентская среда — это совокупность клиентов, регулярно пользующихся фиксированным набором ресурсов (услуг, товаров, сервисов). Клиентскими средами обладают торговые сети, операторы связи, интернет-магазины, поисковые машины, электронные библиотеки, эмитенты пластиковых карт, и т. д.

Анализ клиентских сред (АКС) — это технология обработки исходных данных о действиях клиентов, направленная на решение таких задач, как персонализация предложений клиентам, поиск схожих ресурсов, каталогизация ресурсов, сегментация клиентской базы, формирование клиентских сообществ, выявление нетипичного поведения клиентов, и т. д. Технология АКС основана на вычислении оценок сходства между ресурсами и между клиентами согласно *принципу согласованности*: «ресурсы схожи, если ими пользуются схожие клиенты; в то же время, клиенты схожи, если они пользуются схожими ресурсами» [1].

В данной работе предлагается новый алгоритм АКС, основанный на выявлении интересов (скрытых профилей) клиентов и ресурсов. Алгоритм напоминает генеративные (generative model based) методы коллаборативной фильтрации (collaborative filtering) [2], но отличается от них применением принципа согласованности.

Восстановление профилей клиентов и ресурсов

Пусть заданы: U — множество клиентов, R — множество ресурсов, $D = (u_i, r_i)_{i=1}^{\ell} \subset U \times R$ — выборка (протокол) действий клиентов. Пара (u_i, r_i) означает, что клиент u_i воспользовался ресурсом r_i в момент времени i. Задано множество возможных интересов или тем T. Допустим, что каждый клиент $u \in U$ интересуется темой $t \in T$ с вероятностью $p_{tu} = p(t \,|\, u)$, и каждый ресурс $r \in R$ способен удовлетворить интерес t с вероятностью $q_{tr} = q(t \,|\, r)$. Здесь и далее все вероятности, относящи-

еся к ресурсам, обозначаются буквой q. Задача состоит в том, чтобы по протоколу D восстановить неизвестные *скрытые профили клиентов* $\{p_{tu}\colon t \in T\},\ u \in U$ и *скрытые профили ресурсов* $\{q_{tr}\colon t \in T\},\ r \in R$.

Вероятность $p(u,r)$ того, что клиент u выберет ресурс r, выписывается по формуле полной вероятности, причём сразу двумя способами:

$$p(u,r) = \textstyle\sum_{t\in T} p_u p_{tu} q(r \mid t,u) = \tag{1}$$
$$= \textstyle\sum_{t\in T} q_r q_{tr} p(u \mid t,r), \tag{2}$$

где $p_u = p(u)$ и $q_r = q(r)$ — априорные вероятности появления клиента u и ресурса r в записи протокола. Эти вероятности легко оцениваются по протоколу D как соответствующие частоты. Апостериорные вероятности выписываются по формуле Байеса:

$$q(r \mid t,u) = q(r \mid t) = q_{tr} q_r \,/\, \textstyle\sum_{r'\in R} q_{tr'} q_{r'};$$
$$p(u \mid t,r) = p(u \mid t) = p_{tu} p_u \,/\, \textstyle\sum_{u'\in U} p_{tu'} p_{u'}.$$

Подстановка апостериорных вероятностей в (1) и (2) позволяет выразить вероятность $p(u,r)$ через неизвестные профили $\{p_{tu}\}$ и $\{q_{tr}\}$. Предположим, что протокол D охватывает такой промежуток времени, в течение которого все рассматриваемые вероятности остаются неизменными. Тогда для восстановления скрытых профилей можно применить принцип максимума правдоподобия:

$$\textstyle\sum_{i=1}^{\ell} \ln p(u_i, r_i) \to \max,$$

где максимум берётся по всем профилям $\{p_{tu}\}$ и $\{q_{tr}\}$, удовлетворяющим ограничениям-равенствам $\sum_{t\in T} p_{tu} = 1$ для всех $u \in U$ и $\sum_{t\in T} q_{tr} = 1$ для всех $r \in R$. Для решения данной оптимизационной задачи предлагается алгоритм, в котором чередуются два шага:

— оптимизация профилей $\{p_{tu}\}$ при фиксированных $\{q_{tr}\}$;
— оптимизация профилей $\{q_{tr}\}$ при фиксированных $\{p_{tu}\}$.

На каждом шаге для оптимизации профилей выполняется несколько итераций EM-алгоритма. На каждой EM-итерации возникает более простая задача максимума правдоподобия, которая решается аналитически.

Скрытыми переменными в EM-алгоритме являются апостериорные вероятности того, что клиент u, выбирая ресурс r, удовлетворяет свой интерес t. Эти оценки крайне важны для многих приложений.

Главной особенностью алгоритма является его «симметричность»: разложения по клиентам (1) и по ресурсам (2) наравне используются в итерациях, благодаря чему и достигается согласованность профилей.

Начальное приближение задаётся либо случайным образом, либо исходя из априорной информации. В частности, во многих приложениях ресурсы и/или клиенты изначально классифицированы по набору тем T. Это позволяет устранить т. н. «проблему холодного старта» — когда надо принимать решения относительно ресурса или клиента, для которого в D не зафиксировано ни одной записи. В таких случаях вычислить искомый профиль невозможно, но его можно заменить априорным профилем, имеющим ту же самую структуру.

Эксперименты на модельных данных ($|R| = 200,\ |U| = 1000,\ |T| = 10,\ \ell = 50\,000$), в которых истинные профили были известны изначально, и точность их восстановления оценивалась в среднеквадратичном, показали, что 3–4 итераций на внешнем цикле и 4–5 EM-итераций на внутренних циклах вполне достаточно для восстановления, причём дальнейшее увеличение числа итераций может даже немного ухудшить точность.

Оценивание сходства клиентов и ресурсов

Существуют различные способы определить расстояние между клиентами $\rho(u, u')$ и между ресурсами $\rho(r, r')$: через корреляцию [3], через проверку статистической гипотезы о независимом совместном выборе [4], через ассоциативные правила, и др [5]. Имея профили, расстояние естественно определить как евклидову метрику в пространстве профилей.

Был проведён эксперимент на протоколах поисковой машины Яндекс. В данной задаче клиентами являются пользователи, делающие поисковые запросы; ресурсами являются документы, выдаваемые в результате поиска; действием пользователя считается переход по гиперссылке. Протокол охватывал семь дней регулярной работы Яндекса в 2005 году; $|R| = 1024,\ |U| = 7292,\ |T| = 20,\ \ell = 47\,000$. Часть ресурсов $R' \subset R$, $|R'| = 396$, была заранее классифицирована на 8 классов по тематике. Строились две метрики на R: по вероятности независимых совместных выборов [4] и по профилям ресурсов. Обе метрики использовались для классификации множества ресурсов R' методом k ближайших соседей. Число k настраивалось по скользящему контролю для каждой метрики отдельно. Доля ошибок оказалась равной 22% при $k = 12$ для первой метрики и 11% при $k = 15$ для второй метрики.

Таким образом, предложенный алгоритм не только агрегирует информацию о пользователях и о ресурсах в виде хорошо интерпретируемых тематических профилей, но и строит более адекватные оценки сходства.

Работа выполнена при поддержке РФФИ, проект № 05-07-90410.

Литература

[1] *Рудаков К. В., Воронцов К. В., Вальков А. С., Чехович Ю. В.* Технология анализа клиентских сред. — Форексис, 2005. — `www.forecsys.ru/cea.php`.

[2] *Marlin B.* Modeling user rating profiles for collaborative filtering // Neural Information Processing Systems (NIPS-16). — MIT Press, 2004.

[3] *Sarwar B. M., Karypis G., Konstan J. A., Reidl J.* Item-based collaborative filtering recommendation algorithms // World Wide Web. — 2001. — Pp. 285–295.

[4] *Воронцов К. В., Рудаков К. В., Лексин В. А., Ефимов А. Н.* Выявление и визуализация метрических структур на множествах пользователей и ресурсов Интернет // Искусств. Интеллект. — Донецк, 2006. — № 2 — С. 285–288.

[5] *Symeonidis P., Nanopoulos A., Papadopoulos A., Manolopoulos Y.* Collaborative filtering: Fallacies and insights in measuring similarity // PKDD Workshop on Web Mining, Berlin, Germany. — 2006.

Классификация биомагнитных данных и диагностика патологий

Махортых С. А., Семечкин Р. А.

ruslan-impb@mail.ru

Пущино, Институт математических проблем биологии РАН

Исследование причин развития нейродегенеративных заболеваний — одно из приоритетных направлений нейронауки. Проблема нейродегенеративных заболеваний стала в последние годы особенно острой в связи с увеличением продолжительности жизни в развитых странах Запада и ростом числа больных. Например, в США более 2 млн. человек страдают нейродегенеративными заболеваниями.

Актуальность данной проблемы обусловлена не только значительным числом больных, но и открывшимися во второй половине XX века возможностями добиваться существенного улучшения качества жизни этих пациентов. В данной исследовательской работе предложен комплексный метод классификации типов биомагнитного сигнала головного мозга по данным магнитной энцефалографии.

Магнитная энцефалография (МЭГ) — область современной математической биологии, занимающаяся изучением магнитных полей, связанных с высшей нервной деятельностью человека. МЭГ позволяет проводить исследование функциональных областей мозга и диагностику различных патологий [1, 3].

Исходные экспериментальные данные получены в Медицинской школе Нью-Йоркского университета. Измеряемый сигнал представляет собой пространственно-временную структуру: 148-мерный вектор измерений в 148 точках на поверхности головы, развернутый во временной ряд с частотой опроса датчиков 500 Гц.

Классификация сигнала

В последнее время интенсивно развивается подход к распознаванию, использующий спектральное представление сигнала ортогональными функциональными разложениями. Для распознавания типа активности сигнала предлагается следующий метод:

1. Векторизация данных МЭГ. Исходная функция $f(\theta, \varphi)$ определяется формулой
$$f(\theta, \varphi) = \sum_{l=0}^{l=N} \sum_{m=-l}^{l=l} a_{lm} Y(\theta, \varphi),$$
где коэффициенты разложения $a_{lm} = \int_0^{2\pi} \int_0^{\pi} f(\theta, \varphi) \sin\theta \, d\theta \, d\varphi$.
2. Выделение наиболее информативных гармоник.
3. Удаление шума из выбранных гармоник при помощи дискретного вейвлет-преобразования.
4. Проведение кластерного анализа итеративным методом k-средних. Найденные моменты времени, в которых присутствует аномальная компонента, становятся исходными данными в задаче локализации участков мозга, связанных с рассматриваемой патологией.
5. Локализация источника повышенной биомагнитной активности в выбранные моменты времени.
6. Анализ стохастической динамики биомагнитного сигнала.

Результаты исследования

Распознавание типа сигнала методом кластерного анализа проведено в зависимости от спектральной характеристики сигнала. Всего выделено четыре типа сигнала: A, B, C и D.

Решение обратной задачи представлено для каждого момента времени в виде токового диполя с переменным моментом. Для моментов времени, в которых наблюдается сигнал типа A, источники биомагнитной активности находятся в мозжечке, для сигнала B — в мозжечке и стволе головного мозга, а для сигнала типа C — в стволе (черная субстанция). Для моментов времени, в которых наблюдается сигнал типа D, источники биомагнитной активности основную часть времени находятся в коре большого мозга.

Результаты локализации источников повышенной активности в записи магнитной энцефалограммы подтверждают существующее в медицине мнение о связи болезни Паркинсона с поражениями подкорковых областей мозга. В частности, имеются данные о связи заболевания с гибелью меланинсодержащих нейронов одного из подкорковых ядер головного мозга — черной субстанции. Высшим уровнем регуляции движений

являются кора большого мозга, базальные ядра и мозжечок. Поэтому вызывает интерес то, что начало приступа Паркинсонизма, по результатам наших исследований, связано с одной из важнейших областей головного мозга — мозжечком, который не только обеспечивает непрерывный контроль двигательной активности, но и принимает участие в реализации когнитивного кодирования и памяти.

В ранее проведенных исследованиях [3] впервые было показано, что переключение из нормальной в аномальную активность приводит к упрощению динамики сигнала. Чем меньше корреляционная размерность аттрактора сигнала, тем более активным является источник этого сигнала. Таким образом, источник повышенной биомагнитной активности перемещается из мозжечка в ствол головного мозга во время вспышки паркинсонической активности. Об этом свидетельствует уменьшение значения корреляционной размерности: сигнал типа A — 5.04, сигнал типа B — 3.42, сигнал типа C — 1.86.

Работа выполнена при поддержке РФФИ, проекты № 07-01-00564, № 06-01-08039.

Литература

[1] *Устинин М. Н., Махортых С. А., Молчанов А. М. и др.* Задачи анализа данных магнитной энцефалографии // Компьютеры и суперкомпьютеры в биологии, М.: Институт компьютерных технологий, 2002. С. 327–349.

[2] *Дедус Ф. Ф., Махортых С. А., Устинин М. Н., Дедус А. Ф.* Обобщенный спектрально-аналитический метод обработки информационных массивов. Задачи анализа изображений и распознавания образов // М.: Машиностроение, 1999. — 357 с.

[3] *Дергузов А. В., Махортых С. А., Семечкин Р. А.* Комплексная диагностика Паркинсонизма по данным магнитной энцефалографии // Электронный журнал «Исследовано в России», 2006. — С. 646–659. zhurnal.ape.relarn.ru/articles/2006/065.pdf.

[4] *Sarvas J.* Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem // Phys. Biol. — Vol. 32. — 1987. — pp. 11–22

[5] *Youdim M., Riederer P.* Understanding Parkinson's disease // Scientific American. — Vol. 276. — 1997. — pp. 52–59.

Концепция построения систем анализа и фильтрации Интернет трафика на основе методов интеллектуального анализа данных

Машечкин И. В., Петровский М. И., Глазкова В. В., Масляков В. А.
Москва, МГУ им. М. В. Ломоносова

Проблема контроля доступа к Интернет ресурсам возникает при решении следующих важных задач: блокирование доступа к нелегальной (экстремистской, антисоциальной и другой) информации; пресечение утечек конфиденциальной информации через Интернет; ограничение использования Интернет ресурсов не по назначению, в частности, блокирование доступа к развлекательным ресурсам в рабочее время.

Традиционно в существующих системах анализа и фильтрации Интернет информации применяется подход, основанный на применении *экспертных баз знаний* адресов Интернет ресурсов, где для каждого ресурса эксперт задает набор релевантных тем (категорий). Однако такие системы обладают рядом недостатков:

— не поддерживается анализ содержимого трафика в реальном времени, что необходимо, поскольку контент одного и того же ресурса может динамически изменяется;
— не поддерживается контентный анализ исходящего Интернет трафика (для предотвращения утечки конфиденциальной информации);
— необходимо использовать обновляемые извне базы знаний, что может быть недопустимо по соображениям безопасности, и в целом, качество фильтрации зависит от оперативности работы организаций, поддерживающих подготовку обновлений;
— невозможно классифицировать Интернет ресурс, данных о котором нет в текущей базе знаний.

Для преодоления данных недостатков авторами предлагается подход, основанный на *применении методов машинного обучения для анализа содержимого Интернет трафика в режиме реального времени*, что позволяет построить систему фильтрации, обладающую такими свойствами, как: *адаптируемость* — способность дообучаться и анализировать содержимое Интернет ресурсов в динамике; *автономность* — независимость от внешних баз знаний и экспертов; независимость системы от языка анализируемых ресурсов.

В процессе обучения на основе обучающей совокупности, состоящей из заранее рубрицированных HTML-документов, строится классификатор, который позволяет определять релевантные категории для произвольных ресурсов аналогичного содержания (Рис. 1). Впоследствии

класификатор может *дообучаться* на новых ресурсах. В предлагаемом подходе учитывается, что Интернет ресурсы являются *многотемными (multi-label)*, то есть каждый Интернет ресурс может быть отнесен к нескольким категориям. Для решения задачи многотемной классификации реализован подход на основе декомпозиции multi-label проблемы в набор задач бинарной классификации на основе подходов «каждый-против-остальных» и «каждый-против-каждого» [1]. Для начального обучения бинарных классификаторов используется метод SVM, а для дообучения — Kernel Perceptron.

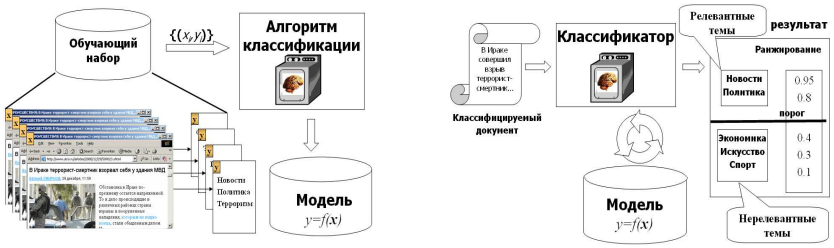


Рис. 1. Процесс обучения (слева) и процесс классификации (справа).

В процессе классификации построенный классификатор для новомого документа (или фрагмента) выдает релевантности всех тем (из предопределённого на этапе обучения набора). Далее находится *порог отсечения по релевантности* и отбрасываются наименее релевантные темы (Рис. 1). Пороговое значение, также определяется на основе методов машинного обучения, т. е. на обучающей совокупности строится модель, предсказывающая для классифицируемого документа значение порога отсечения по релевантности [2].

Для представления HTML-документов реализованы подходы на основе ключевых слов (с поддержкой стемминга) и n-грамм. В качестве меры сходства используется частотная мера сходства типа TF-IDF, а также модифицированная нами мера сходства на основе k-spectrum kernel. Кроме того, учитывается ссылочная структура HTML-документов. Для этого гиперссылки в данном документе заменяются на идентификаторы тем, релевантные документу, на который указывает ссылка (если эти темы известны, т. е. если документ, на который указывает ссылка, уже был классифицирован).

Авторами была предложена и реализованна в виде прототипа *архитектура системы интеллектуального анализа Интернет трафика* (Рис. 2).

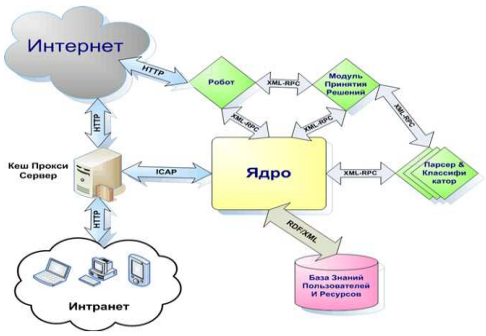


Рис. 2. Архитектура системы.

Модуль лексического разбора (парсинга) и классификации осуществляет преобразование HTML-документов во внутреннее представление и осуществляет многотемную (multi-label) классификацию. *Кеш-прокси сервер* предназначен для кеширования и фильтрации трафика локальной сети и перенаправляет запрашиваемые ресурсы ядру системы. *Ядро* координирует организацию работы модулей системы, а также сохраняет в *автоматически формируемой базе знаний* параметры ресурсов результаты анализа и классификации. *Модуль принятия решений* осуществляет разрешения или блокирования доступа к Интернет ресурсу с учетом результатов классификации и заданных политик доступа для конткретного пользователя. *Робот* используется в системе для скачивания содержимого ресурсов для последующего обучения на них и для отложенной классификации с целью пополнения базы знаний и классификации документов, ссылки на которые были выявлены ранее в процессе анализа других документов.

Работа выполнена при поддержке РФФИ, проект №06-01-00691, гранта Президента РФ МК-4264.2007.9, а также в рамках госконтракта с Федеральным агентством по науке и инновациям №02.514.11.4026.

Литература

[1] *Petrovskiy M. I.* Paired Comparisons Method for Solving Multi-label Learning Problem. — Hybrid Intelligent Systems, IEEE Press, 2006. — Pp. 42–48.

[2] *Petrovskiy M. I., Glazkova V. V.* Linear Methods for Reduction from Ranking to Multilabel Classification. — Springer-Verlag, LNAI, 2006. — V. 4304 — Pp. 1152–1156.

Анализ влияния психофизиологических параметров участников на агрегированное поведение рынка методами экспериментальной экономики

Меньшиков И. С.
ivanm@ccas.ru
Москва, Вычислительный центр РАН

В работе описаны исследования, проводимые в Лаборатории экспериментальной экономики (ЛЭЭ) МФТИ и ВЦ РАН, оборудованной системой психологического тестирования и аппаратно-программным комплексом из пяти стабилографических кресел, разработанных по специальному заказу в ОКБ «Ритм» г. Таганрог. Приводятся некоторые результаты, характеризующие влияние психофизиологических факторов на принятие экономических решений участниками лабораторных рынков. Анализ результатов экспериментов стимулировал новые исследования на стыке теории игр и экспериментальной экономики.

Цели и методы

Цель работы состоит в том, чтобы проанализировать влияния психофизиологических факторов на процесс принятия экономических решений человеком. Традиционно в исследовании операций предполагается, что процесс принятия решений можно формализовать в виде оптимизационной задачи, которую остается только решить. Психологический тип ЛПР не учитывается. Ситуация осложняется, если имеются неконтролируемые факторы, поскольку приходится моделировать отношение ЛПР к риску. Здесь уже необходимо касаться психологии принятия решений. Еще сложнее ситуация, когда есть несколько участников экономической ситуации. Несмотря на успехи теории игр ясно, что однозначные рекомендации по принятию решений теория игр может дать только в отдельных простых случаях. В типичной экономической ситуации решение принимать надо, причем в ограниченное время, а заведомо оптимального решения нет. Нам хочется изучить, как разные люди это делают.

Большие возможности в исследовании данного вопроса дает метод экспериментальной экономики, который состоит в том, чтобы создать экономическую ситуацию в контролируемых условиях лаборатории. Для этого обычно в лаборатории с помощью сети компьютеров конструируется рынок, приглашаются участники, которым детально поясняются все правила, и определяется финансовая мотивация участников. Иногда финансовая мотивация заменяется учебной: заработанные на лабораторном рынке деньги переводятся в учебные очки, на основании которых выставляется оценка.

Для определения психологического типа перед началом эксперимента все участники проходят психологическое тестирование как минимум по двум взаимодополняющим тестам. Во время эксперимента физиологическое состояние участника может быть измерено полностью безопасным для него образом с помощью специального стабилографического кресла. Сигналы со всех кресел собираются на центральном компьютере у диспетчера эксперимента, который с соседнего компьютера контролирует лабораторный рынок. В результате от каждого эксперимента остаются три группы данных: результаты психологического тестирования, посекундная запись всех экономических действий каждого участника, данные с кресел (частота 50 раз в секунду).

На данном этапе вряд ли можно рассчитывать на создание точной модели поведения человека при принятии экономических решений. Скорее, можно надеяться на открытие закономерностей на основе анализа разнородных экспериментальных данных, опираясь на хорошо разработанные методы распознавания. Изложенная программа исследований начала реализовываться в Лаборатории экспериментальной экономики (ЛЭЭ) МФТИ и ВЦ РАН. Получены первые обнадеживающие результаты [3].

Междисциплинарный проект. При лабораторном исследовании процесса принятия экономических решений возникает уникальная возможность системно обследовать лиц, принимающих решения. В ЛЭЭ такое обследование участников экспериментов ведется по двум направлениям: психологическое тестирование и компьютерная стабилография.

Результаты тестирования служат основанием для выделения психологических типов. Основной гипотезой для нас является предположение о том, что психологический тип участника определенным образом влияет на процесс принятия решений и на его потенциальную результативность в заданной экономической ситуации [1].

Система стабилографических кресел позволяет на основе агрегирования первичной информации определять параметры функционального состояния участников как в покое, до и после эксперимента, так и во время активной фазы эксперимента. Динамика функционального состояния участников в процессе эксперимента показывает, как процесс принятия решений отображается на биомеханическом уровне. В определенных случаях можно говорить о функциональном состоянии группы в процессе эксперимента.

Методы распознавания. На данном этапе трудно рассчитывать на построение достаточно полной модели поведения участников лабораторных рынков с учетом их психофизиологических параметров, поэтому целесообразно проверять гипотезы с помощью методов распознавания [2].

Сопоставление данных тестирования и стабилографии дает инструментальное подтверждение психологической типологии. Выявленная связь психологических параметров и результативности участников в длительной серии однородных экспериментов показывает исходную асимметрию участников по отношению к заданному типу экономических ситуаций.

Развитие методов теории игр и экспериментальной экономики

Агрегированное равновесие. В процессе анализа экспериментальных данных было обнаружено, что условие наилучшего ответа на действия остальных, на котором основано равновесие Нэша, гораздо лучше выполняется в среднем по группе, чем для отдельного участника. Определение агрегированного равновесия (АР) адекватно лабораторным экспериментам со случайным выбором состава группы. АР является естественным расширением теоретико-игровых понятий равновесия. Для проведенной серии экспериментов по однозвенным сетевым рынкам показано соответствие динамики лабораторных рынков АР.

Модифицированное равновесие. Понятие модифицированного равновесия основано на модифицированной игре, в которой каждый участник фактически заменяется группой, обладающей случайным разбросом действий в соответствии с заданной функцией распределения. Для сетевых рынков такой подход позволяет выделить из множества равновесий единственное. Результаты проведенной серии экспериментов показывают, что разброс индивидуальных действий во круг среднего значения согласуется с гипотезой о нормальности распределения.

Работа выполнена при поддержке РФФИ, проекты № 07-01-00605а, № 06-01-08057-офи, гранта Президента РФ по государственной поддержке ведущих научных школ (код проекта НШ-5379.2006.1) и по программе «Развитие научного потенциала высшей школы (2006–2008 годы)» Федерального агентства по образованию (код проекта РНП.2.2.1.1.2467).

Литература

[1] *Меньшиков И. С., Меньшикова О. Р.* Лабораторные исследования информационной эффективности рынков. — М.: ВЦ РАН, 2006. — 58 с.

[2] *Журавлёв Ю. И., Рязанов В. В., Сенько О. В.* РАСПОЗНАВАНИЕ. Математические методы. Программная система. Применения. — М.: ВЦ РАН, 2006. — 306 с.

[3] *Лукьянов В. И., Максакова О. А., Меньшиков И. С., Меньшикова О. Р., Сенько О. В., Чабан А. Н.* Функциональное состояние и эффективность участников лабораторных рынков // Известия Академии наук. Теория и системы управления. — 2007. — № 6. — (в печати).

Кластеризация семантических знаний в задаче распознавания ситуаций смысловой эквивалентности

Михайлов Д. В., Емельянов Г. М.
mdv@novsu.ac.ru
Великий Новгород, ГОУ ВПО НовГУ им. Ярослава Мудрого

Центральной задачей анализа смысла высказывания Естественного Языка (ЕЯ) является выделение класса Семантической Эквивалентности (СЭ). Наиболее известная система классов СЭ в ЕЯ определяется множеством Π^R правил синонимических преобразований ЕЯ-высказываний в рамках стандартных Лексических Функций (ЛФ) [1]. При этом для каждого правила $\pi \in \Pi^R$ задается условие $r(\pi)$ его применимости. Условие $r(\pi)$ есть совокупность требований к синтаксическим и семантическим свойствам лексических единиц исходного ЕЯ-высказывания, заменяемых посредством π. В содержательном плане такие требования определяются Смысловыми Отношениями (СО) между некоторым ЕЯ-словом, относительно которого задается Ситуация СЭ (ССЭ), и его лексико-семантическими производными (лексическими коррелятами), которые входят в заменяемый комплекс лексических единиц. В лексической семантике такие СО описываются стандартными ЛФ. Ставится задача: на основе признаков слов пары T сравниваемых ЕЯ-высказываний T_1 и T_2 отнести T к одному из известных классов $\pi \in \Pi^R$, либо образовать с помощью T новый класс. Данная задача есть классическая задача распознавания образов. В настоящей работе мы рассмотрим метод, основанный на теории Анализа Формальных Понятий (АФП) [1, 2], для автоматизации формирования $r(\pi)$ как прецедента класса π.

Наибольший интерес для описания $\{r(\pi)\}$ представляют ситуации с ЛФ-параметрами. Посредством этих ЛФ описываются Расщепленные Значения (РЗ). Фактически РЗ есть конструкция, непосредственно задающая СО в рамках $r(\pi)$. Это позволяет поставить задачу его выявления и обобщения по аналогии с описанием семантики Именных Групп на основе формализованного представления толкований Лексических Значений (ЛЗ) слов в виде теорий [1]. Сказанное подтверждается наработками по Русскому общесемантическому словарю (РОСС): ЛФ используются в качестве Семантических Характеристик отдельных слов в РОСС [2]. Следовательно, такие слова могут быть и названиями отношений в утверждениях теорий других слов. Пример — значение Лексической Функции $Oper_1$ для ЛЗ «эксперимент» (т. е. «осуществлять»). Как видно из Рис. 1, оно присутствует в одном из утверждений теории ЛЗ «экспериментировать». Данное ЛЗ эквивалентно РЗ «осуществлять эксперимент», где значением ЛФ $Oper_1$ задается СО типа «операция с. . . » между 1-м участником ССЭ (кто осуществляет эксперимент) и её названием.

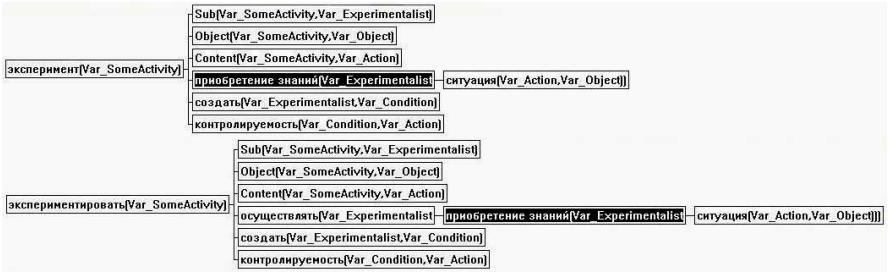


Рис. 1. Теории ЛЗ «эксперимент» и «экспериментировать».

Пусть для слов $w_1 \in T_1$ и $w_2 \in T_2$ мы имеем описания теорий ЛЗ. При этом утверждение (постулат значения [1]) каждой теории либо задает бинарное отношение R_2 между понятиями C_1 и C_2 посредством тройки:

$$M_p = (R_2, C_1, C_2), \tag{1}$$

либо представляет собой описание отношения R_n произвольной арности:

$$M'_p = (R_n, C, L^M). \tag{2}$$

В этом случае постулат значения определяется рекурсивно на основе списка L^M утверждений вида (1) и (2). Посредством L^M задается связь понятия C с другими словами и понятиями. При этом смысл слова $w_i \in T_1 \cup T_2$ определяется множеством функций [1], которые задаются утверждениями теории ЛЗ w_i и составляют для него набор признаков.

Утверждение 1. *Смысловое отношение F, значимое для формирования $r(\pi)$, между некоторым словом $w_1 \in T_1$ и его лексическим коррелятом $w_2 \in T_2$, входящим в РЗ, будет иметь место тогда, когда*

$$L_1^M = L_{11}^M \cup \left\{(F, C, L_{22}^M)\right\} \cup L_{12}^M;$$
$$L_2^M = L_{11}^M \cup L_{22}^M \cup L_{12}^M;$$
$$L_{11}^M \cap L_{22}^M = \varnothing, \quad L_{11}^M \cap L_{12}^M = \varnothing, \quad L_{12}^M \cap L_{22}^M = \varnothing;$$

где L_1^M — набор утверждений теории ЛЗ для w_1, а L_2^M — для w_2.

Правильность выявления F зависит от корректности ЕЯ-толкований w_1 и w_2. Рассмотрим решение этой задачи методами АФП.

Лексическими Функциями описывается в первую очередь лексическая сочетаемость. Последняя определяется Лексическим Значением

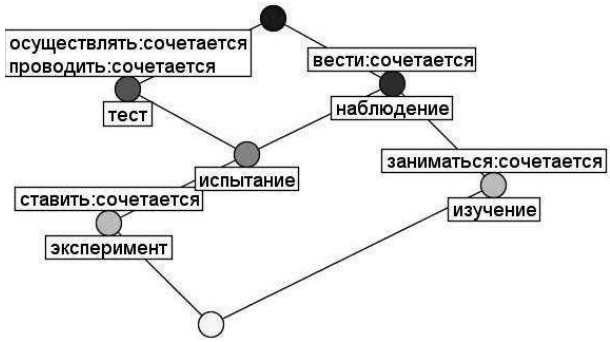


Рис. 2. Слова-аргументы Лексическо Функции $Oper_1$ из верхней окрестности для ЛЗ «эксперимент».

ключевого слова ЛФ-синонимической замены. Более узкое ЛЗ (в терминологии АФП— гипоним) включает более широкие ЛЗ (гиперонимы), которые упоминаются в толковании рассматриваемого ЛЗ, а, следовательно, и в его теории. Таким образом, слово-гипоним в большинстве случаев будет иметь в качестве значений ЛФ-параметра значения этой же ЛФ тех слов-гиперонимов, которые упоминаются в его толковании (теории). Для предикатных слов отношение гипонимии может быть выявлено анализом содержания их семантических валентностей [2].

Применением специализированного ПО ToscanaJ (`http://toscanaj.sourceforge.net`) для заданной ЛФ строится модель системы слов — ее аргументов, Рис. 2.

При этом слова-аргументы ЛФ выступают в качестве объектов, а её значения — в качестве атрибутов. На основе полученной модели критерий адекватности $r(\pi)$ формулируется следующим образом.

Утверждение 2. *Теории ЛЗ w_1 и w_2 адекватно задают $r(\pi)$ при условии существования отношения F, если F принадлежит множеству формальных атрибутов того ЛЗ, которое является наименьшей верхней гранью (супремумом) множества слов верхней окрестности [2] ЛЗ w_2.*

Требования к РЗ с ЛЗ-супремумом определяются аналогично.

Работа поддержана УНИК НовГУ и РФФИ, проект № 06-01-00028.

Литература

[1] *Михайлов Д. В, Емельянов Г. М.* Модель сортовой системы языка в задаче построения семантического образа высказывания на уровне глубинного синтаксиса // Таврический Вестник Информатики и Математики. — Симф., 2006. — № 1. — С. 79–90.

[2] *Михайлов Д. В., Емельянов Г. М.* Применение семантических полей словаря РОСС в задаче построения Модели Управления предикатного слова // Всеросс. конф. ММРО-12. — Москва: Макс Пресс, 2005. — С. 382–385.

Метод непараметрического многофакторного анализа вызванных ответов в ЭЭГ человека

Морозов А. А., Морозов В. А., Обухов Ю. В., Строганова Т. А., Обухова Е. Ю.
{morozov,vmorozov,obukhov}@cplire.ru, stroganova@pirao.ru
Москва, ИРЭ РАН, ПИ РАО

Разработанный авторами метод непараметрического многофакторного анализа (НМА) позволяет анализировать влияние различных факторов на частотно-временную динамику волновых процессов коры головного мозга, порождаемых стимулом, а также эффекты взаимодействия различных факторов. Метод был реализован и успешно применён для анализа данных, собранных в ходе экспериментов с иллюзорными изображениями, проводимых Психологическим институтом РАО. Анализировались фазово-связанные и фазово-несвязанные компоненты электроэнцефалограмм (ЭЭГ), а также спектрограммы полной мощности ЭЭГ [1, 2, 3, 4].

Метод НМА направлен на преодоление следующих проблем, связанных с применением существующих статистических методов анализа вызванных ответов в электроэнцефалограммах:

1. Наличие пространственных (между различными каналами ЭЭГ), а также временны́х причинно-следственных (статистических) зависимостей в ансамблях экспериментальных данных.
2. Существенная негауссовость ансамблей экспериментальных данных, во многих случаях приводящая к невозможности применения наиболее мощных (для гауссовых выборок) и хорошо отработанных параметрических методов статистического анализа.
3. Нестационарность ЭЭГ, проявляющаяся на различных временны́х масштабах и, во многих случаях, обуславливающая некорректность применения существующих методов обработки сигналов.

Учёт причинно-следственных зависимостей

Для нейрофизиологических исследований наличие в экспериментальных данных причинно-следственных зависимостей опасно, прежде всего, возможностью получения заниженных оценок ошибки первого рода, то есть, вероятности того, что наблюдаемые эффекты возникли случайно и не отражают объективно существующие закономерности.

Причинно-следственные зависимости, обусловленные «размытием» сигнала по времени, мы устраняем с помощью прореживания последовательностей измеренных значений. Наличие причинно-следственных зависимостей между каналами ЭЭГ учитывается с помощью преобразования многомерных (статистически связанных) исходных данных в одномерный массив. Это преобразование осуществляется с помощью метода анализа главных компонент (principal component analysis, PCA). В качестве результата преобразования мы берём проекцию исходного массива данных на ось одного из главных компонентов. Обычно, выбирается первый главный компонент, который вносит наибольший вклад в изменчивость анализируемых данных.

Анализ эффектов взаимодействия факторов

Для проверки статистических гипотез мы используем непараметрические критерии. Для анализа эффектов взаимодействия внутригрупповых и межгрупповых факторов мы разработали специальный метод, основанный на том факте, что многие важнейшие факторы, влияющие на интерпретацию результатов нейрофизиологического эксперимента, являются бинарными, т. е. имеют два значения (например, тестовый и контрольный стимулы, левое и правое полушария). Поэтому для учёта влияния таких факторов достаточно вычислить парные разности значений, соответствующих противоположным значениям бинарного фактора, и проверять те или иные статистические гипотезы на ансамблях вычисленных разностей с помощью критерия знаков, критерия парных сравнений Вилкоксона, критерия Манна-Уитни, критерия Флайгера-Полицелло (Fligner-Policello) или перестановочного метода (permutation).

Учёт нестационарности ЭЭГ

В случаях, когда нестационарность сигнала проявляется в достимульном интервале, возникает вопрос, какие именно интервалы времени можно использовать в качестве референтной области, для сравнения с ними сигнала после подачи стимула? Для решения этой проблемы мы разработали метод сравнения исследуемой величины с её значениями в многосегментной референтной области. Этот метод основан на следующих исходных предположениях:

1. Сегменты референтной области являются квазистационарными.
2. Рассматриваемый набор сегментов референтной области является репрезентативным по отношению к решаемой задаче, то есть, адекватно описывает все возможные состояния достимульного интервала.
3. Все сегменты референтной области содержат результаты независимых друг от друга наблюдений.

Исходя из этих предположений, осуществляется проверка статистических гипотез о стохастическом равенстве исследуемого ансамбля данных (соответствующего некоторой послестимульной пространственно-временной области) ансамблям различных сегментов референтной области. На основе полученных оценок $p_1, \ldots, p_N$ (N — общее количество сегментов референтной области) минимальной (для рассматриваемого теста) статистической значимости различий сравниваемых ансамблей вычисляется $F_c(K, p_1, \ldots, p_N)$ — интегральная функция распределения вероятностей ошибки первого рода для утверждения, что рассматриваемый ансамбль данных стохастически больше (меньше) ансамблей некоторых K сегментов референтной области ($K \leqslant N$). Функция F_c определяется с помощью обобщённого биномиального распределения.

Трёхмерная визуализация результатов

Авторами работан метод трёхмерной визуализации результатов анализа, который пригоден как для визуализации результатов однофакторного анализа данных (вырожденный случай), так и для визуализации результатов анализа эффектов взаимодействия факторов. Идея состоит в том, что на горизонтальных осях координат откладываются значения факторов, не являющихся бинарными (например, «электроды» или «время»), а на третьей оси координат откладывается статистическая характеристика (среднее или медиана) исследуемой величины или разность значений (если анализируется взаимодействие бинарных факторов). Кроме того, с помощью цвета отображается информация о наличии статистически значимого отличия исследуемых выборок от референтной области, а также о знаке отличия (больше, меньше).

Работа выполнена при поддержке РФФИ, проекты № 06-07-89302 и № 05-01-00651, РГНФ, проект № 07-06-00208, а также программы Президиума РАН «Фундаментальные науки — медицине».

Литература

[1] *Morozov A. A., Obukhov Yu. V., Stroganova T. A., Tsetlin M. M., Orekhova E. V.* The search of the regularity in the spatio-temporal dynamics of the human visual cortex oscillations // Pattern Recognition and Image Analysis. — 2005. — Vol. 15, № 4. — pp. 697–699.

[2] *Морозов А. А., Морозов В. А., Обухов Ю. В., Строганова Т. А.* Метод многофакторного анализа электроэнцефалограмм человека на основе вейвлет-спектрографии и непараметрической статистики // Докл. VII межд. науч.-техн. конф. «Физика и радиоэлектроника в медицине и экологии», Владимир: Изд-во «Собор», 2006. — Книга 1. — С. 145–147.

[3] *Морозов А. А., Морозов В. А., Обухов Ю. В., Строганова Т. А.* Непараметрический метод многомерного многофакторного анализа электроэнцефалограмм человека // Искусственный интеллект. — 2006. — № 3. — pp. 603–612.

[4] *Stroganova T. A., Orekhova E. V., Prokofyev A. O., Posikera I. N., Morozov A. A., Obukhov Y. V., Morozov V. A.* Atypical event-related potentials response to illusory contour in boys with autism // NeuroReport. — 2007. — Vol. 18, № 9. — pp. 931–935.

Распознавание неоднородностей, определение их геометрических характеристик и построение 3D геометрических моделей в задачах неразрушающего контроля

Николаев А. А.
beloian@km.ru
Москва, МГТУ им. Н. Э. Баумана

Разработан новый подход к реализации систем технического машинного зрения на основе контурного анализа [1] в задачах неразрушающего контроля магнитным и тепловым методами [2, 3]. Предложена и реализована автоматизированная технологическая цепочка построения 3D геометрической модели (ГМ) обнаруженных неоднородностей на фоне модели объекта контроля. Разработаны элементы программно-математического обеспечения (ПМО) расшифровки зашумленных изображений и построения ГМ с дефектами, ориентированные на работу в реальном времени.

Задача построения геометрической модели неоднородностей, представление неоднородностей как системы контуров

Задача построения 3D ГМ неоднородностей (дефектов) на фоне ГМ объекта контроля является задачей компьютерного зрения. Имеется изображение, получаемое в режиме реального времени. В рассматриваемых методах неразрушающего контроля (магнитном и тепловом) задачей предобработки часто является задача подавления шума (например, в [3] используются медианные фильтры). В разрабатываемой системе к задачам предобработки относятся также адаптивная бинаризации и морфологическое расширение. В качестве задачи извлечения признаков рассматривается задача выделения контуров. Для выделения контуров использованы алгоритмы «Жука» [4, 5], Розенфельда [6] и метод активных контуров [7]. Задача обнаружения неоднородностей (дефектов) является задачей обнаружения объектов. В данной работе задача обнаружения объектов рассматривается как задача распознавания зашумленных контуров. Построение ГМ неоднородности является задачей анализа характеристик объекта. Исследуемые эталонные ГМ дефектов представляются в виде классов систем эквализированных контуров. Система эквализированных контуров является набором линий уровней, которые путем

пространственной триангуляции преобразуются в набор граней, описывающих дефект.

Задача классификации дефектов

Рассмотрим задачу классификации дефектов как задачу распознавания зашумленных контуров.

Воспользуемся обозначениями, принятыми в [1].

Пусть $\Gamma_{(j)}^{(c)} = |\mu_{(j)}| \exp\{i\Delta\varphi_{(j)}\}\{\gamma_{(j)}(n + d_j)\}_{0,k-1}$ — сигнальный контур, состоящий из k векторов, и полученный путем преобразований (поворота, масштабирования и сдвига) из исходного эталонного контура $\Gamma_{(j)} = \{\gamma_{(j)}(n)\}_{0,k-1}$ класса A_j, где $j = 1, \ldots, M$, M — общее число классов распознаваемых контуров и $\gamma_{(j)}(n)$ — n-й вектор j-го контура в комплекснозначном представлении [1]. Скалярные параметры $|\mu_{(j)}|$ — масштаб, $\Delta\varphi_{(j)}$ — поворот и $d_{(j)}$ — сдвиг для каждого контура предполагаются неизвестными. Подаваемый на вход устройства распознавания зашумленный контур имеет вид $N_{(j)} = \Gamma_{(j)}^{(c)} + Z$, где Z — шумовой контур с дисперсией σ^2. Для получения оценок параметров масштаба, поворота и сдвига используется метод максимального правдоподобия. Для распознавания строятся контурные согласованные фильтры (КСФ) [1] для алфавита эталонных контуров из классов $A_1, \ldots, A_M$. Решающее устройство определяет номер фильтра с максимальным по модулю выходным сигналом, и при превышении порогового уровня выносится решение в пользу класса, номер которого совпадает с номером фильтра. После выбора номера класса решается задача об отнесении контура к зашумленному или шумовому [1, 8]. В виду краткости изложения, приведем только результаты данной задачи. Найдены геометрические параметры: класс эталонного контура, координаты геометрического центра масс, оценки параметров масштаба $|\mu|$ и поворота $\Delta\varphi$.

Построение 3D геометрических моделей обнаруженных неоднородностей на фоне модели объекта контроля

Входной информацией данной задачи являются результаты предыдущей задачи. Входные линии уровней (система контуров) подаются на вход процедуре пространственной триангуляции, работающей на основе триангуляции Дэлоне, производится разбиение и создание набора граней (треугольников), описывающих неоднородность (дефект).

Программный комплекс для решения задач классификации контуров (дефектов) и определения их геометрических параметров

На основе созданного алгоритма в среде Microsoft VS 7.0 разработаны элементы программно-математического обеспечения (ПМО). На вход

ПМО подается набор классов (дефектов), распознаваемая матрица результатов измерений, исходные данные для алгоритма принятия решений (априорные вероятности, значение отношения сигнал/шум и др.). Результатом работы комплекса ПМО является решение об отнесении распознаваемого контура к одному из классов и оценки линейных преобразований ($|\hat{\mu}|$, $\Delta\hat{\varphi}$ и $\hat{d}$, смещение контура относительно начала координат), или решение о том, является ли распознаваемый контур шумовым. По обработанным данным модуль триангуляции производит построение 3D ГМ системы дефектов.

Выводы

Разработанный комплексный алгоритм на основе теории распознавания образов является быстрым и позволяет создавать 3D ГМ систем неоднородностей (дефектов) на фоне ГМ объекта контроля. Дальнейшее увеличение скорости обработки, по мнению автора, реализуемо путем включения в комплексный алгоритм элементов адаптивного обучения под конкретную задачу неразрушающего контроля. Автоматически получаемая 3D ГМ системы дефектов на фоне объекта контроля в дальнейшем импортируется в программный комплекс, решающий трехмерные и оболочечные задачи напряженно-деформированного состояния и задачи прогнозирования остаточного ресурса исследуемых объектов контроля.

Работа выполнена при поддержке РФФИ, проект № 07-08-00574-а.

Литература

[1] *Фурман Я. А.* Введение в контурный анализ. Приложения к обработке изображений и сигналов. — М.: Физматлит, 2003. — 592 с.

[2] *Клюев В. В.* Неразрушающий контроль и диагностика. — М.: Машиностроение, 2005.

[3] *Будадин О. Н.* Тепловой неразрушающий контроль изделий: Научно-методическое пособие. — М.: Наука, 2006. — 472 с.

[4] *Дуда Р., Харт П.* Распознавание образов и анализ сцен. — М.: Мир, 1976.

[5] *Klette R., Rosenfeld A.* Digital geometry. Geometric methods for digital picture analysis. — San Francisco, CA: Morgan Kaufmann, 2004.

[6] *Розенфельд А.* Распознавание и обработка изображений. — М.: Мир, 1987.

[7] *Blake A., Isard M.* Active Contours. — Springer-Verlag, 1998.

[8] *Николаев А. А.* Математические аспекты классификации дефектов и определения их геометрических параметров при диагностировании магнитным методом // Информационно-математические технологии в экономике, технике и образовании — 2006, Екатеринбург: УГТУ-УПИ, 2006. — С. 35–36.

Цифровая диагностика остеопороза в программном комплексе для медицинской цифровой рентгенографии

Ольшевец М. М., Устинин М. Н.
om@impb.ru
Пущино, ИМПБ РАН

В Пущинском научном центре РАН разработана цифровая компьютерная приставка к стандартному медицинскому рентгеновскому аппарату для получения рентгенограмм без применения рентгеновской пленки. Важной частью созданного аппаратно-программного комплекса является компьютерная программа получения, просмотра, анализа, обработки, хранения и распечатки цифровых рентгеновских изображений.

Созданный программный комплекс объединяет в себе функции управления камерой, графического редактора и системы управления архивом цифровых снимков, сведений о проведенных обследованиях и личных данных пациентов [1].

Программа обеспечивает выполнение следующих задач:

— получение цифрового снимка;
— хранение снимков и организация быстрого доступа к ним;
— визуализация снимков и операции цифровой диагностики;
— цифровая обработка снимков и печать на бумаге или пленке.

Относительно высокая зашумленность и низкая контрастность получаемого снимка могут приводить к появлению артефактов при неправильно выбранном сценарии обработки изображения. Поэтому наиболее широкое применение в повседневной рентгенологической практике находят алгоритмы обработки изображения из класса алгоритмов поэлементной обработки, а также некоторые алгоритмы сглаживания. В частности, имеется возможность коррекции изображения по интерактивно задаваемой пользователем передаточной функции.

В состав программного комплекса включены шаблоны нескольких известных операторов свертки при различных размерах окна (гауссиан, лапласиан, и др.), пользователь может выбрать наиболее подходящую операцию, последовательно применять фильтры, задать в диалоговом режиме произвольное окно и матрицу весов фильтра и сохранить созданный шаблон для дальнейшего исследования и анализа снимка.

Также реализованы некоторые нелинейные алгоритмы обработки изображений со скользящим окном.

Наряду с шумоподавлением и повышением контрастности изображения важной задачей является сжатие цифровых рентгеновских снимков. Актуальность проблемы обусловлена необходимостью хранения большого числа полноформатных цифровых изображений и их передачи по ка-

налам связи без потерь значимой диагностической информации. Для решения обеих задач использовались разложения по wavelet-базисам (базисам всплесков) [2], представляющим собой специфические системы ортогональных функций, хорошо подходящих для обработки резко меняющихся данных. Существуют эффективные алгоритмы быстрого преобразования исходного сигнала в пространство коэффициентов разложения. Дальнейшая обработка цифровых массивов с использованием wavelet-базисов ведется в пространстве коэффициентов. Методы разложений по базисам всплесков ведут к экономному решению многих задач обработки, требующих реализации в рамках ограниченных аппаратных или вычислительных ресурсов.

Для осуществления сжатия в разложении цифрового изображения по выбранному базису сохраняют только коэффициенты с амплитудой, превышающей некоторый порог. При этом частично решается и задача шумоподавления. Фактически, с выбором порога решается задача нелинейной аппроксимации по элементам базиса, вносящим наибольший вклад в разложение. В применении к задачам медицинской бесплёночной рентгенографии удается достигнуть значительной степени сжатия хранимых цифровых рентгеновских снимков при сохранении приемлемого качества изображения.

В программе также реализован алгоритм коррекции неравномерности яркости по полю снимка. Яркость снижается по мере удаления от центра изображения. Эта неравномерность является, с одной стороны, следствием высоких требований к используемой в приставке цифровой фотокамере, с другой стороны свойством рентгеновского источника. Для коррекции этого дефекта применяется подход с использованием эталонного снимка, выполняемого при заданных условиях и не содержащего объектов. Тестовый снимок аппроксимируется с использованием различных алгоритмов и полученное гладкое поле яркости используется в качестве корректирующего множителя при обработке целевых снимков. В результате обеспечивается постоянный уровень фона, а также происходит сглаживание точечных шумов.

Такая коррекция необходима для реализованной в программе методики определения рентгеновской плотности снимаемых объектов (денситометрия) по цифровому рентгеновскому снимку. Денситометрия является основным методом диагностики остеопороза — заболевания скелета, характеризующегося снижением плотности кости и нарушением структуры костной ткани. Остеопороз приводит к увеличению хрупкости костей и риска их переломов. Последствия остеопороза в виде переломов тел позвонков и трубчатых костей приводят к значительному подъему заболеваемости и смертности среди лиц пожилого возраста. При цифровой

рентгеновской диагностике остеопороза в поле съемки наряду с диагностируемым объектом включается эталонный объект с нормированной минеральной плотностью. Сравнивая яркости диагностического и эталонного объектов и зная реальную плотность эталона, можно определить плотность ткани.

Строятся математические модели изучаемых объектов и процесса получения рентгеновского снимка, что позволяет выполнять аннотированную сегментацию изображения и проводить денситометрию в полуавтоматическом или автоматическом режиме, что является одним из этапов интеллектуального анализа диагностических снимков.

На основе разработанного аппаратно-программного комплекса возможно создание центров телемедицины, в которых помимо прочих услуг может проводиться массовая диагностика остеопороза с использованием цифровых рентгеновских систем общего назначения и компьютерных приставок к стандартным медицинским рентгеновским аппаратам.

Работа выполнена при поддержке программы Президиума РАН «Фундаментальные науки — медицине» и РФФИ, проекты № 07-07-00280, № 07-07-00313 и № 06-07-89303.

Литература

[1] *Olshevetz M. M., Ustinin M. N., Nikonov I. A.* Software for Digital Filmless Roentgenography // Pattern Recognition and Image Analysis. — 2006. — Т. 16. — С. 23–25.

[2] *Daubechies I.* Ten Lectures on Wavelets. — Philadelphia: SIAM, 1992. — 314 с.

Модели потоков работ

Осипов Г. С.
gos@isa.ru
Москва, Институт системного анализа РАН

В наши дни эффективность работы каждой организации связана с эффективностью управления её ресурсами и процессами.

В большинстве случаев действующие организации не имеют исчерпывающих описаний всех реализуемых ими процессов, поэтому на первый план выходят задачи построения таких описаний на основе примеров или, как иногда говорят, рабочих последовательностей.

Каждая рабочая последовательность представима в виде графа, в вершинах которого находятся некоторые работы (операции, мероприятия), а ребра определяют порядок выполнения работ. Как оказалось, рабочие последовательности, даже решающие одну и ту же задачу, могут различаться как порядком выполнения, так и составом работ, поэтому

встает задача построения описания всего множества рабочих последовательностей, решающих одну и ту же задачу, т. е. такого графа, что граф каждой рабочей последовательности являлся бы его подграфом.

Для этой цели вводится понятие оператора переходов, т. е., по существу, правила (соответствующего некоторой работе), меняющего состояние процесса. Далее с помощью таких операторов определяются примеры или прецеденты потоков работ, тем самым уточняется понятие рабочей последовательности. Далее строятся описания классов эквивалентности прецедентов. Наконец, описания классов эквивалентности используются для синтеза модели потоков работ, которая, в свою очередь, может служить основой для реинжиниринга бизнес-процессов, оптимизации процессов по различным критериям и т. д.

Потоки работ и процессы

Пусть U — множество слов конечной длины над некоторым алфавитом. Зададим на U семейство алгебраических систем с сигнатурами, включающими одно-, двух- и n-местные отношения на U: $P^1, \ldots, P^m$. Для простоты будем полагать, что в сигнатуры входит ровно по одному отношению каждой местности. Каждую такую алгебраическую систему будем называть *состоянием* и обозначать через $\mathbf{s}$. Множество всех состояний обозначим через $\mathbf{E}$. Элементы многоместных отношений будем далее называть *фактами*, элементы одноместных отношений — *признаками*. Если N — дискретное линейно упорядоченное множество, то семейство отображений $\mathbf{O} = \{o_i\},\ i = 1, \ldots, M,\ \mathbf{O}\colon \mathbf{E} \times N \to \mathbf{E}$, таких что $o(s, n) = (s, n + 1)$, где $(s, n) = \langle (z, n), (p, n) \rangle$ — состояние системы в точке n, $z \subseteq P^i,\ i = 2, \ldots, m$ — множество фактов, $p \subseteq P^1$ — множество признаков, будем называть *множеством операторов переходов*. Далее множество N будем называть временем (дискретным), а для (s, n), (z, n), (p, n) используем более привычные обозначения: $s(n)$ либо s_n и т. д.

Если $\pi \subseteq P^1,\ \varphi \subseteq P^2 \cup \cdots \cup P^m$, то оператор $\mathbf{o} \in O$ имеет вид: $\mathbf{o} = \langle \pi, \varphi \rangle$, и $s(n + 1) = \mathbf{o}\, s(n)$, где $\mathbf{o}\, s(n) = \langle z(n + 1),\ p(n + 1) \rangle$, $z(n + 1) = z(n) \cup \varphi$ либо $z(n+1) = \varphi,\ p(n+1) = p(n) \cup \pi$ либо $p(n+1) = \pi$.

Два разных способа действия оператора переходов соответствуют двум различным способам формирования нового состояния: появлению в нем новых фактов и признаков при сохранении имеющихся, либо исчезновению старых признаков и фактов и появлению новых.

Если $\Omega(\mathbf{O})$ — семейство последовательностей вида $\omega = \langle \mathbf{o}_i, \mathbf{o}_j, \ldots, \mathbf{o}_k \rangle$ над множеством $\mathbf{O}$ операторов $\mathbf{o}_j$, где $i, j, \ldots, k$ — элементы множества натуральных чисел $\mathbb{N}$ и $i < j < \ldots < k$, то каждую последовательность ω будем называть *прецедентом* или *примером потока работ*. На $\Omega(\mathbf{O})$ зададим отношение эквивалентности ρ, порождающее фактор-множество Ω_ρ множества $\Omega(\mathbf{O})$.

Описанием $G(\{\omega\})$ каждого класса эквивалентности $\{\omega\} \in \Omega_\rho$ будем называть граф, такой что маршруты, порожденные всеми примерами $\omega \in \{\omega\}$ являются его подграфами.

Классы эквивалентности потоков работ и их построение

Внутри каждого из классов могут оказаться примеры, отличающиеся от других порядком следования операторов, порядком следования их групп, степенью их повторяемости и др. Эти различия приводят к появлению так называемых маршрутов: последовательного, параллельного, конкурентного, итеративного и условного в потоках работ [1] и необходимости представления таких маршрутов в $G(\{\omega\})$. Первые два вида маршрута были описаны в [2]. В клинической медицине, например, в некоторых случаях допускаются различные последовательности лечебных мероприятий для лечения одной нозологической формы.

На множестве матриц инцидентности введем ассоциативную и коммутативную операцию покомпонентного сложения, сохраняющую единицу: если $A = [a_{ij}]$ и $B = [b_{ij}]$ — матрицы инцидентности, то $A + B = [c_{ij}]$, где $c_{ij} = \max\{a_{ij}, b_{ij}\}$.

Пусть $M(\omega_j)$ — матрица инцидентности графа примера ω_j, через $M(G)$ обозначим матрицу инцидентности графа $G(\{\omega\})$.

Теорема 1. *$M(G) = \sum_j M(\omega_j)$, где суммирование в указанном выше смысле выполняется по всем примерам ω_j из класса $\{\omega\}$.*

Эта теорема обосновывает процедуру построения описания класса.

Следующая теорема показывает, что, какой бы пример мы ни взяли, если он принадлежит одному из классов, то он порождает хотя бы один из маршрутов, названных выше.

Теорема 2. *Для любого ω, если $\omega \in \{\omega\}$, то ω порождает хотя бы один из маршрутов в $G(\{\omega\})$.*

Если с каждым оператором связать условие его применимости и состояние, к которому он применяется, то получим понятие процесса над множеством дискретных событий: последовательность $\rho = \langle (s_i, \mathbf{c}_i, \mathbf{o}_i), (s_j, \mathbf{c}_j, \mathbf{o}_j), \ldots, (s_k, \mathbf{c}_k, \mathbf{o}_k) \rangle$ будем называть *примером процесса*, если для каждых двух её элементов $(s_n, \mathbf{c}_n, \mathbf{o}_n)$ и $(s_{n+1}, \mathbf{c}_{n+1}, \mathbf{o}_{n+1})$ справедливо $\mathbf{c}_{n+1} \in \mathbf{o}_n s_n$, где $\mathbf{o}_n s_n$ — результат применения оператора $\mathbf{o}_n$ к состоянию s_n, а $\mathbf{c}_{n+1}$ — условие применимости оператора $\mathbf{o}_{n+1}$ к состоянию s_{n+1}. Легко видеть, что множество примеров потоков работ изоморфно множеству примеров процессов относительно упорядочения, поэтому приведенные выше утверждения тривиальным образом переносятся на процессы.

Модель потоков работ

Моделью потоков работ будем называть динамическую систему $H = \langle X, N, \Psi \rangle$, где X — дискретное множество событий, N — линейно-упорядоченное дискретное множество, Ψ — функция переходов. Наибольший интерес здесь представляет восстановление функции переходов Ψ, так как множество X определено выше ($X = \mathbf{E}$), а в качестве N можно взять множество натуральных чисел. Что касается Ψ, то $\Psi\colon X \times N \to X$, так что для каждого состояния s_i $\Psi(s_i, n) = s_{i+1}$. Так как в принятом здесь представлении переходы реализуются операторами, то Ψ реализуется следующим алгоритмом:

1. Выбрать оператор, условие которого выполняется в текущем состоянии s_i;
2. Применить оператор, т. е. построить состояние $s_{i+1} = \mathbf{o}\, s_i$,
3. Перейти к пункту 1.

Это означает, что задача восстановления функции Ψ сводится к задаче восстановления операторов и условий их применения. Здесь надо заметить, что каждый из классов $G(\{\omega\})$ может содержать свое множество операторов, поэтому речь должна идти о восстановлении множества операторов каждого из классов и последующем объединении этих множеств. Для восстановления операторов следует для каждого оператора найти π и φ. Для этой цели будем сравнивать пары состояний s_i и s_{i+1}.

Случай 1: $s_i \cap s_{i+1} \neq \varnothing$, тогда $\varphi = z_{i+1} \backslash (z_{i+1} \cap z_i)$, $\pi = p_{i+1} \backslash (p_{i+1} \cap p_i)$. В частных случаях, когда имеет место включение в одну, либо в другую сторону, например $s_i \subseteq s_{i+1}$, получаем $\varphi = z_{i+1} \backslash z_i$, $\pi = p_{i+1} \backslash p_i$.

Случай 2: $s_i \cap s_{i+1} = \varnothing$, тогда $\varphi = z_{i+1}$, $\pi = p_{i+1}$.

Вершины графа $G(\{\omega\})$, в которых начинается любой маршрут, кроме последовательного, будем называть точками ветвления. В них определяются условия применимости операторов. Пусть ρ и ρ' — прецеденты, породившие параллельный маршрут в $G(\{\omega\})$, s^0 — состояние $G(\{\omega\})$ в точке ветвления, s и s' — следующие за s^0 состояния прецедентов ρ и ρ', соответственно. Если говорить неформально, то для определения условий применимости операторов $\mathbf{o}$ и $\mathbf{o}'$ следует сравнить состояния s и s' прецедентов ρ и ρ' и определить их различия, которые в простейшем случае и являются искомыми условиями применимости.

В более сложных случаях условия применимости операторов могут описываться логическими выражениями, например, в языке исчисления предикатов первого порядка. В этом случае условием оператора является конъюнкция атомарных формул языка исчисления предикатов первого порядка, каждая из которых интерпретируется одним из сигнатурных отношений и выполняется на элементах, различающих s и s'.

Заключение

В докладе кратко описан способ построения модели потоков работ по прецедентам. Метод позволяет построить описание технологических или бизнес-процессов сложной системы и, на этой основе, выполнить различные их преобразования, такие как уменьшение различных видов операций, уменьшение количества циклов и структурную оптимизацию.

Работа поддержана грантом РФФИ 06-07-89110.

Литература

[1] *Г. С. Осипов* Обнаружение и исследование потоков работ и процессов над множествами дискретных событий // Труды конференции «Системный анализ и информационные технологии 2005», М.: URSS. — 2005.

[2] *Laura Maruster, A. J. M. M. (Ton) Weijters, Wil M. P. van der Aalst, Antal van den Bosch* A rule-base approach for process discovery. Dealing with noise and imbalance in process logs // Data Mining and Knowledge Discovery, Springer-Verlag. — 2006. — 13. — P. 67–87.

Моделирование срока родов по данным сигнала наружного датчика вибраций

Переверзев-Орлов В. С.

peror@iitp.ru

Москва, ИППИ РАН им. А. А. Харкевича

Как ни странным это может показаться, но поведение матки в предродовом периоде оказалось малоизученным, и у специалистов пока нет единого мнения относительно того, как вообще осуществляется управление схватками.

Акушеры обычно строят свои заключения о характере схваток, воспринимая их ладонью, положенной на живот беременной. В случае предполагаемой патологии осуществляются более детальные исследования, в частности, с применением фетальных мониторов, в которых тензодатчик определяет напряжение брюшной стенки, ультразвук определяет частоту биений сердца плода, а специальная кнопка нажимается, когда беременная ощущает движения плода или схватки. Все это позволяет опытному врачу достаточно много узнать о характере схваток и состоянии плода и понять, нормально ли протекает процесс, или же в нём имеется какая-то патология.

С технологией такого рода связаны проблемы, так как патологические проявления могут возникать в любое время и вдали от медицинских центров, но беременной всё равно нужно правильно оценить ситуацию и спланировать соответствующие действия, а существующие мониторы громоздки, малоудобны и рассчитаны на использование медицин-

ским персоналом. В результате только в США это приводит ежегодно к потере примерно 250 миллионов долларов.

Наиболее важная задача, которую должен решать такой монитор — это определение типа возникающих схваток: являются ли они родовыми, предродовыми или же патологическими, предшествующими родовым.

Задача эта достаточно нетривиальна, особенно если за основу принимаются сигналы, считывемые вибродатчиком с поверхности тела беременной, поскольку схватки и мешающие их восприятию сигналы располагаются в общей низкочастотной области от примерно 0.001 до 100 Гц., причем мешающими являются движения беременной и плода и процессы в их органах и системах, пересекающиеся в значительной степени по спектрам и крайне резко различающиеся по мощности. Видимо, именно поэтому во всех известных нам работах исследователи ориентируются в первую очередь на анализ биопотенциалов на поверхности живота беременной.

Время от времени появляются и весьма пессимистические работы, подчеркивающие трудность задачи определения типов схваток, тем не менее, такая работа тоже идет.

Постановка задачи

Итак, наиболее интересной для акушеров задачей оказывается определение типа схваток, возникающих у беременной вне стен медицинского учреждения. То есть речь идёт в основном о преждевременных родах, особенно, когда схватки начались дома, и нужно принимать решение — ехать ли в клинику. Или же вызванная скорая помощь должна принять решение — брать ли пациентку в госпиталь или оставить дома, но с риском разродиться без медпомощи.

Рассматривая эту проблему с разных сторон и ориентируясь главным образом на удобство использования создаваемого монитора, мы остановились на варианте, когда в качестве входного сигнала будут использоваться вибрации передней стенки живота беременной, хотя, на первый взгляд, у такого решения больше минусов, чем плюсов.

Принимая это решение, мы прекрасно понимали, насколько усложняем свою задачу, так как вибродатчик воспринимает все вибрации передней стенки живота, обусловленные множеством источников, включая, помимо схваток, движения беременной, перистальтику ее кишечника, дыхательные движения, её сердцебиения, шевеления плода, его сердцебиения, и т. д. Уже первые исследования показали, что сигнал от движений беременной и схваток может в тысячу и более раз превосходить интенсивность сигнала от сердца плода, превосходя при этом и сигнал самих схваток. Спектральный диапазон смеси воспринимаемых вибродатчиком сигналов простирается при этом в диапазоне от примерно 0.001

до 50–100 Гц, и спектры сигналов, порождаемых различными источниками, в существенной мере пересекаются. В результате суммарный сигнал на выходе датчика очень похож на шумовой, в котором даже схватки далеко не всегда удается увидеть.

Это было кардинальное решение, но наш предыдущий опыт говорил, что такого рода попытка вполне оправдана и имеет реальные шансы на успех, если разумно распространить на эту задачу те технологии решения сложных задач в медицине, разработкой которых мы много лет занимались в рамках проекта «Партнерская система» (ПС). Главная проблема, возникающая при этом перед нами, оказывалась в том, что вибрационный сигнал можно было получать, а специалиста по интерпретации сигнала вибраций брюшной стенки не было. Поэтому нужно было найти такие преобразования суммарного вибрационного сигнала, которые позволили бы максимально подчеркнуть в нём то, что имело отношение к дифференциации типов схваток. Задача усложнялась тем, что анализ вёлся по мере накопления данных от различных типов датчиков, поскольку и здесь было абсолютно неясно, какой их тип может оказаться наиболее перспективным.

Помимо этого, нам нужно было развить уже существовавшие технологии ПС, согласуя их с задачей различения типов схваток и создания монитора для беременной. Главным при этом было то, что, понимая недостаточность существующих методов анализа и разделения сигналов в сложных их смесях, нужно было создать основу для направленного поиска такого рода преобразований на основе методов обучения распознаванию, приспособленных к работе с очень большими массивами данных. Параллельно нужно было совершенствовать и методы работы с клинической информацией о беременной, ориентируя их всё на те же перспективные возможности использования в малогабаритном мониторе.

Подходы к решению и первые результаты

Для сбора первичных данных было создано несколько прототипов монитора, включающих стандартный холтеровский монитор «АннА-3000» фирмы МедС и три типа вибродатчиков — пьезоэлектрический, входящий в комплект поставки «АннА», и несколько вариантов ионных сейсмодатчиков фирмы Р-сенсорс, предназначенных для тонких измерений линейных и крутильных ускорений. Использовавшиеся датчики являются прототипными разработками, не были сертифицированы, и не обладают пока воспроизводимостью своих характеристик при переходе от одного экземпляра к другому того же типа, чем наша задача дополнительно усложнялась. Всего с помощью сотрудничающих с нами врачей МОНИАГ было собрано более 120 записей от более чем 50 беременных.

Поиск преобразований входного сигнала в новые, обладающие смысловым сходством со схватками и их типами, вылился в чрезвычайно громоздкий и трудоемкий полунаправленный поиск таких преобразований сигнала и его спектра, которые визуально на имеющейся выборке с относительной устойчивостью демонстрировали бы желаемые результаты. Достаточно быстро был обнаружен принципиально важный феномен распространения спектра схваток в область высоких для этих сигналов частот (5–35 Гц), казалось бы, совершенно не типичных для схваток. Адекватного физиологического объяснения этому феномену в литературе обнаружить не удалось, но нам представляется вполне адекватным считать, что это является следствием спонтанно возникающих во время схватки коротких щелчков, порождаемых отдельными мышечными волокнами матки.

Следствием этого феномена оказалось то, что для анализа схваточного сигнала можно обойтись его спектральными компонентами, лежащими в области выше нескольких герц, что существенно упрощает, удешевляет и повышает качество анализа этого сигнала. Но при этом всё равно остаётся проблема восстановления по движениям спектра сигнала схваток того по форме сигнала, который регистрируется тензодатчиками фетальных мониторов. Это важно, так как врачи, привыкшие к использованию таких мониторов, предпочли бы и в новом мониторе иметь сходные по виду проявления, чтобы у них не возникало проблем при переходе с одного инструмента на другой. В связи с этим рассматривались различные преобразования сигнала и описания распределения и движения энергии по его спектру.

Было обнаружено, что для каждой беременной удается найти такой специфический для неё фильтр, энергия выходного сигнала которого с достаточной точностью воспроизводит форму сигнала с тензодатчика, обнаруживая при этом и некоторые тонкие детали, явно связанные с типом схваток. Таким образом, оказалось, что, с одной стороны, нет единого преобразования суммарного сигнала с датчиков, которое в равной степени было бы пригодно для требуемого описания схваток у всех беременных, но, с другой стороны, выяснилось, что нужные параметры фильтров образуют небольшой набор, допускающий поиск адекватного преобразования вполне разумным перебором. Иными словами, возникла ситуация, столь типичная при работе по формализации знаний врачей, когда наблюдаются некие качественно однородные проявления, и задача состоит в том, чтобы построить формальную модель, с достаточной точностью воспроизводящую такие проявления.

Наконец, было ясно, что помимо работы с сигналом по формированию его классификационных описаний, потребуется еще и учёт суще-

ственной, по мнению врачей, клинической информации, которая может оказывать критическое влияние на интерпретации сигналов. А поскольку речь шла о создании мобильного монитора, то возникала и задача переноса и доразвития ранее создававшихся нами технологий для Партнерских систем, чтобы всё это можно было бы применять на карманных компьютерах. Соответствующая работа была проведена, и в настоящее время мы располагаем вариантом технологии ПС, пригодным для применения как на настольных, так и на карманных компьютерах и коммуникаторах. Эта работа продолжается.

Для содержательного анализа данных и поиска формальных моделей схваточного сигнала в суммарном сигнале вибродатчика было решено ориентироваться на исследуемый нами уже более десяти лет «Синдромный анализ» (СА), принципиально обладающий как способностью к построению прозрачных моделей, что уже было доказано ранее, так и способностью к синтезу нелинейных фильтров, направляемому смысловой разметкой сигнала во времени. Было ясно, что ранее созданные образцы программ для СА не обладают достаточными ресурсными возможностями, чтобы справиться с теми потоками информации, которые предоставляются суммарным вибросигналом. Поэтому была предпринята разработка пока ещё полностью не завершённой версии новой программы синдромного анализа (ESA), которая при работе с тестовыми массивами объёмом порядка 40–60 MB показала вполне приемлемые результаты по скорости и качеству моделирования, так что теперь есть реальные основания надеяться, что уже к концу текущего года для различных типов схваток будут получены результаты моделирования на этой основе.

Работа выполняется при поддержке РФФИ, проект № 07-07-00407-а.

О выборе модели представления текстовой информации для задач анализа и фильтрации содержимого Интернет трафика

Петровский М. И., Глазкова В. В., Царёв Д. В.
Москва, МГУ им. М.В.Ломоносова

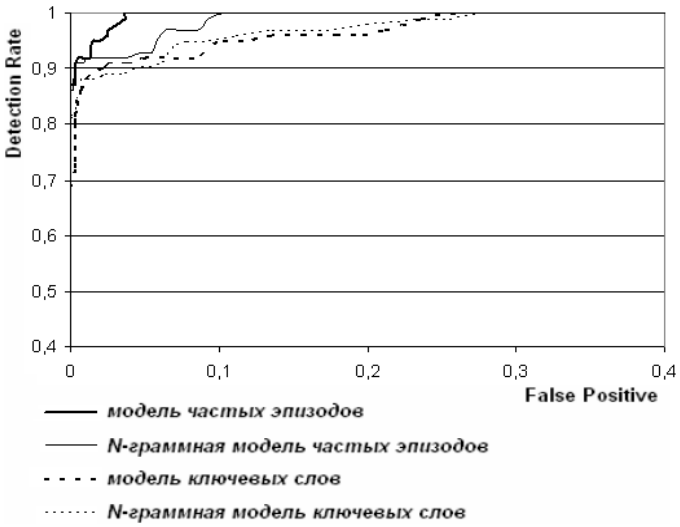
Задача классификации текстовых и гипертекстовых документов — одна из основных алгоритмических подзадач, возникающих при реализации систем фильтрации Интернет трафика, в частности, систем фильтрации спама и web-контента. Объекты классификации — текстовые и гипертекстовые документы и их фрагменты — являются слабо структурированными разнородными данными. Большинство алгоритмов классификации работают с формальным описанием объектов, используя векторную модель представления документа [1]. В данной модели документ

описывается числовым вектором фиксированной длины $\bar{a} \in \mathrm{R}^n$, где размерность вектора n есть число признаков, а i-я координата определяет вес i-ого признака. Соответственно, для реализации модели представления необходимо, во-первых, выбрать признаковое пространство, во-вторых, определить алгоритм вычисления весов. Качество выбранной модели представления при фиксированном алгоритме классификации и фиксированном тестовом наборе документов можно оценить по следующим критериям: точность классификации, размерность признакового пространства, размер получаемой модели классификации, время обучения и классификации, поддержка морфологии.

Самым распространенным способом формирования признакового пространства является *метод ключевых слов*, где в качестве признаков используются лексемы, входящие в документы, а размерность пространства равна размеру словаря. Но данный метод не учитывает морфологию языка. Поддержку морфологии можно реализовать с помощью stemming (все слова приводятся к своим базовым словоформам), что приводит к дополнительной вычислительной нагрузке. Кроме того, построение лексического анализатора для некоторых языков является достаточно сложной задачей. Более просто проблему морфологии решает разбиение лексем на N-граммы. В этом случае в качестве координат в признаковом пространстве рассматриваются все возможные подряд идущие буквосочетания фиксированной длины N. При этом однокоренные слова образуют сходный набор N-грамм.

Необходимо отметить, что основным недостатком обоих подходов является то, что семантические связи между лексемами (или N-граммами) не учитываются. Для преодоления этого недостатка нами был предложен метод построения модели представления, основанный на выделении *частых эпизодов*. В этом случае множество частых комбинаций лексем (или N-грамм) формирует новое признаковое пространство. Для этого на этапе обучения из каждого документа выделяются все предложения, входящие в его состав. Каждое предложение представляет собой отдельную транзакцию t, состоящую из лексем (или N-грамм) данного предложения. Весь тренировочный набор документов представляется в виде множества таких транзакций $\{t\}$. Далее с помощью алгоритма FP-tree [2] в $\{t\}$, выделяются частые эпизоды лексем (или N-грамм), которые удовлетворяют заданным параметрам — минимальной частоте встречаемости и максимальному размеру эпизода. Все полученные эпизоды нумеруются и составляют новое признаковое пространство.

Вес i-го признака определятся как нормированная частота встречаемости этого признака в документе: $a_i = f_i \,/\, \sqrt{\sum_{i=1}^{m} f_i^2}$, где f_i — частота встречаемости i-го признака в документе, m — количество непустых при-



знаков в данном документе. В отличие от традиционных мер сходства типа TF-IDF, предложенную меру можно использовать при дообучении, не пересчитывая веса, поскольку они зависят только от текущего документа, а не от всего набора.

Для сравнения моделей представления нами был использован эталонный тестовый набор документов SpamAssassin public corpus [3], который содержит как текстовые, так и гипертекстовые документы, относящиеся к одному из двух классов: спам и легальная почта. В качестве базового алгоритма классификации был выбран алгоритм на основе метода опорных векторов (SVM), как один из наиболее популярных и точных алгоритмов для классификации данных большой размерности. В качестве меры сходства мы использовали экспоненциальную потенциальную функцию. Результаты экспериментов представлены ниже в виде ROC-кривых для сравнения точности и сводной таблицы по всем основным критериям. Размер модели указан как число опорных векторов в построенной SVM модели. Время обучения для методов на основе частых эпизодов существенно больше, но для задачи фильтрации Интернет трафика это не принципиально, поскольку обучение может производиться в offline. Время классификации у всех алгоритмов получилось примерно одинаковым.

Из результатов эксперимента видно, что предложенная модель на основе частых эпизодов кардинально превосходит традиционные подходы по всем основным критериям. Кроме того, представление на основе ча-

	лексемы	N-gramm	эпизоды	эпизоды+N-gramm
Detection Rate	89%	85%	**92**%	91%
False Positive	1.2%	0.5%	0.5%	**0.3**%
Размерность	4227	5042	**3331**	5172
Размер модели	712	698	**610**	612

стых эпизодов с N-граммами можно использовать для морфологически сложных языков, таких как русский и немецкий.

Работа выполнена при поддержке гранта РФФИ №06-01-00691, гранта Президента РФ МК-4264.2007.9, а также в рамках госконтракта с Федеральным агентством по науке и инновациям №02.514.11.4026.

Литература

[1] *Salton G., McGill J.* An introduction to modern information retrieval. — New York: McGraw-Hill, 1983.

[2] *Jian Pei* Pattern-growth Methods for Frequent Pattern Mining. — Ph.D. Thesis, Simon fraser university, 2002.

[3] Apache Software Foundation. — The Apache SpamAssassin Public Corpus. — `http://spamassassin.apache.org/publiccorpus/`.

Исследование и разработка методов интеллектуального анализа данных для задач компьютерной безопасности

Петровский М. И., Машечкин И. В., Трошин С. В.

Москва, МГУ им. М.В. Ломоносова

Под вторжением в компьютерную систему понимается любая деятельность, нарушающая целостность, конфиденциальность или доступность данных. Для обнаружения вторжений используется специальное программное обеспечение — intrusion detection systems (IDS). В традиционных IDS применяется сигнатурный подход, при котором правила распознавания атаки задаются экспертом «вручную» и хранятся в периодически обновляемой базе знаний. У такого подхода есть ряд серьезных недостатков, в частности, он не устойчив к новым типам атак, поскольку базы знаний еще не содержат соответствующих сигнатур; кроме того, для распределенных и для «замаскированных» атак определение их сценария в виде экспертных правил является нетривиальной задачей. В связи с этим в настоящее время специалистами по компьютерной безопасности большое внимание уделяется применению интеллектуальных методов в IDS. Идея применения этих методов основывается на предположении о том, что активность пользователя или программы может быть отслежена, и на основе прецедентных данных с помощью методов

машинного обучения может быть построена либо модель нормального поведения (для подхода «обнаружения аномалий») либо модель распознавания атаки (для подхода «обнаружения нарушений»).

В рамках данного исследования ставились следующие задачи: разработать архитектуру распределенной интеллектуальной IDS; разработать интеллектуальные методы выявления вторжений, в основе которых лежит идея построения моделей поведения пользователей системы с целью обнаружения аномалий, а также распознавания следов вторжений.

В рамках указанных направлений получены следующие результаты. Был разработан экспериментальный прототип для сбора и анализа информации о поведении пользователей защищаемой компьютерной системы [1]. Он представляет собой мульти-агентную систему консолидации информации из системных журналов и лог-файлов защищаемой компьютерной системы и автоматизированное рабочее место аналитика безопасности (АРМ). Задачей системы консолидации является сбор, нормализация, предобработка и сохранение в едином хранилище в унифицированном XML-подобном представлении информации безопасности из различных источников. АРМ аналитика безопасности позволяет проводить анализ собранных данных с помощью статистических и интеллектуальных методов на предмет выявления следов вторжений, а также для построения моделей поведения пользователей. В состав АРМ включены стандартные методы оперативного статистического анализа данных об активности пользователей с использованием технологии OLAP (online analytical processing), а также реализованы следующие алгоритмы интеллектуального анализа данных.

1. *Алгоритм обнаружения аномалий в разнородных структурированных данных* на основе ассоциативных правил [1]. Идея алгоритма базируется на том, что ассоциативные правила, описывающие корреляции между атрибутами событий, можно использовать для прогнозирования значений одних атрибутов по значениям других. Для этого на основе найденной системы правил строится функция, которая вычисляет распределение условной вероятности значений некоторого атрибута в зависимости от значений остальных атрибутов. В таком случае уровень «ожидаемости» (нормальности) значения этого атрибута события вычисляется как отношение условной вероятности реально наблюдаемого значения к условной вероятности наиболее ожидаемого значения. Аномальность всего события определяется как свертка значений аномальностей всех атрибутов. Такой подход дает возможность не только обнаружить аномальные события, но «интерпретировать» причину их аномальности, т. е. выявлять те атрибуты, которые являются ненормальными с точки зрения предыдущей активности пользователей.

2. *Алгоритм построения вероятностной модели поведения пользователей* позволяет прогнозировать следующее действие пользователя по последовательности предыдущих действий [2]. Он основан на использовании метода потенциальных функций для отображения последовательности событий в конечномерное вещественное пространство признаков, в котором применяется алгоритм построения деревьев решений типа CART для прогнозирования следующего события, при условии, что предыдущая активность описывается заданной последовательностью. В отличие от большинства существующих алгоритмов такой подход учитывает временные интервалы между событиями, а не только порядок событий, и одновременно допускает представление результирующей модели в понятном эксперту виде, в частности, в виде набора правил вида «ЕСЛИ ... ТО».

3. *Алгоритмы анализа «сырого» сетевого трафика* (TCP/IP), основанные на комбинации методов потенциальных функций и теории нечетких множеств [3]. Алгоритм обнаружения аномалий в сетевом трафике базируется на методе поиска исключений, основанном на использовании потенциальных функций и вычислении нечеткой степени «типичности» объектов в анализируемой выборке. На базе данного метода поиска исключений и нечеткого метода опорных векторов (Fuzzy Support Vector Machines) разработан гибридный метод решения задачи бинарной классификации для больших объемов данных в условиях наличия шума, который применяется для обнаружения атак по сетевому трафику в режиме обнаружения нарушений. Разработанные алгоритмы проверены на эталонных тестовых наборах данных DARPA Intrusion Detection Evaluation Program и показали высокую точность по сравнению с существующими методами.

В рамках работ по данному направлению также была спроектирована и реализована *мульти-агентная интеллектуальная подсистема обнаружения и защиты от атак, осуществляемых через несанкционированную рассылку электронных сообщений* [4]. Это потребовало разработать специальные алгоритмы моделирования электронной переписки пользователей защищаемой компьютерной системы и фильтрации нежелательных электронных сообщений на уровне почтового сервера. Эти алгоритмы используют методы машинного обучения на основе опорных векторов, а также оригинальные методы сокращения тренировочного набора и уменьшения размерности пространства признаков, что позволяет применять их в режиме реального времени. Данные алгоритмы были успешно верифицированы на эталонных тестовых наборах данных SpamAssasin и LinqSpam.

Работа выполнена при поддержке РФФИ, проекты №05-01-00744 и №06-07-08035-офи.

Литература

[1] *Машечкин И. В., Петровский М. И., Трошин С. В.* Система мониторинга и анализа поведения пользователей компьютерной системы // Программные системы и инструменты. — 2006. — №7. — С. 95–113.

[2] *Petrovskiy M.* A Data Mining Approach to Learning Probabilistic User Behavior Models from Database Access Log // Proc. of ICSOFT, Portugal, 2006. — V. 2. — Pp. 73–79.

[3] *Петровский М. И.* Применение методов интеллектуального анализа данных в задачах выявления компьютерных вторжений // Труды конф. «Методы и средства обработки информации», Москва, 2005. — С. 158–167.

[4] *Mashechkin I., Petrovskiy M., Rozinkin A.* Enterprise Anti-spam Solution Based on Machine Learning Approach // Proc. of 7th Internat. Conf. on Enterprise Information Systems, USA, 2005. — V. 2. — Pp. 188–193.

Математические методы атрибуции литературных текстов небольшого объема

Рогов А. А., Сидоров Ю. В., Суровцова Т. Г.
rogov@psu.karelia.ru
Петрозаводск, Петрозавоский государственный университет

Изучение литературных произведений с использованием математических методов имеет богатую историю, а появление компьютеров расширило возможности проведения вычислительных экспериментов. Основной целью работы является поиск методов, которые помогут «оценить» стиль литературного текста, выявить закономерности, присущие разным жанрам, авторам, произведениям. Для этого необходимо на едином тестовом материале исследовать методы на надежность и устойчивость.

Материал для исследования

Основой для проведения исследования стала электронная коллекция публицистических статей из петербургских журналов «Время», «Эпоха», «Современник», «Гражданин», и других текстов 60–70-х гг. XIX века в оригинальной орфографии дореволюционной России [2, 3], которая создается в Петрозаводском государственном университете, начиная с 1995 года, включающая синтаксический и морфологический разборы произведений. Богатый материал дает основу для анализа авторского стиля. Причем можно рассматривать не только легко рассчитываемые признаки (длина предложения, средняя длина слова и т. д.), но и более сложные, описывающие грамматику текста. Что позволяет более полно изучать произведения, размер которых не достаточно велик.

Методы и результаты

Был проведен поиск методов, описанных в литературе, которые использовались для атрибуции текстов [1, 4], и их проверка на нашем материале. Предложены модификации методов, а также разработаны собственные подходы [3] и их приложение к атрибуции текстов небольшого объема. Наиболее интересные, с нашей точки зрения, вошли в экспертную систему, входящую в программный комплекс «Статистические методы анализа литературных текстов» (ПК СМАЛТ). В частности для проведения анализа в экспертной системе предлагаются следующие группы методов:

— разбиение анализируемых текстов на однородные группы с близким набором грамматических признаков;
— проверка статистических гипотез об однородности распределения частотных характеристик текстов, таких как распределение частей речи на разных позициях предложения, индекс разнообразия лексики и т. д.;
— метод «сильного графа» для оценки парной связи грамматических классов.

Наборы признаков для анализа, которые можно получить на основе грамматических разборов произведений, очень разнообразны. Сложно определить те, которые описывают авторский стиль, поэтому предусмотрена возможность проводить исследования на наборе морфологических и синтаксических признаков, который определяет самостоятельно специалист-филолог. Выбираются и методы, которые должны быть применены. С использованием экспертной системы ПК СМАЛТ был проведен ряд исследований. Одно из них по поиску авторского инварианта — некоторых особенностей текста, неосознанное предпочтение которым отдает автор при создании своих произведений, описано ниже. Были выбраны публицистические произведения, объемом от 6 до 700 предложений. Проведем краткое описание методов и результатов.

Анализ частоты синтаксических конструкций. Параметры, выбранные для анализа синтаксического разбора, представляют собой относительную встречаемость определенной синтаксической конструкции в тексте (тип предложения, осложненность, наличие второстепенных членов и т. д.). Для получения групп произведений использовался алгоритм иерархического кластерного анализа (методы ближайшего и дальнего соседа с евклидовым расстоянием и расстоянием Чебышева) [3]. Группы объектов, которые получились в результате, не дали четкого разделения по авторам произведений, причем группировка менялась в зависимости

от выбранной метрики и метода построения кластеров. Выбранный для анализа набор признаков оказался неустойчивым.

Метод «сильного графа» основан на определении пороговых значений [1], которые позволяют оставить устойчивые связи, отбрасывая более редкие, как менее значимые. Получаемые в результате графы сравниваются для обнаружения близости между текстами. Была рассмотрена связь между синтаксическими структурами текста, выделены классы предложений: простое односоставное, сложное бессоюзное предложение и т. д. Изучались переходы между классами, которые присутствуют в текстах, строились «сильные графы» с различными пороговыми значениями. В результате был сделан вывод о том, что структура графа в первую очередь зависит от длины рассматриваемого текста. С помощью данного метода не удалось получить достаточно устойчивые результаты, то есть даже при небольших изменениях пороговых параметров, матрица близости текстов сильно меняла свой вид.

Проверка статистических гипотез об однородности распределения частотных характеристик текстов. Были использованы методы, приведенные в исследованиях Г. Хетсо [5]. Эти методы основаны на проверке статистических гипотез о значимости различий рассчитываемых параметров для сравниваемых авторов. Из числа рассматриваемых текстов исключаются те, которые имеют значение критерия, попадающее в критическую область. Полученные результаты согласуются с результатами, полученными Г. Хетсо, о возможной принадлежности некоторых произведений Dubia (спорное авторство) перу Ф. М. Достоевского, не смотря на то, что мы работали с текстами в оригинальной орфографии XIX века и отличными морфологическими разборами текстов. Что говорит об устойчивости методов и о выявленных реальных отличиях или сходстве.

Заключение

Использование ПК СМАЛТ позволяет проводить проверку методов на разных наборах грамматических признаков. Так некоторые из методов оказались неустойчивыми для выбранных нами наборов признаков. Происходит поиск новых методов, которые могли бы работать с произведениями небольшого объема. В настоящее время разрабатывается комплексный подход к проводимым экспериментам, который учитывал бы результаты, показанные каждым из методов. Все полученные данные предложены для рассмотрения специалистам-филологам, которые исследуют творчество Ф. М. Достоевского.

Проект поддержан грантами РГНФ № 02-04-12015в, № 05-04-12418в.

Адрес проекта в Интернете: `http://smalt.karelia.ru`.

Литература

[1] *Бородкин Л. И., Милов Л. В., Морозова Л. Е.* К вопросу о формальном анализе авторских особенностей стиля в произведениях Древней Руси // Математические методы в историко-экономических и историко-культурных исследованиях. — Москва, 1977. — С. 298–326.

[2] *Захаров В. Н., Леонтьев А. А., Рогов А. А., Сидоров Ю. В.* Программная система поддержки атрибуции текстов статей Ф. М. Достоевского // Труды Петрозаводского государственного университета: Сер. Прикладная математика и информатика. Вып. 9. — Петрозаводск: ПетрГУ, 2000. — С. 180–189.

[3] *Rogov A. A., Sidorov Yu. Vl.* Statistical and Information-calculating Support of the Authorship Attribution of the Literary Works // 6th Int. Conf. Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods, Vol. 2: K-S. — Minsk: BSU, 2001. — Pp. 187–192.

[4] *Хетсо Г.* Принадлежность Достоевскому: к вопросу об атрибуции Ф. М. Достоевскому анонимных статей в журналах Время и Эпоха // Oslo: Solum Forlag A. S., 1986. — 82 с.

Задачи анализа изображений в информационно-поисковой системе PIRS

Рогова К. А., Быстров М. Ю.

ksushar@mail.ru

Петрозаводск, Петрозаводский государственный университет

Петроглифы Карелии — ценнейший памятник первобытной эпохи, получивший мировую известность. Рисунки на скалах постоянно подвергаются природным и человеческим воздействиям [1]. Информационно-поисковая система PIRS создана с целью сохранения и доступности для исследователей и рядовых пользователей информации о петроглифах Карелии. Кроме этого, электронная информация об изображениях позволяет проводить новые исследования с использованием математических методов и компьютерных технологий.

Информационно-поисковая система PIRS

Информационно-поисковая система PIRS состоит из четырех блоков: базы данных петроглифов, модуля подготовки изображений для базы данных, модуля онлайнового доступа и модуля локального доступа к базе данных [2,3].

База данных является наиболее важным блоком системы. В нее входят графитные копии, фотографии, карты, черно-белые схемы, характеристики и тестовые описания петроглифов. Выделены: группы петроглифов по местонахождению, сюжетные группы и отдельные петроглифы. Текстовое описание и карта местости сопровождает каждую группу

изображений по местонахождению. Сюжетные группы описаны по следующим параметрам: название, кодовый номер, общая площадь и текстовая информация. Кроме того, для каждой сюжетной группы представлены фотографии и графитные копии (в среднем более чем по 3 для каждой). Отдельные петроглифы описаны по следующим признакам: кодовый номер, название, дистанция, высота над уровнем моря, глубина вырезки, обрастание мхом, сохранность, угол поворота. Так же присутствуют фотографии и графитные копии каждого петроглифа (более 10 для каждого). База данных зарегистрирована в государственном регистре баз данных (№ 0220611248).

Модуль подготовки изображений для базы данных включает в себя подпрограммы, реализующие решение следующих задач: выделение отдельных изображений из группы, сегментация изображений, приведение изображений к единому стандарту, нанесение защитных надписей на фотографии.

Сегментация изображений производится пороговым методом и методом градиентного спуска. Позволяет выделить группы петроглифов на фотографии. На входе и выходе — графические файлы.

Для подпрограммы выделения отдельных изображений на вход подается черно-белая копия группы, а на выходе — графические файлы с отдельными изображениями. При этом возникает следующая проблема: если петроглифы (разные) соединены линией, то они воспринимаются как одно изображение. В этом случае необходимо разделение вручную. Заметим, что таких изображений не больше 5%.

Нанесение защитных надписей на фотографии необходимо для фотографий, представленных на сайте для защиты авторских прав. Защитой является адрес сайта на середине фотографии.

Модуль онлайнового доступа через Интернет реализован в виде раздела «Каталог» сайта «Петроглифы Карелии». Поиск происходит по признаку «местонахождение петроглифа» путем перемещения с помощью гиперссылок по картам. В настоящий момент на сайте находится более 500 фотографий групп петроглифов с их описаниями. Общее количество представленных петроглифов превышает 2000 фигур. Адрес сайта: `http://smalt.karelia.ru/~petroglyphs`.

Модуль локального доступа к базе данных состоит из следующих подпрограмм: поиска по признакам и поиска по изображениям. Рассмотрим их подробнее.

Поиск по признакам изображений реализован на примере петроглифов лосей и оленей Беломорья. Было выделено 16 признаков. Примерами признаков являются: тип головы, тип корпуса, наличие/отсутствие

холки, изгиб передней и задней пар ног и т. д. Все признаки были проанализированы на статистическую зависимость с помощью критерия χ^2 Пирсона. Далее все петроглифы были разбиты на несколько групп по степени сходства, с выделением типичных лосей и типичных оленей при помощи иерархического кластерного анализа.

Общая схема работы с данным поиском следующая: пользователю предлагается выбрать значения признаков, точность поиска (количество совпадений признаков). После выбора всех условий поиска будут найдены петроглифы, соответствующие выделенным критериям. Пользователю доступны изображения найденных петроглифов и полная информация о них.

Поиск по изображениям предназначен для поиска изображений, похожих на данное. На вход подается исследуемое изображение, а на выходе должны появиться изображения из базы данных, наиболее похожие на исходное.

На сегодняшний день все новейшие материалы по петроглифам Карелии представляют собой набор цветных фотографий. Определенную сложность поиска создает фактическое отсутствие некоторых частей изображения. Поиск также осложняется тем, что часто невозможно определить, где верх, а где низ изображения. При этом, требование, что при поиске необходимо только совпадение контура изображения, позволяет упростить поиск, а значит, изображение петроглифа можно рассматривать, как бинарное (скале соответствует белый цвет, а петроглифу — черный). В зависимости от выбранных параметров поиска (точность поиска, процент совпадений элементов изображений) будет найдено одно или несколько изображений. Для поиска используется сеть адаптивного резонанса.

Пользователю предоставляется доступ к информации о кодовом номере, месторасположении, характерных признаках найденного петроглифа и петроглифах, близких к нему по ранее описанным признакам.

В настоящее время проводится тестирование изложенных выше методов поиска и разработка новых.

Работа выполнена при поддержке РГНФ, проект № 05-01-12118в.

Литература

[1] *Савватеев Ю. А.* Залавруга. — Ленинград: Наука, 1970. — 250 с.

[2] *Рогова К. А.* Информационно-поисковая система «Петроглифы Карелии» // Интеллектуальные системы и компьютерные науки. — Москва: ЦПИ при мех.-мат. факультете МГУ, 2006. — С. 262–264.

[3] *Рогов А. А., Рогова К. А.* Возможности современных информационных технологий при изучении петроглифов Беломорья. // Сборник научных статей и докладов. — Архангельск: СОЛТИ, 2006. — С. 473–479.

Проектирование новых материалов с заданными свойствами и оптимизация существующих технологий их изготовления с помощью систем интеллектуального анализа данных

Саакян Р. Р., Шпехт И. А., Яхшибекян М. Р.
rsahakyan@yahoo.com,mish9@rambler.ru
Анапа, филиал Российского государственного социального университета
Ванадзор, Армянский государственный инженерный университет

Требования современного рынка (как промышленной, так и потребительской продукции) ставят перед предприятиями-производителями задачи разработки и производства качественно новой продукции, соответствующей современным требованиям инновационной составляющей на этапе их разработки.

Решение указанной задачи тесно связано с классификацией и прогнозированием поведения существующих материалов с целью оптимизации существующих технологий изготовления продукции.

Одним из направлений решения вышеуказанных задач является разработка прикладных систем интеллектуального анализа экспериментальных данных (для конкретных предметных областей исследования) и создание средств поддержки вычислительных экспериментов.

Внедрение в производство наукоемких информационных технологий нового поколения делает реальной возможность использования автоматизированной системы проектирования технологических режимов не только при проведении научных исследований, но и непосредственно в производстве. Это значительно уменьшит время на освоение нового ассортимента продукций, позволит избежать многочисленных ошибок при организации производства нового ассортимента.

Постановка задачи

Целью настоящей работы является формирование общей методологии расширения приложений систем интеллектуального анализа данных применительно к проектированию новых материалов с заданными свойствами и оптимизации существующих технологий их изготовления. При этом содержание указанного расширения относится ко всем этапам методологии:

— к обработке первоначальной экспериментальной информации при опоре на информационно-ориентирующую схему для малого объема требуемых прецедентов;
— к составлению аналитических представлений закономерностей (АПЗ) изменения переменных типа «эксплуатационные (потребительские) свойства материала — параметры технологического процесса».

Проблема проектирования новых материалов с заданными свойствами и оптимизации существующих технологий их изготовления выдвигает на первый план решение следующих задач:

1. Исследование и моделирование существующих процессов изготовления (формирования) и изменения (при хранении и эксплуатации) прочностных характеристик материалов.
2. Обобщение и систематизация имеющихся экспериментальных данных (поведения потребительских свойств готовой продукции в зависимости от изменения технологических параметров и человеческого фактора, изменение структуры материалов под воздействием внешних эксплуатационных факторов) и создание на их основе классификационно-ориентирующих баз экспериментальных данных (КОБЭД).
3. Прогнозирование поведения готового изделия в зависимости от основных структурных параметров и технологических особенностей процесса изготовления с использованием КОБЭД.
4. Оценка изменения структурных и прочностных характеристик при проектировании новых материалов с заданными свойствами на основе КОБЭД.

Применение систем интеллектуального анализа данных для составления АПЗ при обработке полученных экспериментальных результатов позволяет учесть влияние разнообразных факторов, не увеличивая количества дорогостоящих экспериментов.

Методы решения

В работе для решения задач классификации, распознавания и прогнозирования свойств и характеристик исследуемых объектов на основе формируемых КОБЭД при составлении АПЗ используется расширенный метод линейных направлений — согласующих функций (ЛН-СФ,Р).

Метод ЛН-СФ,Р позволяет решать прямую задачу для функции многих переменных — составление АПЗ — связи выходных переменных с входными переменными по известным описаниям искомой функции по линейным направлениям в рассматриваемой области входных переменных.

В работе предлагается для совершенствования АПЗ опираться на описанные в литературе ансамбли функций, из которых наиболее подходящие варианты для аппроксимируемых функций выбираются варьированием (подбором) числовых значений параметров обозначенных ансамблей. Здесь в первую очередь рекомендуется воспользоваться ансамблями ортогональных базисов (базисами Чебышева и Лагерра), при этом не исключаются и базисы полиномиального варианта.

На первом этапе в известной области описания признаков объектов решается задача определения аналитических представлений закономерностей по методу ЛН-СФ [1] на основе малого объема экспериментальных данных.

На втором этапе на основе расчетных значений функции в узловых точках, которые несут в себе информацию о поведении функции также в промежуточных точках сетки, при опоре на использование ортогональных аппроксимаций изменения основных переменных [2], определяется описание функции по границам области изменения признаков объектов (метода ЛН-СФ,Р). Далее, на основе указанных описаний функции по границам, получают уточняющее описание во всей области определения признаков. Для коррекции результатов полученного описания используются описания функции по промежуточным линейным направлениям, определенные на первом этапе (проверочная выборка линейных направлений).

Заключение

На основе представленной методологии решения прямой задачи можно представить решение обратной задача методом ЛН-СФ,Р, когда в результате пассивного эксперимента или математических расчетов (численный эксперимент), получены значения искомой функции (прецеденты) в промежуточных точках ячейки локальной области входных переменных, которые в общем случае не являются узловыми. Неизвестными в этом случае являются коэффициенты разложения по базису, которые находятся методом наименьших квадратов.

Полученные на основе использования ортогональных аппроксимаций результаты можно использовать для сглаживания описания внутри области определения выходных признаков (на стыках ячеек), а также при расширении области прогноза поведения искомой функции.

На основе АПЗ, полученных при решении обратной задачи, можно так же осуществлять прогнозирование значений искомой функции для любого сочетания значений входных переменных.

Литература

[1] *Саакян Р. Р.* Неклассические информационные технологии в управлении машинными агрегатами и производственными технологиями. — Благовещенск: АмГУ, 2004.— 216 с.

[2] *Дедус Ф. Ф., Махортых С. А., Устинин М. Н., Дедус А. Ф.* Обобщенный спектрально-аналитический метод обработки информационных массивов.— М.: Машиностроение, 1999.— 356 с.

Применение вероятностного алгоритма фильтрации в задачах обработки данных телеметрии космических спутников

Серостанов А. С., Ветров Д. П., Кропотов Д. А.
serostanov@gmail.com, vetrovd@yandex.ru, dkropotov@yandex.ru
Москва, ВМиК МГУ, ВЦ РАН

Одной из наиболее важных задач, возникающих при цифровой обработке сигналов, является задача их фильтрации — выявления и устранения помех с минимальными изменениями полезной части. Существует множество подходов к решению данной задачи.

Самый простой предполагает использование низкочастотных фильтров, которые просты в реализации и показывают отличные результаты в скорости работы. Однако алгоритмы подобного оказывают влияние не только на помехи, но искажают и полезную часть сигнала. Это обусловлено тем, что низкочастотные фильтры основаны на том или ином сглаживании сигнала, а также используют априорные предположения о модели шума.

В силу вышесказанного, алгоритмы низкочастотной фильтрации нельзя использовать в задачах, в которых невозможны априорные соображения о виде шумов (например, они возникают в результате сбоев в измерении и передаче информации). Следовательно, необходимо использовать фильтры, которые не зависят от модели шума. Примером такого алгоритма может служить метод, основанный на вероятностной фильтрации.

Особенностью этого подхода является то, что зашумленные точки не обрабатываются, а исключаются из рассмотрения, а полученные таким образом пропуски заполняются путем линейной интерполяции.

Постановка задачи

Телеметрия представляет собой сбор данных о шести показателях по времени. В процессе измерений и передачи информации могут происходить сбои, в результате чего данные искажаются, и появляются шумы. Требуется построить фильтр, устраняющий эти выбросы с сохранением полезной формы сигнала.

Алгоритм максимизации правдоподобия

Рассмотрим следующую задачу. Пусть имеется зашумленный сигнал $x[i]$, представленный моментами времени $t_1, \ldots, t_T$. Необходимо построить фильтр, формирующий на выходе сигнал $y[i]$ таким образом, чтобы фильтрованный сигнал не содержал в себе существенных неоднородностей, связанных с ошибками в измерении данных. В случае отсут-

ствия подобных неоднородностей сигнал не должен подвергаться изменениям, т. е. $y[i] = x[i]$.

Рассмотрим последовательную процедуру фильтрации сигнала. Предположим, что имеется некоторая точка $x[i_0]$, про которую известно, что она является первой точкой полезного сигнала, т. е. $y[i] \neq x[i]$ при $1 \leqslant i < i_0$; $y[i_0] = x[i_0]$. Далее, определим меру правдоподобия принадлежности точки полезному сигналу при условии, что точка $x[i_0]$ является незашумленной, следующим образом :

$$l(i, i_0) = -\frac{\big(x[i] - x[i_0] - M_{i,i_0}\big)^2}{2\Sigma_{i,i_0}^2} - \ln\big(\sqrt{2\pi}\Sigma_{i,i_0}\big), \tag{1}$$

где M_{i,i_0} и Σ_{i,i_0} выражаются формулами через так называемые центроидное и вариационное поля соответственно :

$$M_{i,i_0} = \textstyle\sum_{k=i_0}^{i} \mu[k], \qquad \Sigma_{i,i_0} = \sqrt{\textstyle\sum_{k=i_0}^{i} \sigma^2[k]}.$$

При фиксировании очередной полезной точки сигнала следующая точка находится путем максимизации функции правдоподобия (1). Совокупность построенных таким образом точек назовем *допустимыми*. Остальные точки $x[i]$ рассматриваются как помехи и игнорируются. Таким образом, полезный сигнал $y[i]$ представляет собой последовательность допустимых точек, в промежутках между которыми сигнал достраивается путем интерполяции.

Для реализации указанного алгоритма фильтрации необходимо оценить центроидное и вариационное поля. Центроидное поле $\mu[i]$ можно рассматривать как ожидание относительно тенденции изменения значения полезного сигнала на том или ином участке, а вариационное поле, в свою очередь, характеризует предположения о локальной мере изменчивости полезного сигнала в точке. Для определения значений $\mu[i]$ и $\sigma[i]$ можно воспользоваться априорными соображениями о приблизительной форме сигнала или использовать предварительную обработку низкочастотными фильтрами, например, скользящим средним. Пусть $z[i]$ — сигнал, полученный путем сглаживания исходного. Тогда центроидное и вариационное поля можно определить по формулам:

$$\begin{aligned} \mu[i] &= z[i+1] - z[i], \quad i = 1, \ldots, T-1; \\ \mu[T] &= 0; \\ \sigma[i] &= \lambda + \delta|\mu[i]|, \quad i = 1, \ldots, T. \end{aligned}$$

Здесь λ и δ — действительные параметры, которые подбираются с учетом конкретного вида сигнала и уровня шумов.

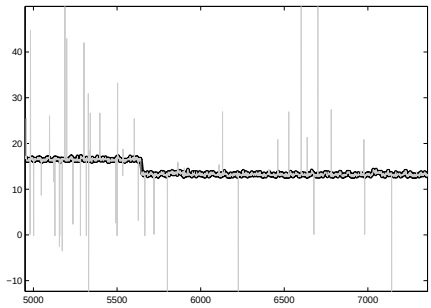


Рис. 1. Пример фильтрации с помощью разработанного алгоритма

Эвристические модификации алгоритма

Постановка задачи обработки данных телеметрии предполагает введение определённых ограничений. Так, уровень зашумленности сигнала позволяет предположить, что не имеет смысла для каждой точки просматривать весь сигнал до конца, так как следующая точка полезного сигнала будет находиться рядом с исследуемой. Значения λ и δ подбираются исходя из априорных соображений о структуре анализируемого сигнала.

Заключение

Метод вероятностной фильтрации, основанный на максимизации правдоподобия, является мощным инструментом для решения задач цифровой обработки сигналов. Относительно несложный в реализации, он способен показывать хорошие результаты при правильном подборе параметров. Исследования показали, что алгоритм хорошо решает задачи с единичными выбровами, даже если сам сигнал достаточно сложен.

Литература

[1] *Vetrov D. P., Kropotov D. A.* Application of Probabilistic Filter to Signal Filtartion Tasks // Pattern Recognition and Image Analysis. — 2006. — Vol. 16, № 3. — P. 478–485.

Прогнозирование результатов хирургического лечения атеросклероза на основе анализа клинических и иммунологических данных

Кузнецов М. Р., Туркин П. Ю., Воронцов К. В., Дьяконов А. Г., Ивахненко А. А., Сиваченко Е. А.
pavelturkin@km.ru, voron@ccas.ru
Москва, Российский государственный медицинский университет;
Москва, Вычислительный центр РАН

Проблема облитерирующего атеросклероза артерий считается одной из наиболее важных в современной медицинской науке. Проявлениями данной патологии являются такие социально значимые заболевания, как ишемическая болезнь сердца, острый инфаркт миокарда, острое и хроническое нарушение мозгового кровообращения и др. В возрасте 50 лет ими в той или иной форме страдают до 90%, а в возрасте свыше 60 лет — 100% населения. В структуре смертности атеросклероз и его проявления занимают основное место, являясь причиной 80% летальных исходов. Тем не менее, до сих пор не выявлены ключевые механизмы возникновения атеросклероза. Вызвано это отчасти тем, что фундаментальные и клинические изыскания в данной области практически не имеют точек соприкосновения. Клинические исследования в основном рассматривают вопросы терапии частных проявлений атеросклероза и практически не изучают основы патогенеза атеросклероза в целом.

Задача прогнозирования отдалённых последствий анастомоза

В данной работе решается задача прогнозирования результатов хирургического лечения атеросклероза, в ходе которого производится имплантации шунта с наложением соустий (анастомозов) между артерией и протезом [1]. Рассматриваются только двузначные прогнозы — либо шунт приживётся, либо нет, т. е. прогнозирование сводится к классификации на два класса.

Данная задача классификации характеризуется большой размерностью (число признаков превышает число прецедентов), разнотипностью признаков, наличием пропусков в данных, низкой точностью измерения признаков. Перечисленные особенности свойственны многим задачам медицинской диагностики. Отличительной особенностью данной задачи является то, что данные собираются из двух независимых источников:

— *Клинические данные* [2]: скорости кровотока в зоне анастомоза, степень сужения линии анастомоза, свёртываемость, вязкость, агрегация тромбоцитов и эритроцитов, и др. Всего 18 признаков, измеряемых сразу после операции, спустя неделю, 1, 2, 3 и 6 месяцев.

— *Данные иммунологического обследования*: показатели местного иммунного статуса в поражённой зоне, показатели гуморального иммунитета, цитокинового статуса и т. д. Всего 17 признаков, измеряемых до и после операции.

Совместный анализ этих данных позволяет впервые в практике изучения атеросклероза оценить, какие иммунные факторы и особенности клеточных реакций в действительности оказывают значимое влияние на течение атеросклероза.

Задачи анализа данных

С точки зрения анализа данных цели исследования заключаются в следующем:

— Построить алгоритм прогнозирования (классификации), обладающий достаточно высокой точностью предсказания.
— Выявить значимые, нетривиальные, интерпретируемые закономерности и взаимосвязи между данными клинических и иммунологических обследований.

Для достижения указанных целей не достаточно только лишь выбрать наиболее точный (для данной конкретной задачи) метод классификации. Выбранный метод должен суметь «извлечь максимум выгоды» из совмещения клинических и иммунологических данных. Таким образом, появляется дополнительный критерий выбора модели классификации: алгоритм, построенный по объединённому набору признаков должен заметно превосходить по качеству прогнозов алгоритмы, построенные только по клиническим или только по иммунологическим данным.

Вычислительный эксперимент

Исходная выборка данных, собранная в клинике факультетской хирургии РГМУ, содержала описания 72 случаев. Долгосрочный результат операции был благоприятным в 54 и неблагоприятным в 18 случаях.

Тестировались следующие методы классификации:

— MLP — классический двухслойный персептрон;
— RBF — нейронная сеть с радиальными базисными функциями;
— SVM — метод опорных векторов;
— SVM-f — SVM с предварительной фильтрацией объектов-выбросов;
— DT — решающие деревья.

Для всех методов, кроме DT, выполнялся предварительный отбор признаков. Сначала все признаки были проранжированы по индивидуальной разделяющей способности. Затем были найдены парные корреляции признаков и все признаки были проранжированы по числу значимых корреляций с другими признаками. Далее производилась генерация

Алгоритм	Иммуннологические	Клинические	Совмещённые
MLP	81 ± 5	76 ± 4	82 ± 4
RBF	78 ± 4	77 ± 4	71 ± 4
SVM	82 ± 3	78 ± 4	81 ± 3
SVM-f	87 ± 3	81 ± 3	84 ± 2
DT	87 ± 5	85 ± 6	88 ± 3

Таблица 1. Оценки частоты правильных прогнозов (%) по скользящему контролю для 5 алгоритмов классификации и 3 вариантов состава данных: только иммунологические, только клинические, и полные данные. Указаны 90%-е доверительные интервалы.

наборов признаков, причём признакам с высокой разделяющей способностью и бóльшим числом корреляций назначалась более высокая вероятность включения в набор. Для каждого из полученных наборов производилась оценка качества классификации по 10-кратному скользящему контролю (10-fold cross-validation). Описанная эвристическая стратегия направлена на построение информативных наборов признаков и одновременно на получение не сильно смещённых оценок качества классификации в условиях выборки малой длины.

Анализ результатов, представленных в Таблице 1, позволяет сделать следующие выводы. Предсказательная способность иммунологических данных несколько выше, чем клинических. Фильтрация выбросов (нетипичных объектов) из обучающей выборки позволяет построить алгоритм, точность которого на несколько процентов выше. Наилучшим качеством предсказаний обладают решающие деревья, однако они наименее устойчивы к изменению состава данных; для их адекватной настройки требуется выборка большей длины. Другие методы не позволяют увеличить точность прогнозов при совмещении клинических и иммунологических данных.

Работа выполнена при поддержке РФФИ, проекты № 06-07-08102-офи, № 07-01-00734, № 07-07-00380.

Литература

[1] *Кузнецов М. Р., Хаитов М. Р., Туркин П. Ю., Москаленко Е. П., Пинегин Б. В.* Роль нарушений гуморального и клеточного иммунитета в генезе стеноза сосудистых анастомозов после реконструктивных вмешательств на артериях таза и нижних конечностей // Грудная и сердечно-сосудистая хирургия. — 2005. — № 2. — С. 29–33.

[2] *Кузнецов М. Р., Вирганский А. О., Капранов С. А., Москаленко Е. П., Туркин П. Ю. и др.* Способ диагностики функциональной полноценности сосудистого анастомоза после реконструктивных хирургических вмешательств. Патент РФ № 2266711 от 27.12.2005.

Система анализа данных и определения параметров биологических объектов на основе компьютерной модели

Устинин Д. М., Грачев Е. А., Копит Т. А., Черемухин Е. А.
ustinin@mail.ru
Пущино, ИМПБ РАН, Москва, МГУ

В настоящее время основной проблемой в исследовании сложных внутриклеточных биологических систем, таких как фотосинтетическая и дыхательная цепи, является анализ большого количества известных экспериментальных данных различной природы и их интеграция в цельную картину. Целью настоящей работы является построение компьютерной модели процессов в фотосинтетической и митохондриальной мембранах клетки и разработка методов, позволяющих определять параметры модели по экспериментальным данным различной природы. При моделировании явлений в хлоропластах и митохондриях основную сложность представляет их сложная пространственная структура, определяющая протекающие в них процессы. Энергопреобразующие мембраны митохондрий и хлоропластов устроены сходным образом и представляют собой двойной липидный бислой, в который встроены пигмент-белковые комплексы, осуществляющие перенос заряда через мембрану, таким образом создавая трансмембранный электрохимический потенциал, служащий источником энергии при синтезе АТФ. В компартментах, прилегающих к мембране, движутся мобильные переносчики электронов. Дополнительную сложность при моделировании вносит тот факт, что характерные времена переноса заряда внутри белковых комплексов и между различными типами комплексов посредством мобильных переносчиков различаются на несколько порядков.

Методы и алгоритмы

Имитационная компьютерная модель. Для преодоления трудностей, связанных с пространственной неоднородностью изучаемой системы построена новая имитационная модель, непосредственно моделирующая поведение участников электронного и протонного транспорта в хлоропластах и митохондриях. В имитационной модели фотосинтеза мембранные белки и мобильные переносчики электрона описываются как объекты в трехмерном пространстве. Моделируемая область представляет собой двойную фотосинтетическую мембрану, межмембранное пространство (люмен) и область, прилегающую к мембране снаружи (строму). Расположение мембранных белков генерируется псевдослучайным образом, с учетом данных электронной микроскопии. Движение мобильных переносчиков моделируется методом броуновской динамики с уче-

том столкновений. Это позволяет учесть затрудненность диффузии из-за высокой плотности расположения мембранных белков. Перенос электронов внутри белковых комплексов моделируется вероятностным образом. Диффузия протонов в люмене описывается дифференциальным уравнением в частных производных. Модель реализована в виде программного комплекса на языке C++. Время счета сильно сокращается за счет того, что задача естественным образом распараллеливается по реализациям.

Моделирование экспериментально измеряемых сигналов. Основными источниками информации о процессах в хлоропластах и митохондриях являются данные, получаемые методами электронного парамагнитного резонанса, измерениями флуоресценции (для фотосинтеза), а также измерения концентраций протонов и АТФ в различных компартментах органелл. Подробная вероятностная модель процессов внутри белковых комплексов позволяет моделировать сигналы ЭПР и флуоресценции. Модель синтеза АТФ дает возможность вычислить концентрации протонов и АТФ. По этим данным ставится обратная задача оценивания параметров исследуемых биологических объектов.

Полученные результаты

Построенная имитационная модель естественным образом воспроизводит некоторые экспериментальные факты, которые невозможно воспроизвести на более простых кинетических моделях без привлечения дополнительных предположений. Например, это двухфазный характер кинетической кривой восстановления пигмента P700 в циклическом электронном транспорте, а также наличие лаг-периода при синтезе АТФ в хлоропластах в экспериментах с короткой вспышкой света. Эти факты легко объяснить неоднородностью пространственной структуры хлоропласта, что и воспроизводится нашей моделью, учитывающей эту неоднородность. В настоящее время идет работа над развитием методов анализа экспериментальных данных от моделируемых объектов, с целью оценки их параметров.

Работа выполнена при поддержке РФФИ, проект № 06-07-89303.

Формирование признаков распознавания изображений ультразвуковых исследований методами стохастической геометрии

Федотов Н. Г., Шульга Л. А., Смолькин О. А., Кольчугин А. С., Романов С. В.
abrist@pevek.ru, fedotov@diamond.stup.ac.ru
Пенза, Пензенский государственный университет

После катастрофы на Чернобыльской АЭС на территории Белоруссии и соседних государств резко возросло количество заболеваний щитовидной железы. В связи с чем возникла задача массовых обследований населения в целях ранней диагностики. Это позволит снизить риск тяжелых последствий для здоровья и повысит вероятность полного выздоровления.

Основным методом предварительной диагностики заболеваний щитовидной железы является ультразвуковое исследование, оно является ведущим и чрезвычайно информативным подходом выявления заболевания этого органа, особенно при его бессимптомном течении. Квалифицированный специалист, руководствуясь свои опытом и интуицией, описывает характеристики объектов на снимке УЗИ. Для определения данных параметров предлагается использовать методы стохастической геометрии — трейс-преобразование.

В основе метода, в данном случае, лежит сканирование изображения детерминированной решёткой прямых с последующей обработкой полученной трейс-матрицы для получения композиции трех функционалов.

Пусть $F(x, y)$ — функция изображения на плоскости (x, y). Определим на плоскости сканирующую прямую $l(\theta, \rho, t)$:

$$x \cos\theta + y \sin\theta = \rho,$$

где θ и ρ — нормальные координаты, t — естественная координата прямой.

Определим функцию двух аргументов $g(\theta, \rho) = T\big(F \cap l(\theta, \rho, t)\big)$ как результат действия функционала T при фиксированных значениях переменных θ и ρ.

В результате действия функционала T получаем матрицу, элементами которой являются значения $t_{ij} = T\big(F \cap l(\theta_j, \rho_i, t)\big)$, при этом параметры сканирующей линии θ и ρ определяют позицию этого значения в матрице. Назовем эту матрицу *трейс-матрицей*. Последующее вычисление признака заключается в последовательной обработке столбцов матрицы с помощью функционала P. Результатом его действия является вектор значений, непрерывным аналогом которого является 2π-периодическая кривая. Затем применяем к нему функционал Θ (результатом его применения является число — признак). Таким образом, признак вычисляется

Рис. 1. Очаговые образования правильной и неправильной формы.

как последовательная композиция трех функционалов:

$$\Pi(F) = \Theta \circ P \circ T\left(F \cap l(\theta, \rho, t)\right).$$

Определение размеров очагового образования. В качестве размеров очагового образования выступают максимальная протяжённость объекта в любом направлении (назовём её длиной) и максимальная протяженность объекта в перпендикулярном длине направлении (назовём её шириной).

Если в качестве T функционала взять длину большего отсекаемого отрезка прямой $l(\theta, \rho, t)$ на объекте, а функционалы P и Θ — как функции максимума, то получим максимальный диаметр объекта (длину). Зная значение параметра для данного признака и рассматривая в трейсматрице только столбец с параметром $\theta = \theta_k + 90°$, применяя тот же функционал P, получим ширину объекта.

Определение формы очагового образования. При проведении ультразвукового исследования форма очагового образования в щитовидной железе характеризуется как правильная или неправильная, Рис. 1. В общем случае правильной можно назвать форму, близкую к эллипсу.

В качестве функционала T выступает функция количества точек пересечения прямой $l(\theta, \rho, t)$ с объектом, P и Θ — средние значения. Тогда полученный признак будет являться числовой характеристикой формы объекта. А именно: если значения признака близко к двум — объект имеет правильную форму, если значительно больше двух — неправильную. При этом полученный признак позволяет более точно описать форму объекта.

Характеристика границы очагового образования. Граница образования рассматривается как ровная, либо неровная, Рис. 2.

В качестве T функционала возьмем длину большего отсекаемого отрезка прямой $l(\theta, \rho, t)$ на объекте. P функционал определим как оценку вариабельности $t_{i,j}$. В качестве оценки вариабельности может выступать количество перемен знака в векторе значений $\triangle t_{i,j}$, где $\triangle t_{i,j} =$

Рис. 2. Очаговые образования с ровной и неровной границей.

$t_{i,j} - t_{i-1,j}$. Функционал Θ примем как среднее значение этой оценки. Полученный числовой признак будет представлять оценку неровности границы объекта. Если полученное значение близко к единице — объект имеет ровную границу, если значение значительно больше единицы — неровную.

Таким образом, проведённое исследование показало эффективность применения методов стохастической геометрии для определения характеристик и распознавания очаговых образований в щитовидной железе. Благодаря структуре триплетного признака в виде композиции трёх функционалов возможно получение большого числа признаков. Опора на большое число признаков ведет к повышению гибкости и надежности распознавания. Благодаря тому, что анализируются свойства окрестности точки пересечения изображения со сканирующей прямой, при надлежащем выборе функционала можно детально описать свойства окрестности, это также является источником достижения эффективности и универсальности распознавания.

Работа выполнена при поддержке РФФИ, проект № 05-01-00991.

Литература

[1] *Федотов Н. Г.* Методы стохастической геометрии в распознавании образов. — М.: Радио и связь, 1990. — 114 с.

[2] *Паршин В. С., Цыб А. Ф., Ямасита С.* Рак щитовидной железы. Ультразвуковая диагностика. Клинический атлас. По материалам Чернобыля. — Обнинск: МРНЦ РАМН, 2002. — 238 с.

[3] *Федотов Н. Г., Шульга Л. А., Смолькин О. А., Курынов Д. В., Кольчугин А. С., Романов С. В.* Предварительная обработка изображений ультразвуковых исследований в системах медицинской диагностики // Надежность и качество. Труды межд. симпоз., Пенза, 2006. — Т. 2. — С. 247–248.

[4] *Федотов Н. Г., Шульга Л. А., Моисеев А. В.* Теория признаков распознавания и предварительной обработки изображений на основе стохастической геометрии // Измерительная техника. — 2005. — № 8. — С. 8–13.

Формирование признаков распознавания гистологических изображений на основе стохастической геометрии и функционального анализа

Федотов Н. Г., Шульга Л. А., Кольчугин А. С., Смолькин О. А., Романов С. В.

ec@diamond.stup.ac.ru

Пенза, Пензенский государственный университет

Задача распознавания цитологических и гистологических изображений возникает при диагностике онкологических заболеваний. Суть цитологического и гистологического анализов заключается в подготовке препарата и рассмотрении его под микроскопом при различных увеличениях на предмет выявления морфологических признаков, характерных для онкологических заболеваний. В настоящей статье рассмотрена идея создания интеллектуальной системы диагностики, автоматически формирующей триплетные признаки распознавания. Данные признаки базируются на аппарате стохастической геометрии, эффективность которого была подтверждена в [1, 2]. Признаки распознавания в рассматриваемом подходе имеют структуру в виде композиции трех функционалов $\Pi(F) = \Theta \circ P \circ T\left(F \cap l(p,\theta)\right)$, где p, θ — нормальные координаты сканирующей прямой $l(p,\theta)$, с которыми связаны функционалы P и Θ; функционал T связан с естественной координатой t сканирующей прямой $l(p,\theta)$; и F — обозначение изображения распознаваемого объекта. В связи с характерной структурой такие признаки были названы триплетными, их подробное рассмотрение приведено в [2].

Применение данного аппарата непосредственно к исходным гистологическим изображениям затруднительно, поскольку на них изображены ядра, фолликулы, соединительная ткань и другие виды объектов, каждый из которых имеет свои собственные значимые характеристики. Триплетные признаки хорошо «схватывают» геометрические особенности изображенных объектов, но для этого сначала необходимо выполнить их выделение. Эта задача была решена предварительной обработкой изображений в соответствии с процедурой, описанной в [4]. В результате были получены отдельные изображения фолликул и отдельные изображения ядер препарата.

При практическом решении задачи распознавания всегда стоит проблема выделения наиболее информативных признаков. Триплетная структура позволяет получить тысячи различных признаков (для этого достаточно использовать всего 10 функционалов каждого типа), причем в режиме автоматической генерации. Однако вычислительная сложность получения такого числа признаков для каждого распознаваемого изображения, а также сложность построения решающей процедуры при таком

числе признаков требуют от нас ограничиться небольшим количеством наиболее информативных признаков.

Наш подход основывается на формальной генерации большого числа триплетных признаков, формируемых на основе имеющейся библиотеки функционалов, и последующем отборе согласно некоторому критерию эффективности как можно меньшего числа наиболее полезных для распознавания признаков. Отбор признаков часто называют процессом минимизации признакового пространства.

Минимизация обычно включает преобразование кластеризации и выбор признаков. Идея преобразования кластеризации заключается в том, чтобы обеспечить группировку точек, представляющих выборочные образы одного класса. В результате такого преобразования максимизируются расстояния между множествами и минимизируются внутримножественные расстояния.

С точки зрения теории информации критерием оптимизации выбора признаков может служить понятие энтропии. Признаки, уменьшающие неопределенность заданной ситуации, считаются более информативными, чем те, которые приводят к противоположному результату. Таким образом, если считать энтропию мерой неопределенности, то разумным правилом является выбор признаков, обеспечивающих минимизацию энтропии рассматриваемых классов. Это правило эквивалентно минимизации дисперсии в различных совокупностях образов, образующих классы. Выражения для энтропии дают полное представление об информативности описания. Но оценка по этим формулам затрудняется большим объемом вычислений, с учетом того, что в решаемой нами задаче изначально генерируется более 10000 признаков. Это делает задачу определения набора информативных признаков в рамках концепции минимизации энтропии неразрешимой за реальное время. Кроме того, концепция минимизации энтропии основывается на предположении о нормальности распределения образов, составляющих заданные классы, в то время как в реальных задачах законы распределений неизвестны. Объем обучающей выборки часто бывает небольшим, и делать оценки параметров распределения довольно рискованно. В этих условиях целесообразно использовать методы, которые не требуют построения модели распределения и опираются на объекты, имеющиеся в обучающей выборке.

Таким методом является разложение по системе ортогональных функций. При выборе признаков используют обобщенное разложение Карунена-Лоэва, поскольку оно обладает следующими оптимальными свойствами [3]:

1) минимизирует среднеквадратичную ошибку при использовании лишь конечного числа базисных функций в разложении;

2) минимизирует функцию энтропии, выраженную через дисперсии коэффициентов разложения.

Важность первого свойства заключается в том, что оно гарантирует невозможность получения меньшей в среднеквадратичном смысле ошибки аппроксимации с помощью другого разложения. Важность второго свойства заключается в том, что оно связывает с коэффициентами разложения оценку минимальной энтропии или дисперсии.

При генерации признаков распознавания для гистологических изображений изначально было получено 13 500 признаков. На предварительном этапе были отсеяны все вырожденные признаки, значения которых оказались постоянными для всех образов. К оставшимся признакам была применена процедура минимизации на основе разложения Карунена-Лоэва. В результате, для изображений фолликул при коэффициенте $k = 0.8$ было отобрано 59 признаков. Коэффициент k задает долю общей суммы дисперсий $D_j(E[f_{ji}])$ математических ожиданий всех признаков, которая обеспечивается за счет отобранных признаков. Соотношение внутриклассовых и межклассовых дисперсий для отобранных признаков позволяет эффективно организовать процедуру распознавания с использованием простых решающих правил.

Таким образом, можно сделать следующие выводы:

— применение признаков со структурой в виде композиции трех функционалов (триплетных признаков) позволяет формировать большое количество признаков в режиме автоматической компьютерной генерации;
— для отбора наиболее информативных признаков применима процедура, основанная на обобщенном разложении Карунена-Лоэва, которая обеспечивает минимизацию внутриклассовой энтропии, выражаемой через дисперсии коэффициентов разложения.

Работа выполнена при поддержке РФФИ, проект № 05-01-00991.

Литература

[1] *Федотов Н. Г.* Методы стохастической геометрии в распознавании образов. — М: Радио и связь, 1990.

[2] *Федотов Н. Г., Шульга Л. А.* Теория распознавания и понимания образов на основе стохастической геометрии // Искусственный интеллект. — 2002. — № 2. — С. 282–289.

[3] *Ту Дж., Гонсалес Р.* Принципы распознавания образов. — М.: Мир,1978.

[4] *Федотов Н. Г., Шульга Л. А., Кольчугин А. С., Романов С. В., Смолькин О. А., Курынов Д. В.* Предварительная обработка гистологических изображений в системе распознавания заболеваний щитовидной железы // сб. тр. «Надежность и качество–2006». — Пенза, 2006. — Т. 2. — С. 245–246.

Разработка новых методов непрерывной идентификации и прогнозирования состояния динамических объектов на основе интеллектуального анализа данных

Хачумов В. М., Виноградов А. Н.
vmh@vmh.botik.ru
Переславль-Залесский, Институт программных систем РАН

Задачи принятия решений в таких областях, как космос, медицина, машиностроение тесно связаны с необходимостью анализа и интерпретации многомерных данных в реальном времени. Под непрерывной идентификацией здесь понимается мониторинг и определение текущего состояния динамического объекта в течение его жизненного цикла на основе интеллектуального анализа данных. Компонентами интеллектуальной технологии являются [1, 2]: способы когнитивной визуализации параметров динамических объектов, методы анализа изображений, методы диагностики и управления.

Когнитивная визуализация состояний контролируемых объектов

Многомерные данные с помощью ЭВМ могут быть соотнесены в когнитивный графический образ в виде интегральных функциональных профилей или сцен. Например, когнитивная графика обеспечивает непрерывный контроль состояния пациента, визуализируя состояние и допустимые пределы процесса. На Рис. 1 изображены проекции трехмерных образов («звезд») состояний здорового человека, пациентов с легким обострением и тяжелым обострением бронхиальной астмы, которые можно наблюдать в разных плоскостях.

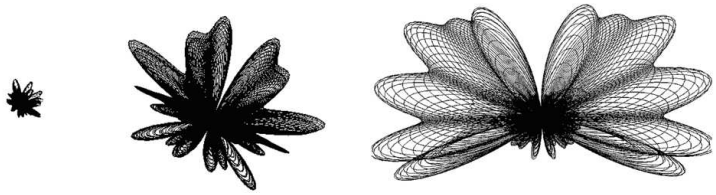


Рис. 1. Когнитивная визуализация состояний пациента.

При ухудшении дыхания «звезда», увеличиваясь, становится более цельной; при росте температуры или частоты пульса концы «звезды» вытягиваются; в случае увеличения параметров газов в крови «звезда» увеличивается с возможным изменением общей структуры, но без выраженного эффекта сглаживания или разделения.

Для космических приложений разработаны методы отображения сложной динамической ситуации в виде соответствующих годографов и их обобщённых графиков. Годографы являются эффективным методом отображения контролируемых и регулируемых параметров относительного движения.

Комплекс алгоритмов для определения ориентации и распознавания объектов

Разработан и исследован комплекс алгоритмов для обнаружения, выделения и распознавания локальных объектов на снимках большой размерности. На Рис. 2 показаны исходный космический снимок и результат выделения локальных объектов (самолетов).

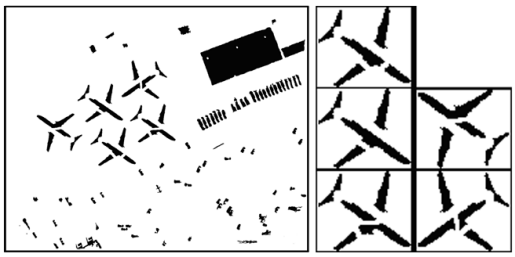


Рис. 2. Нахождение локальных объектов.

Каждый выделенный объект после нормализации проходит далее процедуру распознавания искусственной нейронной сетью (ИНС). Нормализация основывается на определении линий положения объекта и эталона, что позволяет выполнить правильный относительный разворот объекта для последующего корректного сравнения с эталоном (Рис. 3).

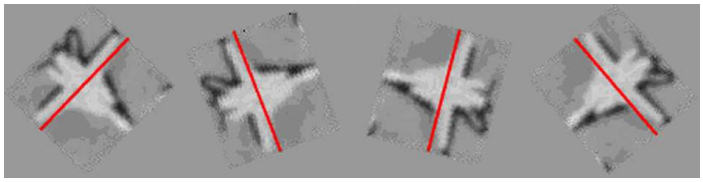


Рис. 3. Нормализация с помощью линий положения.

Расширением метода служит задача определения ориентации тела в трехмерном пространстве, например, для стыковки космических аппаратов. Задача сведена к системе вида: $\mathbf{w} \times \mathbf{I}\mathbf{w} = 0$, $|\mathbf{w}| = 1$, где

$\mathbf{w} = (l\ m\ n)^{\mathrm{T}}$, $\mathbf{I} = \begin{pmatrix} A & D & E \\ D & B & F \\ E & F & C \end{pmatrix}$, где $A, \ldots, F$ — коэффициенты, учитывающие геометрию объекта; l, m, n — коэффициенты прямой.

Полученные собственные значения и векторы матрицы позволяют определять ориентацию и проводить нормализацию. Построена и исследована метрика Евклида–Махаланобиса, позволяющая вести классификацию с большей точностью и универсальностью. Обобщенная метрика определяет расстояние между двумя классами X_1 и X_2 в виде квадратичной формы $R^2_G(X_1, X_2) = (\bar{x}_1 - \bar{x}_2)^{\mathrm{T}} \mathbf{A}^{-1} (\bar{x}_1 - \bar{x}_2)$, где $\bar{x}_1$ и $\bar{x}_2$ — средние выборочные классов, $\mathbf{A} = (\mathbf{C_1} + \mathbf{E})(\mathbf{C_2} + \mathbf{E})$, где $\mathbf{C_1}$ и $\mathbf{C_2}$ — ковариационные матрицы для классов X_1 и X_2 соответственно. Предлагаемая метрика учитывает корреляционные свойства классов таким образом, что расстояние между точкой и классом стремится к расстоянию Евклида, когда дисперсии параметров класса стремятся к нулю.

Комплекс алгоритмов для диагностики объектов

Рассматривается методология интеллектуального анализа рабочих характеристик и выделения информативных параметров. Полученные параметры используются для лингвистического описания процесса, служащего входной информацией для экспертной системы или ИНС, ставящей окончательный диагноз. Управление возлагается на нечеткий контроллер, для оптимизации настроек которого предлагается использовать генетический алгоритм.

Заключение

Разработанные новые механизмы интеллектуального анализа данных обеспечивают понимание ситуаций и качественное управление динамическими объектами. В целом перечисленные компоненты образуют интеллектуальную технологию анализа данных для непрерывной идентификации и принятия решений [1, 2].

Работа выполнена при поддержке РФФИ, проект № 06-07-89083.

Литература

[1] *Бурдаев М. Н., Виноградов В. Ф., Заднепровский В. Ф., Захаров А. В., Куршев Е. П., Хачумов В. М.* Комплекс программно-инструментальных средств для создания интеллектуальных систем контроля и управления объектами аэрокосмического назначения // Авиакосмическое приборостроение. — 2006. — № 8. — С. 24–33.

[2] *Амелькин С. А., Захаров А. В., Хачумов В. М.* Обобщенное расстояние Евклида–Махаланобиса и его свойства // Информационные технологии и вычислительные системы. — 2006. — № 4. — С. 40–44.

Извлечение таблиц из неформатированного текста

Хмельнов А. Е., Шигаров А. О.
hmelnov@irk.ru, shigarov@icc.ru
Иркутск, Институт динамики систем и теории управления СО РАН

В данной работе представлены основные концепции разработанного нами эвристического метода извлечения таблиц из неформатированного текста. Метод использует особенности структуры статистических таблиц, публикуемых Росстатом. Также эти особенности в полной мере относятся к статистическим таблицам, представленным в государственных статистических отчетах США (www.fedstats.gov), Евросоюза (Eurostat yearbook 2006-07) и Японии (Statistical Handbook of Japan 2006). Метод может быть применен к подобным таблицам, представленным, как неформатированный текст.

Извлечение таблиц из документов является одной из задач, решаемых в системах анализа и обработки документов. Обзоры работ по данной проблеме [1, 2, 3], появившиеся за несколько последних лет, показывают растущий интерес к данной проблематике. В литературе выделяются следующие основные стадии обработки, которые могут быть выполнены при извлечении таблиц: обнаружение таблиц в документах, сегментация таблиц на отдельные клетки, функциональный анализ — определение роли клеток, структурный анализ — определение зависимостей между клетками, и интерпретация — преобразование табличной информации к требуемому виду. В работах [4, 5, 6] предложены различные подходы к обнаружению таблиц в неформатированном тексте. В работе [7] предложен метод извлечения таблиц из неформатированных текстов, в котором реализованы все вышеупомянутые стадии обработки, но при этом используются слишком сильные предположения о структуре обрабатываемых таблиц.

Как правило, методы извлечения таблиц из документов ориентируются на определенные среды и форматы представления документов, а также на определенную структуру таблиц, которая обычно определяется стандартами и соглашениями, принятыми в той предметной области, где используются эти таблицы.

В данной работе рассматривается метод, учитывающий особенности статистических таблиц, и позволяющий выполнить все стадии их обработки, результатом применения которого является извлечение информации из текстовых таблиц в реляционную БД. На Рис. 1 показан пример статистической таблицы. Рассматриваемые таблицы состоят из шапки и тела, кроме того, они могут иметь боковик и перерезы. Тело таблицы содержит только числовые данные. Заголовки столбцов обычно выделяются линейками, составленными из символов псевдографики или

Рис. 1. Пример статистической таблицы

подобных им символов набора ASCII. Пересекаясь, линейки образуют клетки, которые ограничивают отдельные заголовки столбцов. Один или несколько заголовков могут быть вложенными в другой заголовок; в этом случае, клетки, ограничивающие их, лежат сразу под клеткой, ограничивающей заголовок, в который они вложены. Заголовки строк также имеют вложенность, которая определяется отступом от левого края таблицы.

Особенности компоновки заголовков столбцов позволяют представить шапку в виде дерева, узлами которого являются заголовки столбцов, а ребрами — пары заголовков (h_a, h_b), где h_a — заголовок, вложенный в h_b. Корнем этого дерева является пустой элемент, заголовки верхнего уровня являются его подузлами. Подобным образом, о боковике также можно думать, как о дереве, в котором заголовки строк являются узлами, а пары заголовков строк, в которых один вложен в другой — ребрами. Перерезы также удобнее рассматривать как дерево, хотя они не имеют вложенности.

В результате для представления рассматриваемых таблиц может быть предложена следующая модель. Пусть H^t, S^t и C^t — деревья, представляющие соответственно шапку, боковик и перерезы, а H, S и C — множества узлов, соответствующие этим деревьям. Пусть $V\colon V \subset \mathbb{R}$ — множество всех значений из тела таблицы. Пусть $L \subseteq H \times S \times C$ — подмножество таких элементов из $H \times S \times C$, для которых определено значение $v \in V$. Тогда множество $T = \{H^t, S^t, C^t, L \to V\}$ составляет модель таблицы.

Прежде всего, решается задача обнаружения таблиц в тексте. При этом используется ряд предположений о наличии в шапке таблицы символов-разделителей. После обнаружения и сегментации заголовка таблицы выполняется сегментация строк текста, составляющих боковик и тело, а также выделяются перерезы. Таким образом, приходим к описанной модели таблицы. Используемые предположения выполняются для абсолютного большинства рассматриваемых таблиц, в случае их нарушения результаты сегментации могут редактироваться пользователем.

Следующим шагом является классификация и нормализация обнаруженных узлов в деревьях заголовков. Для этих целей используются правила, задаваемые при помощи регулярных выражений. Могут выделяться значения, относящиеся к различным измерениям (временной интервал, территория), а также игнорируемые узлы. При обнаружении в заголовке значения, оно сопоставляется всем клеткам таблицы, подчинённым рассматриваемому узлу, а сам узел исключается из дерева (с сохранением подчинённых узлов).

На основе предложенного метода разработано приложение для перевода текстовых таблиц в реляционную БД, с использованием которого в сжатые сроки было обработано около 2800 таблиц, содержащих более 21000 показателей и более 300000 значений.

Литература

[1] *Lopresti D., Nagy G.* A tabular survey of automated table processing // Lecture Notes in Computer Science. — 2000. — Vol. 1941 — pp. 93–120.

[2] *Zanibbi R., Blostein D., Cordy J. R.* A survey of table recognition: Models, observations, transformations, and inferences // International Journal on Document Analysis and Recognition. — 2004. — Vol. 7, No. 1. — pp. 1–16.

[3] *Embley D. W., Hurst M., Lopresti D., Nagy G.* Table-processing paradigms: a research survey // International Journal on Document Analysis and Recognition. — 2006. — Vol. 8, No. 2. — pp. 66–86.

[4] *Tupaj S., Shi Z., Chang C. H., Alam H.* Extracting Tabular Information From Text Files. — 1996. — `citeseer.nj.nec.com`.

[5] *Hu J., Kashi R., Lopresti D., Wilfong G.* Medium-Independent Table Detection // Document Recognition and Retrieval VII (IS&T/SPIE Electronic Imaging), San Jose, 2000. — pp. 291–302.

[6] *Pinto D., McCallum A., Wei X., Croft B.* Table Extraction Using, Conditional Random Fields // 26th Annual International ACM SIGIR, Conference on Research and Development in Information Retrieval, 2003.

[7] *Douglas S., Hurst M., Quinn H.* Using Natural Language Processing for Identifying and Interpreting Tables in Plain Text // 4th annual symposium on document analysis and information retrieval, Las Vegas, 1995. — pp. 535–546.

Использование методов распознавания при прогнозировании радиационной обстановки на долговременных пилотируемых космических станциях

Цетлин В. В., Носовский А. М., Сенько О. В., Кузнецова А. В.
vtsetlin@mail.ru, nam@imbp.ru, senkoov@mail.ru, azfor@narod.ru
Москва, ИМБП РАН, ВЦ РАН, ИБХФ РАН

Одной из задач, связанных с обеспечением безопасности пилотируемых космических полетов является прогноз радиационной обстановки на околоземных орбитальных космических станциях. Целью настоящего исследования явилось создание алгоритма осуществляющего прогноз среднесуточной мощности дозы космического ионизирующего излучения в отсеках Международной космической станции (МКС) на основе результатов многолетних наблюдений. Нами был использован массив данных, включающий ежесуточные значения приращения дозы ионизирующего излучения, измеренные с помощью двух независимых каналов измерений штатного радиометра Р16, результаты ежесуточных измерений гелиогеофизических данных солнечной активности а также набора баллистических параметров орбиты станции. Стандартные методы прогнозирования временных рядов не привели к желаемому результату. Неудача по-видимому была связана с такими факторами как наличие значительной по величине случайной составляющей, связанной с непредсказуемыми вспышками на Солнце, а также сложный нелинейный характер зависимостей. В связи с этим был предложена методика, основанная на теории распознавания, позволяющая исключить влияние больших по величине «выбросов». Дни наблюдений были разбиты на три подгруппы в зависимости от уровня суточного приращения дозы: с высоким, средним и низким уровнями радиационной активности. Прогнозирование основывалось на решении задачи распознавания первой и третьей групп. В качестве прогностических переменных использовались:

а) значения баллистических параметров орбиты МКС на момент прогноза (в течение суток, для которых делается прогноз);
б) значения показателей солнечной активности, измеренные за семь и более дней до момента прогноза;
в) значения так называемых циклических (фазовых) переменных, описывающих циклические изменения уровня радиации.

На предварительном этапе с использованием метода оптимальных достоверных разбиений [3, 4] производилась оценка ценности всевозможных предполагаемых прогностических переменных и формировался оптимальный набор прогностических переменных. По результатам реше-

ния задачи распознавания с использованием изложенного ниже подхода строился собственно алгоритм для прогнозирования суточного приращения дозы. В результате была достигнута точность средненедельного прогноза, оцениваемая коэффициентом линейной корреляции 0.56 между прогнозом и реальными значениями суточных приращений.

Циклическая переменная

Значение циклической переменной $C_p[j]$, соответствующей предполагаемой длине цикла p для суток с номером j относительно некоторой точки отсчета вычисляется как остаток от деления j/p или $C_p[j] = j/p - [j/p]$. При наличии в динамическом ряду периодического чередования с длиной цикла p временных интервалов с высокими и низкими уровнями значений прогнозируемой величины на отрезке значений циклической переменной C_p выделяется интервал, в котором значения прогнозируемой величины отличаются от значений в остальной части интервала. Подобные неоднородности в данных выявляются с помощью метода оптимальных разбиений с использованием одномерной модели с двумя граничными точками [3, 4].

Использование методов распознавания при решении задач динамического прогнозирования

В случае непрерывной прогнозируемой переменной использование методов распознавания позволяет выделить набор прогностических уровней с различными средними значениями прогнозируемой величины. На первом шаге производится разбиение интервала допустимых значений прогнозируемой величины Y. В частности могут быть выделены интервалы с $Y < a$ и $Y > b$, где $b > a$. Моменты времени с $Y < a$ образуют класс K_1, а моменты времени с $Y > b$ образуют класс K_2. Далее строится алгоритм $A(a, b)$, распознающий классы K_1 и K_2. Производится распознавание объектов, соответствующих каждому из моментов анализируемого временного ряда. Причем объекты, соответствующие моментам с $Y < a$ и $Y > b$, распознаются в режиме скользящего контроля. Предположим, что алгоритм $A(a, b)$ представим в виде произведения оператора $R(a, b)$ и решающего правила $C(a, b)$, см. [2]. Причем оператор $R(a, b)$ для объектов, соответствующих моментам рассматриваемого временного ряда, вычисляет значения из отрезка $[G_0, G_f]$. Произведем разбиение отрезка $[G_0, G_f]$ с помощью пороговых значений $G_1, \ldots, G_{f-1}$. Упомянутым выше прогностическим уровням соответствуют подмножества $[G_0, G_f]$: $[G_0, G_1), \ldots, [G_{f-1}, G_f]$. Для каждого из подмножеств $[G_{i-1}, G_i)$ вычисляется среднее по обучающей информации значение прогнозируемой переменной $Y - \hat{y}_i$. При прогнозировании Y для нового момента времени t с помощью оператора $R(a, b)$ вычисляется оценка $g(t)$. Выясняется, како-

му из подмножеств $[G_0, G_1), \ldots, [G_{f-1}, G_f]$ принадлежит $g(t)$. В качестве прогнозируемого значения Y для момента t выбирается среднее значение для подмножества $[G_{i(t)-1}, G_{i(t)})$, содержащего $g(t)$. Таким образом алгоритм прогнозирования, основанный на методах распознавания, ставит в соответствие вектору значений прогностических показателей дискретный прогнозный уровень. Для каждого из дискретных прогнозных уровней предварительно по имеющимся результатам накопленных наблюдений вычисляются средние значения уровня радиации, которые и используются далее в качестве прогнозов.

Работа выполнена при поддержке РФФИ, проект № 05-07-90333.

Литература

[1] *Цетлин В. В., Акатов Ю. А., Архангельский В. В., Митрикас В. Г., Бондаренко В. А., Макин А. С.* Результаты мониторинга радиационных условий внутри РС МКС (2000–2005 гг.) // Авикосмическая и экологическая медицина. — 2006. — № 5. — С. 21–26.

[2] *Журавлёв Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов. I. // Кибернетика. — 1977. — № 4. — С. 5-17.

[3] *Сенько О. В.* Перестановочный тест в методе оптимальных разбиений. // Журнал вычислительной математики и математической физики. — 2003. — № 9. — С. 1438–1447.

[4] *Senko O. V., Kuznetsova A. V.* The Optimal Valid Partitioning Procedures. // Statistics on the Internet. — 2006. April. — `statjournals.net`.

Статистические методы выявления паттернов скрытой периодичности биологических последовательностей в условиях недостаточного объема выборки

Чалей М. Б., Назипова Н. Н., Кутыркин В. А.

`maramaria@yandex.ru`

Пущино, Институт Математических Проблем Биологии РАН,
Московский Государственный Технический Университет им. Н. Э. Баумана

В настоящее время выявление скрытой периодичности в биологических последовательностях основано на понятии размытого тандемного повтора [1, 2], представленного линейным списком поврежденных копий исходного текстового фрагмента (паттерна). Внутренние повреждения копий паттерна могут быть обусловлены не только заменами его отдельных символов, но и вставками или выпадением букв. Тандемные (т. е. идущие один за другим без перерывов) повторы играют существенную роль в процессах регуляции, функционирования и структурирования геномной ДНК, связаны с рядом наследственных болезней.

Для выявления тандемных повторов широко применяются комбинаторные методы, в том числе и методы динамического программирования [1, 2, 3]. Альтернативные методы выявления скрытой периодичности используют статистические критерии проверки однородности строк [4, 5] и различные методы спектрального анализа [6].

Как показано в настоящей работе, комбинаторные методы не всегда оптимально выявляют паттерн периодичности тандемного повтора [7]. Альтернативные методы, строго говоря, фиксируют неоднородности в биологических последовательностях. Но одна только фиксация неоднородностей еще не означает, что последовательность является размытым тандемным повтором. Кроме того, на практике, для достоверной фиксации неоднородности объем статистических данных, как правило, оказывается недостаточным. В результате возможны не только значительные погрешности при оценке размера паттерна периодичности, но и ошибочное признание наличия в последовательности размытого тандемного повтора.

В работе для выявления скрытой периодичности нестандартным образом (подробнее см. [7]) используются статистики стандартных критериев проверки однородности текстовых строк: критерий Пирсона (P-критерий), нормализованный критерий Пирсона (NP-критерий) и информационный критерий (IC-критерий [8]). Для текстовой строки длины n, записанной в алфавите из K букв, анализируемой на тест-периоде L (предварительная оценка длины паттерна периодичности), эти стандартные статистики имеют вид:

$$\begin{aligned}
\nu_P &= R_L \sum_{j=1}^{L}\sum_{i=1}^{K}(\pi_j^i - p^i)^2/p^i;\\
\nu_{NP} &= R_L \sum_{j=1}^{L}\sum_{i=1}^{K}(\pi_j^i - p^i)^2/p^i(1-p^i);\\
\nu_{IC} &= 2R_L \sum_{j=1}^{L}\sum_{i=1}^{K}\big(\pi_j^i \ln(\pi_j^i) - p^i \ln(p^i)\big);
\end{aligned} \tag{1}$$

где $R_L = \frac{n}{L}$ — число копий тест-периода, π_j^i — частота встречаемости i-й буквы алфавита в j-й позиции тест-периода, $p^i = \frac{1}{L}\sum_{j=1}^{L}\pi_j^i$ — оценка частоты встречаемости i-й буквы алфавита во всем тексте.

Следует отметить, что статистика ν_{NP} ранее не использовалась для выявления скрытой периодичности в биологических последовательностях, а статистика ν_{IC} представлена здесь в виде, отличном от того, как она была введена в работе [8] и использована в работе [4]. Формула (1)

показывает, что IC-статистика суммирует по позициям тест-периода отклонения энтропии реального распределения букв алфавита в каждой позиции тест-периода от ожидаемой энтропии.

В настоящей работе при недостатке объема статистического материала проблема достоверного выявления неоднородности в биологических последовательностях решается на основе модели дополнительных статистических экспериментов, использующих метод Монте-Карло. Эта модель адекватно описывает проявление неоднородностей при использовании статистических критериев проверки однородности текстовых строк. На основе этой модели предлагаются нестандартные двухэтапные поликритерии, удобные для предварительного автоматизированного выявления скрытой периодичности в условиях недостаточного объема выборки [7]. На первом этапе этих критериев используются результаты большого числа предварительных статистических экспериментов. В процедуре второго этапа используется метод Монте-Карло, основанный на дополнительных статистических экспериментах. Объем дополнительных экспериментов сокращается за счет совместного использования статистик различных критериев для проверки однородности текстовых строк. Предлагаемые в работе нестандартные поли-критерии обеспечивают уровень значимости выявляемой неоднородности порядка 10^{-6}. Введение второго этапа существенно повышает мощность поли-критерия.

Во многих случаях применение разработанных поли-критериев позволило более оптимально оценить размер и состав паттерна периодичности для известных тандемных повторов из базы данных TRDB (`http://tandem.bu.edu/cgi-bin/trdb/trdb.exe`). Наиболее характерные примеры такой переоценки паттерна приведены в работе [7].

Работа выполнена при поддержке РФФИ, проекты № 06-07-89274, № 06-01-08039.

Литература

[1] *Benson G.* Tandem repeats finder: a program to analyze DNA sequences // Nucl. Acids Res. — 1999. — V. 27. — P. 573–580.

[2] *Kolpakov R., Bana G., Kucherov G.* mreps: efficient and flexible detection of tandem repeats in DNA // Nucl. Acids Res. — 2003. — V. 31. — P. 3672–3678.

[3] *Boeva V., Regnier M., Papatsenko D., Makeev V.* Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression // Bioinformatics. — 2006. — V. 22. — P. 676–684.

[4] *Korotkov E. V., Korotkova M. A., Kudryashov N. A.* Information decomposition method to analyze symbolical sequences // Phys. Lett. A. — 2003. — V. 312. — P. 198–210.

[5] *Gatherer D., McEwan N.* Analysis of sequence periodicity in E. coli proteins // J. Mol. Evol. — 2003. — V. 57. — P. 149–158.

[6] *Li W.* The study of correlation structures of DNA sequences: a critical review // Computers Chem. — 1997. — V. 21. — P. 257–271.

[7] *Чалей М. Б., Назипова Н. Н., Кутыркин В. А.* Совместное использование различных критериев проверки однородности для выявления скрытой периодичности в биологических последовательностях // Мат. биол. и биоинф. (эл.журнал) — 2007. — Т. 2. — С. 20–35. — `www.matbio.org/downloads/Chaley2007(2_20).pdf`

[8] *Кульбак С.* Теория информации и статистика // М.: Наука. — 1967.

Метод быстрой корреляции с использованием множества шаблонов в задачах анализа изображений

Чичева М. А., Глумов Н. И., Копенков В. Н., Мясников Е. В.

`mchi@smr.ru, nglu@smr.ru, vkop@smr.ru, mevg@smr.ru`

Самара, Институт систем обработки изображений РАН

В настоящей работе рассматривается задача поиска и распознавания на изображениях фрагментов, соответствующих одному из множества шаблонов. Такая задача возникает, например, при распознавании текста. Другим очевидным примером является поиск характерных точек на изображениях лиц, таких как координаты зрачков, уголков губ, и т. п. Эти задачи похожи тем, что искомый образец может принимать существенно разный вид, при этом, возможно, следует определить не только его местоположение, но и меру сходства с каждым из образцов, которая впоследствии может служить критерием при принятии решения или использоваться в качестве признака.

Общепринятым подходом для решения таких задач является хорошо известный из литературы метод [1, 2], когда ищется корреляция входного изображения с каждым из шаблонов, после чего анализируются полученные корреляционные поля. Независимо от того, какой способ вычисления корреляции будет выбран (прямое вычисление, через дискретное преобразование Фурье и т. п.), такой подход требует решения ряда проблем. Первая из них — это высокая вычислительная сложность, которая естественно возрастает с ростом числа шаблонов. Вторая — поиск критерия, который позволит из множества точек корреляционных полей выбрать наиболее соответствующую истинному значению. Наконец, в связи с тем, что на реальных изображениях объекты, как правило, находятся на сложном фоне, искажены, зашумлены, необходимо уменьшить влияние мешающих факторов.

К настоящему времени существует ряд методов [3], в той или иной мере решающих изложенные проблемы. Так, использование параллельно-рекурсивной реализации при помощи аппроксимации шаблонов специ-

альным базисом позволяет существенно снизить время обработки. Упрощение шаблонов (например, бинаризация), изменение формы учитываемой области шаблона (уход от прямоугольного «окна») позволяет как частично решить проблему быстродействия, так и уменьшить влияние мешающих факторов. Специальная обработка корреляционных полей может облегчить поиск и расчет критериев. Однако, ни один из названных методов в чистом виде не позволяет эффективно решить поставленную задачу. В докладе предлагается метод, сочетающий в себе перечисленные подходы, а также демонстрируется его применение к двум задачам анализа изображений.

Предлагаемый метод быстрой корреляции со множеством шаблонов состоит из следующих шагов.

1. *Предварительные действия* включают в себя подготовку набора шаблонов (выполняется один раз для всего набора изображений) и определение (поиск или задание) области, в которой будет производиться сравнение с шаблонами (выполняется для каждого изображения).
2. *Формирование корреляционных полей* заключается в расчете корреляции с каждым из шаблонов для всех его положений в области поиска.
3. *Обработка корреляционных полей* выполняется с целью отбора перспективных точек.
4. *Принятие решения* делается на основе анализа корреляционных полей в области выбранных точек.

Характерной особенностью метода является введение тринарных шаблонов, которые содержат только три значения: 1 — для области объекта, -1 — для области фона и 0 — для части, неучитываемой при анализе. Такой способ формирования шаблонов позволяет реализовать алгоритм быстрой корреляции, при котором вычисление свертки изображения символа с шаблоном реализуется без операций умножения, что позволяет существенно ускорить процесс распознавания символов. При этом количество отсчетов изображения, реально участвующих в формировании ответа, невелико. Наличие же неучитываемой части позволяет снизить влияние шумов, искажения толщины линий, исключить иные мешающие факторы.

Задача распознавания машиночитаемых строк на сканированных изображениях документов. Для документов, удостоверяющих личность (паспортов международного образца), Международной Ассоциацией Гражданской Авиации (ICAO) разработан стандарт MRTD [4], который распространяется на документы, предъявляемые при путешествиях (паспорта и визы). В соответствии с ним, изображение документа личности, в частности, содержит семантическую информацию в виде машиночитаемых строк.

Подробно алгоритм распознавания машиночитаемых строк на изображении будет изложен в докладе. В частности, будут показаны сформированные шаблоны, изложен алгоритм выбора решения из списка возможных вариантов, приведены результаты экспериментального исследования.

Задача поиска положения глаз на документальных фотографиях лиц. Эта задача возникает в рамках практически всех алгоритмов распознавания лиц, первым шагом в которых выполняется геометрическая нормализация с привязкой центров роговиц к заранее заданным точкам.

Предварительно формируются тринарные шаблоны для набора различных радиусов роговиц. В расчете свертки участвует небольшое количество отсчетов, причем с ростом радиуса роговицы число таких отсчетов растет несущественно. Это обеспечивает высокую скорость вычисления.

В докладе будет подробно изложен метод определения зоны поиска глаз, метод обработки корреляционных полей, а также критерий, по которому осуществляется выбор верного варианта. Особенностью решения этой задачи является учет ограничений, которые обусловлены расположением глаз, и требований к фотографиям на документы.

Экспериментальное исследование алгоритма показало, что с вероятностью 0.967 максимальное отклонение найденных координат центров роговиц от их истинных значений не превышает $0.25\,r$, где r — радиус роговицы.

Работа выполнена при поддержке РФФИ, проекты № 06-01-00722, № 07-01-96612, № 07-07-97610.

Литература

[1] *Прэтт У. К.* Цифровая обработка изображений. В 2 томах. Пер. с англ. — М.: Мир, 1982.— 310 с.

[2] *Гонсалес Р., Вудс Р.* Цифровая обработка изображений. Пер. с англ. — М.: техносфера, 2005. — 1072 с.

[3] Методы компьютерной обработки изображений // Под ред. Сойфера В. А. — М.: Физматлит, 2003. — 784 с.

[4] Документ ICAO 9303 «Machine Readable Travel Documents (MRTD)» `http://www.icao.int/mrtd/publications/doc.cfm`.

Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематического состава потока новостей

Шмулевич М. М., Киселев М. В.
mark.shmulevich@gmail.com, mkiselev@megaputer.com
Москва, Московский Физико-Технический Институт,
компания «Megaputer Intelligence»

Данная работа посвящена автоматической смысловой кластеризации текстов. Рассмотрено её применение к анализу тематического состава потока новостей. Предложен новый метод, названный *островной кластеризацией*, который основан на статистической мере корреляции встречаемости в текстах термов, характеризующихся значимым превышением их частот над средним уровнем. Показано, что он обладает набором качеств, необходимых для успешного решения задачи кластеризации текстов, и может быть применен для анализа тематической структуры новостного потока.

В настоящее время объем массивов текстовых документов в научной сфере, бизнесе, и других областях человеческой деятельности неуклонно растёт. Этим обусловлен растущий интерес к методам автоматической текстовой кластеризации. Наиболее часто используемым подходом к представлению текстов при кластеризации является подход, в котором текст рассматривается как неупорядоченный набор начальных форм входящих в него слов (Bag of Words).

При формировании деревьев решений для кластеризации текстов возможно учитывать агрегированные правила, построенные на основе статистических и семантических характеристик термов, выделяемых из текстовых документов. Один из таких методов, названный *островной кластеризацией*, рассмотрен в данной работе. Алгоритм основан на использовании статистической меры корреляции встречаемости в текстах термов, характеризующихся значимым превышением их частот над средним уровнем.

Первая часть алгоритма состоит в построении так называемого графа корреляций термов. Этот граф задается матрицей парных корреляций булевых переменных a_{ip}, отражающих наличие терма i в документе p, так что связь между термами i и j считается существующей при достаточно сильной (большей, чем пороговое значение) корреляции между переменными a_i и a_p. Степень корреляции между термами i и j определяется следующим образом. Пусть n — общее количество термов во всех документах, n_i — количество термов в документах, в которых встречается терм i. Обозначим общее число термов j во всех текстах как N_j,

а количество термов j в документах, содержащих терм i — как N_{ij}. Если принять гипотезу, что термы i и j распределены в документах независимо друг от друга, то вероятность того, что в документах, содержащих терм i, окажется N_{ij} или более термов j — это вероятность получения не менее N_{ij} успехов в серии из N_j испытаний при вероятности успеха одного испытания, равной $\frac{n_i}{n}$. Эта вероятность есть $p_{ij} = P_B(N_{ij}, N_j, \frac{n_i}{n})$, где $P_B(n, N, p) = \sum_{i=n}^{N} b(i; N, p)$ — биномиальное распределение. Вероятность p_{ij} может быть принята в качестве основы для расчета меры корреляции между термами i и j — чем она меньше, тем более коррелированы эти термы. Однако величина p_{ij} все же не совсем подходит для описания силы связи термов i и j, в частности, потому что она, как легко видеть, не симметрична: $p_{ij} \neq p_{ji}$. Поэтому в качестве меры корреляции термов берется $\tilde{p}_{ij} = \max(p_{ij}, p_{ji})$.

Одним из важных применений данного метода может быть анализ динамики тематической структуры потока новостей. Показано, что рассматриваемый метод удовлетворяет основным свойствам, которыми должна обладать процедура кластерного анализа для того, чтобы быть практически применимой к кластеризации больших массивов текстов вообще, и к анализу динамики тематической структуры потока новостей в частности:

— интерпретируемость найденных кластеров в терминах смысла содержания относящихся к ним документов;
— статистическая значимость группирования текстов в кластеры;
— возможность отнесения документа более, чем к одному кластеру;
— не более чем логлинейный рост времени работы кластеризатора с увеличением количества текстов;
— минимальная (или вообще отсутствующая) необходимость настройки со стороны пользователя.

Применение процедуры островной кластеризации было проиллюстрировано с использованием публично доступного массива новостей Reuters–21578. Было показано, что метод островной кластеризации может успешно решать задачу тематической кластеризации потока новостей, давая описание полученных результатов в понятных человеку терминах.

В работе были использованы средства пакета для анализа данных PolyAnalyst.

Работа выполнена при поддержке компании Яндекс, грант № 102903.

Литература

[1] *Fellbaum C.* WordNet: An Electronic Lexical Database. — MIT Press, 2005.

[2] *Hofmann T.* Probabilistic Latent Semantic Indexing // 22-nd Ann. ACM Conf. on Research and Development in Information Retrieval, 1999. — Pp. 50–57.

[3] *Apte C. Weiss S.* Data Mining with Decision Trees and Decision Rules // Corpus Linguistics: Investigating Language Structure and Use. — 1997. — № 13. — Pp. 197–210.

[4] PolyAnalyst data/text mining system. User manual — `www.megaputer.com`.

Модифицированное равновесие в лабораторных сетевых рынках

Яминов Р. И.
`rinaty@gmail.com`
Москва, МФТИ

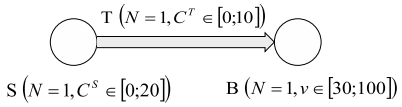
Характерной особенностью сетевого рынка [1, 2] является отсутствие возможности у продавца и покупателя взаимодействовать друг с другом непосредственно, а только через транспортную сеть. Транспортная сеть представляет ориентированный граф, в вершинах которого расположены продавцы и покупатели, ребра соответствуют третьему типу агентов — транспортировщикам, а направление показывает, в какую сторону может передаваться продукт.

1. Покупатель (**B**uyer) хочет приобрести N единиц продукта на рынке для конечного использования. Он обладает некоторыми выкупными стоимостями товара $v_1, \ldots, v_N$, и выигрыш его в случае приобретения $k > 0$ единиц товара равняется $\Pi^B = \sum_{i=1}^{k} \left(v_i - p_i^B\right)$, где p_i^B — цена, по которой была куплена i-я единица товара. Если покупатель не приобрел продукт, то его выигрыш равен нулю. Выкупная стоимость товара может являться случайной величиной, которая реализуется в момент начала аукциона до выставления участниками своих ставок. При этом только покупатель знает реализации выкупных стоимостей. Остальным участникам аукциона известно лишь распределение.
2. Продавец (**S**eller) обладает N единицами неделимого продукта и готов продать их на рынке. Издержки продавца: $C_1^S, \ldots, C_N^S$. Прибыль в случае продажи $k > 0$ единиц, равна $\Pi^S = \sum_{i=1}^{k} \left(p_i^S - C_i^S\right)$, где p_i^S — цена, по которой была продана i-я единица товара. В случае, если продавец ничего не продал, его выигрыш равен 0.
3. Транспортировщик (**T**ransporter) может доставить N единиц товара из одной вершины в другую. Издержки транспортировки: $C_1^T, \ldots, C_N^T$. Прибыль от транспортировки k единиц равна $\Pi^T = \sum_{i=1}^{k} \left(p_i^T - C_i^T\right)$, где p_i^T — цена, транспортировки из одной вершины в другую. В случае, если транспортировщик ничего не перевёз, его выигрыш равен 0.

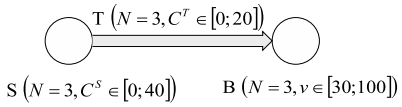
Опишем механизм заключения сделки. Все участники, независимо друг от друга, выставляют заявки. Продавец на каждую единицу товара выставляет цену, за которую он готов ее продать. Транспортировщик на каждую единицу товара выставляет цену, за которую он готов её перевезти. Покупатель на каждую единицу товара, которую он может купить, выставляет максимальную цену, за которую он еще готов купить. После чего по правилам аукциона выбираются те заявки, которые максимизируют общий дополнительный доход.

В работе рассматриваются три теоретико-игровые ситуации.

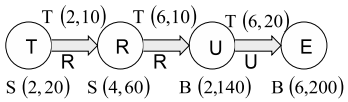
Базовый эксперимент. Сделка происходит, если заявка покупателя больше суммы заявок транспортировщика и продавца. Цена определяется следующим образом: $p_B = Q_B - \frac{1}{3}\Delta$, $p_S = Q_S + \frac{1}{3}\Delta$, $p_T = Q_T + \frac{1}{3}\Delta$, где $\Delta = Q_B - Q_S - Q_T$, а Q_S, Q_T, Q_B – заявки игроков.



Увеличены отрезки затрат для транспортировщика и продавца. Теперь не все сделки могут заключиться. У каждого игрока теперь по 3 заявки.



Более сложный граф. Теперь для нахождения тех заявок, которые должны быть удовлетворены, нужно решать задачу линейного программирования и двойственную — для нахождения цен. У одного игрока не одна роль, а может быть несколько.



Во всех этих играх есть бесконечно много равновесий Байеса-Нэша, поэтому возникает проблема селекции равновесия. Для этих целей, по аналогии с Perfect Equilibrium [5] и Proper Equilibrium [6], вводится модифицированное равновесие.

Для байесовской игры с неполной информацией $G = \langle N, A, T, p, u\rangle$, в которой $A_i \subseteq \mathbb{R}^k$ — непустые выпуклые компакты, введем дополни-

тельно некое семейство ξ_m^d независимых непрерывных векторных распределений с дисперсией d и матожиданием m такое, что распределения непрерывно зависят от дисперсии d и матожидания m.

Введем модифицированные выигрыши игроков:

$$\widetilde{u}_i\big(a, t \,|\, \xi_m^d\big) = \mathsf{E}u\big(a_1 + \xi_m^d, \ldots, a_i, \ldots, a_n + \xi_m^d, t\big),$$

где матожидание берется по ξ_m^d. Обозначим через $S = (s_1, \ldots, s_n)$ профиль стратегий $s_i \colon T_i \to A_i$ всех игроков.

Определение 1. *Профиль стратегий S будем называть модифицированным равновесием для байесовской игры G, если существуют равновесия Байеса-Нэша $S\,(m,d) = \big(s_1(m,d), \ldots, s_n(m,d)\big)$ для игр $\tilde{G}_m^d = \langle N, A, T, p, \tilde{u}\rangle$ такие, что $\lim\limits_{\substack{m\to 0\\ d\to 0}} S(m,d) = S$.*

Игру $\tilde{G}_m^d$ будем называть *модифицированной игрой*, а $S\,(m,d)$ — *равновесием в модифицированной игре*.

Для данных сетевых рынков проведен численный расчет модифицированных равновесий и равновесий в модифицированных играх. Вычислительный эксперимент показал, что в модифицированных играх имеется только по одному равновесию.

На базе системы Z-Tree [4] были написаны программы, реализующие данные сетевые аукционы. С их помощью было проведено несколько серий лабораторных экспериментов в компьютерных классах на базе ВЦ РАН, МГИМО, лаборатории экспериментальной экономики МФТИ и центра переподготовки персонала ЦБ РФ. Результаты экспериментов согласуются с понятием модифицированного равновесия при предположении о нормальности отклонений от наилучшего ответа.

Работа выполнена при поддержке РФФИ, проекты №07-01-00605а, №06-01-08057-офи, и программы «Развитие научного потенциала высшей школы (2006–2008 годы)» Федерального агентства по образованию, код проекта РНП.2.2.1.1.2467.

Литература

[1] *Журавель Ю. Ю., Меньшиков И. С.* Двойной аукцион для сетевых рынков. — М.: ВЦ РАН, 2004.

[2] *Меньшиков И. С.* Анализ влияния психофизиологических параметров участников на агрегированное поведение рынка методами экспериментальной экономики // ММРО-13 (в настоящем сборнике). — 2007. — С. 497–499.

[3] *Myerson R.* Game Theory: Analysis of Conflict — 1991.

[4] *Fischbacher U.* Z-Tree — Zurich Toolbox for Readymade Economic Experiments // Experimenter's Manual, Working Paper No. 21, Institute for Empirical Research in Economics, University of Zurich, 1999.

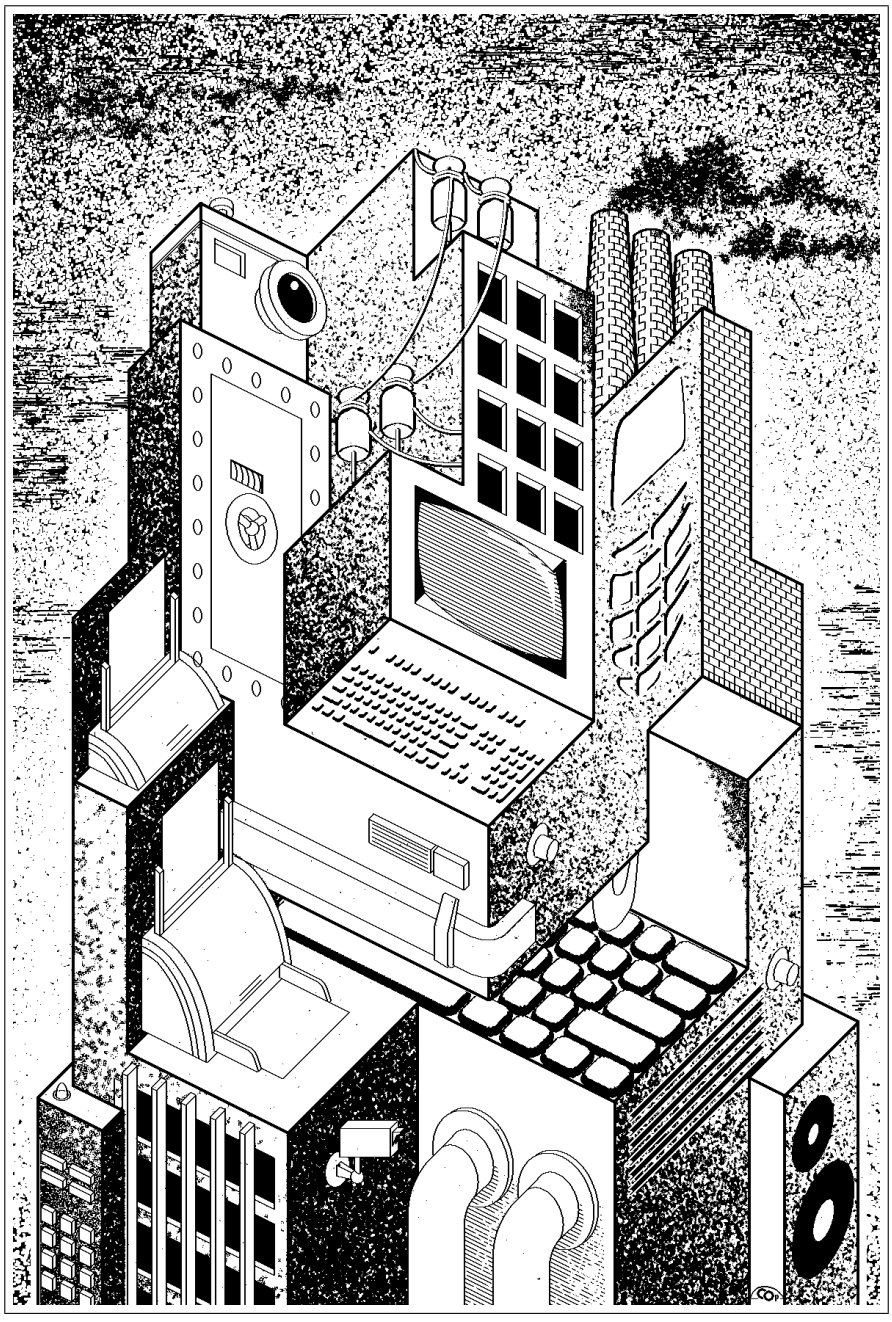
[5] *Selten R.* Reexamination of the perfectness concept for equilibrium points in extensive games // Int'l Journal of Game Theory. — No. 4, 1975. — Pp. 25–55.

[6] *Myerson R.* Refinements of the Nash equilibrium concept // Int'l Journal of Game Theory. — No. 7. — 1978. — Pp. 73–80.

Прикладные системы распознавания и прогнозирования

Код раздела: AS (Applied Systems)

- Реализации прикладных систем интеллектуального анализа данных.
- Информационные технологии.
- Применение методов распознавания в системах защиты информации.
- Средства поддержки вычислительных экспериментов.
- Средства научной визуализации.



Архитектура и разработка системы имитационного моделирования для многофакторных моделей социальной динамики

Бабкин Э. А., Козырев О. Р.
okozyrev@hse.nnov.ru
Нижний Новгород, Государственный университет, Высшая школа экономики

Постановка задачи

Мы сосредоточим наши исследования на решении важной научной проблемы по разработке единой методологии создания архитектуры информационно-вычислительных средств для математического, информационного и компьютерного моделирования сложных процессов в социальной динамике с учетом влияния многих факторов внешней среды.

Для эффективного разрешения указанных проблем требуются информационно-вычислительные средства нового поколения на основе современных технологий компонентных и распределенных вычислений. При разработке систем имитационного моделирования, за исключением единичных примеров системного, концептуального подхода к построению моделей, часто выбирается традиционный путь. Знания о структуре математической модели и границах ее применимости выражаются в виде описаний на естественном языке, а также неявно содержатся в программном коде, который самостоятельно создается авторами моделей. Это приводит к ряду серьезных проблем (практически невозможно получить отчуждаемое описание семантики соответствующей математической модели; ручное кодирование моделей противоречит современным тенденциям развития программной инженерии).

Особые требования предъявляются к моделям социальной и культурной сфер. Здесь для эффективного разрешения указанных проблем требуются разработанные модели и информационно-вычислительные средства нового поколения на основе знаний [2, 3]. Они должны позволять исследователям проводить разработку и интеграцию математических моделей, независящих от особенностей программной реализации, с последующей автоматизированной генерацией кода для проведения численных экспериментов в различных системах имитационного моделирования. Необходимо наличие единого системного подхода ко всем этапам математического моделирования, что обуславливает большую актуальность разработки методологии унифицированного семантического описания содержательных моделей и алгоритмов по их преобразованию в различные формы.

Отдельные задачи абстрактного системного описания математических моделей социально-экономических систем решаются в рамках

подхода «Системный анализ развивающейся экономики», развиваемого ВЦ РАН. Однако этот подход нацелен на описание ограниченного числа типов моделей экономики, а решение по сопровождению жизненного цикла распределенных систем моделирования и автоматизированная генерация программного кода агентов для многоагентных имитационных систем там отсутствуют.

Определенной степенью близости обладает также проект AuctionBot Мичиганского университета, США, но в нем отсутствуют общие принципы описания моделей взаимодействия. В ряде международных проектов ведется работа по систематизации и предоставлению доступа к различным информационным ресурсам на основе протокола Z39.50. Наиболее известными из них являются проекты Aquarelle и AHDS.

Предполагаемые подходы

В изучении феноменов социальной динамики важную роль играют функционально-структурные модели. При этом одним из наиболее перспективных направлений считается многоаспектное математико-информационное моделирование, воплощенное в форме распределенных имитационных систем на основе парадигмы взаимодействия индивидуальных сущностей (individual-based systems). Для построения и анализа таких моделей с помощью методов имитационного моделирования требуется наличие единой информационно-вычислительной системы, предоставляющей удаленным пользователям удобный интерфейс для самостоятельного расширения возможностей взаимодействия с системой, обладающей развитыми средствами автоматизированной разработки и анализа имитационных многоаспектных моделей. Такой подход носит название генеративного описания моделей (generative models description).

Задача, которая решается в данном проекте, состоит в построении методологии моделирования и программной архитектуры распределенной информационно-вычислительной системы для генеративного описания и изучения многофакторных моделей социальной динамики. Эти функции предполагается эффективно реализовать на основе развития принципов многоагентного моделирования и средств распределенного программирования в стандарте CORBA, развивая полученные ранее результаты в области распределенных систем имитационного моделирования [2, 3].

Результатом явится достижение следующих научных целей:

— создание новой методологии построения распределенных систем имитационного моделирования на основе совместного применения CORBA и многоагентных систем;
— получение объективных критериев производительности класса подобных систем имитационного моделирования;

— разработка новых математических моделей и алгоритмов масштабируемых вычислительных экспериментов на моделях социальной динамики в локальных сетях.

Новизна решаемой задачи обусловлена отсутствием соответствующего программного и информационного обеспечения и цельной методологии проведения имитационных экспериментов в социально-экономических исследованиях. Нами развиваются методы специализации группы современных информационных технологий (CORBA и многоагентные системы) для решения специфических задач, возникающих в ходе проведения имитационных экспериментов с многофакторными функционально-структурными моделями социальной динамики. Особенность используемых подходов заключается в применении теоретических принципов декларативного унифицированного описания семантики математических моделей, а также алгоритмов интеграции и трансформации декларативных описаний моделей в набор программных агентов для «сквозной» автоматизации задач имитационного моделирования.

Важным научным вкладом предлагаемого проекта будут являться новые принципы описания структуры и поведения распределенной информационно-вычислительной системы, использующей архитектуру CORBA и технологию многоагентных вычислений, с учетом требований математико-информационного моделирования социальных систем. Эта система предположительно будет состоять из трех составных частей:

— подсистема декларативных описаний семантики математических моделей;
— подсистема сопровождения жизненного цикла;
— подсистема проведения имитационных экспериментов и справочная подсистема (Web-портал).

Для взаимодействия подсистем планируется использовать объектно-ориентированные распределенные технологии: связь с имитационным сервером реализовать с помощью технологии CORBA (TAO ORB), Web-портал является сервером приложений, построенным на принципах трехуровневой архитектуры и обеспечивающим четыре вида открытых интерфейсов для внешних пользователей и внутренних подсистем (динамический HTML, Java-апплеты, EJB, SOAP-CORBA). Банк знаний обеспечивает функции специализированного графического редактора моделей и функции по долговременному хранению описаний моделей в СУБД. Для реализации банка знаний используется язык Java и объектно-реляционная СУБД PostgreSQL. В качестве лингвистических средств в банке знаний используются известные стандарты представления знаний: XML, RDF, DAML+OIL. Будет также обеспечена возможность экспорта знаний в различном представлении (онтологии, модели) из ряда широко

распространенных инструментальных средств (в частности, из системы Protege). Для построения эффективного имитационного сервера применяется технология многоагентных распределенных систем с программной реализацией на основе расширения системы SWARM.

Таким образом, автоматизированная генерация кода по моделям будет приводить к созданию наборов автономных агентов. Их реализация будет сохраняться в информационно-логической системе для повторного использования в составе других моделей. Работа выполнена при финансовой поддержке гранта РФФИ № 07-07-00058.

Литература

[1] *Бабкин Э. А., Козырев О. Р.* Система моделирования микроэкономических сценариев: общая концепция и принципы программной реализации // Известия АИН РФ. Сер. Прикладная математика и информатика, Т. 6. Москва, Н. Новгород, 2006. — С. 22–30.

[2] *Бабкин Э. А., Козырев О. Р., Куркина И. В.* Принципы и алгоритмы искусственного интеллекта // Н. Новгород: Изд-во НГТУ, 2006. — 132 С.

[3] *Бабкин Э. А., Козырев О. Р.* Методы представления знаний и алгоритмы поиска в задачах искусственного интеллекта. — Н. Новгород, Изд-во Талам, 2005. — 146 С.

[4] *Babkin E. A., Kozyrev O. R., Logvinova K. V., Zubov M. L.* Ontology-based Modeling of Micro Economics Scenarios // Proceedings of BIR-2004, Shaker Verlag, 2004. — p. 33–44.

Разработка прототипа интеллектуальной системы прогнозирования исхода беременности

Берестнева О. Г., Шаропин К. А., Добрянская Р. Г., Муратова Е. А.

shar@am.tpu.ru

Томск

К настоящему времени накоплены факты, свидетельствующие о том, что мать и ребёнок представляют собой единый нейрогуморальный организм. Каждый из них в равной степени страдает от неблагоприятного влияния внешнего мира, которое записывается в долговременной памяти, оказывая воздействие на всю последующую жизнь ребёнка. Основной тенденцией перинатальной демографии, наблюдаемой в России в течение последних лет, является рост патологии новорожденных. В медицинской модели подробно изучено значение нервной системы в регуляции родовой деятельности, в то время как изучению психологической готовности к родам уделяется недостаточное внимание. Представленная в докладе интеллектуальная система прогнозирования исхода беременности SPM

построена на основе выявленных авторами особенностях течения родов у женщин в зависимости от их психофизиологических особенностей и стратегий копинг-поведения [1–3].

Структура системы SPM

Авторами разработан прототип интеллектуальной системы SPM, предназначенной для решения следующих задач:

1) создания и ведения базы данных психофизиологического состояния беременных женщин;
2) компьютерной обработки и анализа результатов психодиагностического тестирования;
3) выделения групп психологического риска на уровне женской консультации и формирования рекомендаций по программам дифференциальной дородовой подготовки;
4) изучения особенностей течения беременности у женщин с различными психологическими и психофизиологическими особенностями.

Система имеет дружественный пользовательский интерфейс, гибкую систему настроек, а также все составляющие современной модели информационной системы: сбор, хранение, обработку, передачу, выдачу и защиту информации. Система спроектирована с использованием реляционной базы данных, реализованной в СУБД MS Access, с включением программных модулей, написанных на языке Visual Basic for Aplication.

Формирование базы знаний

Для решения задач выделения групп психологического риска на уровне женской консультации и задач изучения особенностей течения беременности в систему интегрированы разработанные авторами алгоритмы и программы на основе методов Data Mining. Этот выбор обусловило наличие в экспериментальном материале количественных, порядковых и качественных признаков и необходимость их совместного анализа. В системе используются:

1) информационная технология извлечения и структурирования знаний, объединяющая несколько методов интеллектуального анализа данных с последующей интеграцией выявленных логических закономерностей [3];
2) усовершенствованный метод локальной геометрии (за счет проведения на начальных этапах конструирования логических правил визуального анализа геометрической структуры исходных и преобразованных данных) [4];

3) технология нахождения устойчивых логических закономерностей (относительно выбранных методов решения), представляющая собой новый подход для интеграции логических моделей и позволяющая исключать ложные закономерности, которые могут проявлять себя в силу имеющихся особенностей исследуемых данных [5].

Получено два вида прогностических моделей: в виде деревьев решений и в виде дискриминантных функций. Для формирования рекомендаций по дифференцированной психологической коррекции беременных женщин в базу знаний будут включены результаты исследований, представленные в [6, 7].

Заключение

Полученные в работе результаты позволили разработать прототип интеллектуальной системы прогнозирования исхода беременности. Дальнейшее развитие системы предполагает наряду с выявлением «групп риска» решать задачу формирования рекомендаций психологам женских консультаций по видам дифференцированной психологической коррекции и психологической поддержке беременных женщин.

Работа выполнена при поддержке РГНФ, проект № 07-06-12143в.

Литература

[1] *Добрянская Р. Г.* Исследование стратегий преодоления эмоционального стресса у беременных женщин // Сибирский психологический журнал. — 2003. — № 18. — С. 65–67.

[2] *Добрянская Р. Г., Евтушенко И. Д.* Влияние дородовой подготовки беременных женщин с различными стратегиями адаптации на поведение в родах, послеродовом периоде и здоровье новорожденных // Вест. перинатологии, акушерства и гинекологии, Красноярск: КрасГМА, 2004. — С. 53–60.

[3] *Берестнева О. Г., Добрянская Р. Г., Муратова Е. А.* Применение технологии Data Mining для прогнозирования исхода родов // Математические методы распознавания образов-12, М.: МАКС Пресс, 2005. — С. 260–264.

[4] *Муратова Е. А., Берестнева О. Г., Янковская А. Е.* Анализ структуры многомерных данных методом локальной геометрии // Известия Томского политехнического университета. — 2003. — Т. 306, № 3. — С. 19–23.

[5] *Муратова Е. А., Берестнева О. Г.* Применение технологии конструирования диагностических шкал в задачах психологии интеллекта // Труды межд. науч.-тех. конф. «Интеллектуальные системы» (IEEE AIS'04) и «Интеллектуальные САПР» (CAD-2004), М.: Физматлит, 2004. — С. 223–228.

[6] *Добрянская Р. Г., Залевский Г. В., Евтушенко И. Д.* Дифференцированная медико-психологическая подготовка беременных к родам в условиях женской консультации // Пятый Всероссийского конгресс по пренатальной

и перинатальной психологии, психотерапии и перинатологии. — М.: Издательство Института психотерапии, 2005. — С. 72–74.

[7] *Добрянская Р. Г., Евтушенко И. Д., Залевский Г. В.* Система дифференциальной медико-психологической помощи беременным женщинам. Методические рекомендации для врачей акушеров-гинекологов и психологов женских консультаций. — Томск: Издательство ТГУ, 2005. — 58 с.

Система эмпирического измерения качества алгоритмов классификации

Воронцов К. В., Инякин А. С., Лисица А. В.

voron@ccas.ru, inyakin@forecsys.ru, lisitsa@forecsys.ru

Москва, Вычислительный Центр РАН, ЗАО «Форексис»

Тестирование в режиме скользящего контроля является стандартной методикой сравнения алгоритмов классификации. Существующие системы поддержки научных исследований MatLab, R, DELVE, WEKA, и др. позволяют проводить такое тестирование, но требуют от пользователя достаточно высокой квалификации. На основе технологии, описанной в [2], авторами разрабатывается открытая распределённая система `AxTTA` (читается «акста», от Algorithms × Tasks Testing Area — полигон для тестирования алгоритмов на задачах), предоставляющая доступ к задачам, методам и результатам тестирования через интуитивно понятный web-интерфейс. Система предназначена как для специалистов по анализу данных, так и для экспертов-прикладников. Она позволяет полностью автоматизировать типовое исследование, цель которого — выяснить, какой из известных методов лучше подходит для решения конкретной прикладной задачи классификации или класса задач.

Набор задач взят из общедоступного репозитория UCI [3] и может пополняться пользователями. Алгоритмы запускаются на удалённых вычислительных серверах. Формирование выборок данных и оценивание качества классификации производится центральным сервером. Результаты тестирования представляются в виде таблицы «алгоритмы × задачи». Пользователь может формировать наборы отображаемых алгоритмов и задач, просматривать исходные данные задач, задавать управляющие параметры алгоритмов, назначать состав информации, отображаемой в ячейках таблицы. Обычно при сравнительном анализе алгоритмов классификации в таблицу выводятся только оценки скользящего контроля. Система `AxTTA` для каждой ячейки вычисляет расширенный набор критериев, позволяющий глубоко исследовать особенности как алгоритмов, так и задач. В сообщении рассматриваются некоторые методологические аспекты системы `AxTTA`.

Класс решаемых задач. Пусть X — множество объектов, Y — конечное множество имён классов, $X^L = \{(x_i, y_i)\}_{i=1}^L \subset X \times Y$ — выборка длины L. Объекты x из X описываются признаками $f_1(x), \ldots, f_n(x)$, возможно, разнотипными. *Задача классификации* задаётся $L \times n$-матрицей данных $\big[f_{ij}\big] = \big[f_j(x_i)\big]$ и целевым вектором $\big[y_i\big]$. Дополнительно может быть задана матрица потерь $\big[C_{yy'}\big]$, где $C_{yy'}$ — штраф за отнесение объекта класса y к классу y', а также некоторая априорная информация о признаках. Матрица данных может содержать пропуски. Требуется построить алгоритм классификации $a\colon X \to Y$, аппроксимирующий неизвестную целевую зависимость $y(x)$ на всём множестве X.

Метод обучения μ по обучающей выборке $X^\ell \subseteq X^L$ строит алгоритм классификации $a = \mu(X^\ell)$.

Качество алгоритма a на конечной выборке U характеризуется *частотой ошибок* $\nu(a, U) = \frac{1}{|U|} \sum_{x \in U} \big[a(x) \neq y(x)\big]$.

Процедура скользящего контроля является основой для вычисления большинства критериев. Производится N разбиений выборки X^L на обучающую подвыборку длины ℓ и контрольную длины k, $X^L = X_n^\ell \cup X_n^k$, $L = \ell + k$, $n = 1, \ldots, N$. *Оценка скользящего контроля* для функции $\xi\colon \{1, \ldots, N\} \to \mathbb{R}$ определяется как среднее $\hat{\mathsf{E}}\xi = \frac{1}{N} \sum_{n=1}^N \xi(n)$. Разбиения строятся по стандартной методике $t \times q$-fold cross-validation [5]: генерируется t случайных разбиений выборки X^L на q блоков примерно равной длины и равными долями классов, и каждый блок поочерёдно становится контрольной выборкой. Таким образом, $N = tq$ и $k = \frac{L}{q}$ с точностью до округления.

Качество классификации на n-м разбиении характеризуется частотой ошибок на обучении $\nu_n^\ell = \nu(a_n, X_n^\ell)$ и на контроле $\nu_n^k = \nu(a_n, X_n^k)$, где $a_n = \mu(X_n^\ell)$. Обобщающая способность метода μ на выборке X^L характеризуется одной из *оценок скользящего контроля*

$$\mathrm{CV}(\mu, X^L) = \hat{\mathsf{E}}\nu_n^k; \qquad \mathrm{CV}_\varepsilon(\mu, X^L) = \hat{\mathsf{E}}\big[\nu_n^k - \nu_n^\ell > \varepsilon\big].$$

Графики зависимости CV и CV_ε от ℓ при фиксированном k позволяют оценивать достаточную длину обучения для данного метода в данной задаче. График $\mathrm{CV}_\varepsilon(\varepsilon)$ позволяет оценивать риск переобучения.

Разложение ошибки на вариацию и смещение (bias-variance decomposition) [5]. Введём функцию *среднего предсказания*

$$\tilde{y}(x) = \arg\max_{c \in Y} \hat{\mathsf{E}}\big[a_n(x) = c\big], \quad x \in X.$$

Назовём $B(x) = \big[\tilde{y}(x) \neq y(x)\big]$ *смещением* метода μ на объекте $x \in X$. Соответственно, объекты выборки X^L разделятся на смещённые и несмещённые. Для произвольной конечной выборки $U \subset X$ определим среднее

смещение $B(U) = \frac{1}{|U|}\sum_{x\in U} B(x)$. Тогда имеет место разложение:

$$\mathrm{CV}(\mu, X^L) = \hat{\mathsf{E}}B(X_n^k) + V(\mu, X^L),$$

где первое слагаемое характеризует смещённость модели классификации, используемой в методе μ; второе слагаемое, называемое *вариацией*, характеризует изменчивость результата обучения по отношению к составу обучающей выборки. Если первое слагаемое велико, то надо менять саму модель. Если второе слагаемое велико, то качество классификации можно улучшить путём регуляризации или композиции алгоритмов.

Профиль устойчивости показывает, насколько изменяются классификации получаемого алгоритма, если состав обучающей выборки изменяется на m объектов:

$$S_m(\mu, X^L) = \hat{\mathsf{E}}_n\hat{\mathsf{E}}_{n'}\frac{1}{|X_{nn'}|}\sum_{x\in X_{nn'}}\big[\rho(X_n^\ell, X_{n'}^\ell) = m\big]\big[a_n(x) \neq a_{n'}(x)\big],$$

где $m = 1, \ldots, \min\{\ell, k\}$, $X_{nn'} = X_n^k \cap X_{n'}^k$, $\rho(U, V)$ — число несовпадающих объектов в выборках U и V. График профиля устойчивости, как правило, монотонно возрастает. Чем ниже проходит начальный (левый) участок профиля, тем устойчивее обучение.

Профиль разделимости определён для вещественнозначных алгоритмов классификации вида $a(x) = \arg\max\limits_{c\in Y}\Gamma_c(x)$, где $\Gamma_c(x)$ — оценка принадлежности объекта x классу c. Степенью граничности или *отступом* (margin) объекта $x_i \in X^L$ называется величина [4]

$$M(a_n, x_i) = \Gamma_{n,y_i}(x_i) - \max_{c\in Y\setminus y_i}\Gamma_{n,c}(x_i).$$

Отступ показывает, насколько близко объект x_i подходит к границе класса y_i. Если объект оказывается за границей, то отступ отрицателен, и на данном объекте алгоритм допускает ошибку. Чем больше отступы, тем лучше качество классификации [4]. Эмпирические распределения отступов (*профили разделимости*) на обучающих и контрольных данных показывают, насколько надёжно данный метод разделяет классы.

Профиль представительности объектов. Для каждого объекта x_i из X^L вычисляется доля разбиений, при которых данный объект попадает в контроль, и на нём допускается ошибка:

$$I_i(\mu, X^L) = \hat{\mathsf{E}}\big[x_i \in X_n^k\big]\big[a_n(x_i) \neq y_i\big], \quad i = 1, \ldots, L.$$

Упорядоченная последовательность значений $I_1 \geqslant \ldots \geqslant I_L$ называется *профилем представительности* объектов. В начальный участок профиля попадают *шумовые* объекты, для которых $I_i > 0.5$. Объекты, оказавшиеся шумовыми для многих методов, по всей видимости, объективно

являются таковыми. Система AxTTA позволяет отбросить шумовые объекты и провести обучение только по представительным объектам.

Профиль информативности признаков — это зависимость CV от числа использованных признаков. Отбор признаков осуществляется последовательным отбрасыванием наименее значимых признаков. Эта процедура подходит для любого метода обучения, но ресурсоёмка, что, впрочем, вполне приемлемо в системе распределённых вычислений.

Временны́е показатели работы алгоритма. Поскольку в системе AxTTA обучение производится многократно на подвыборках разной длины, появляется возможность аппроксимировать время обучения зависимостью вида $T(\ell) = T_0 \ell^{\alpha} (\ln \ell)^{\beta}$, где коэффициенты α и β показывают эффективность алгоритма, а множитель T_0 зависит от параметров вычислительного сервера и не представляет интереса.

Сравнительный анализ методов для фиксированной задачи. Система AxTTA позволяет провести сравнительный анализ заданного пользователем набора методов по заданному набору критериев. При этом соответствующие разным методам таблицы сводятся вместе, а графики — накладываются и изображаются в одном масштабе.

Работа выполнена при поддержке РФФИ, проекты №№ 07-07-00372, 07-07-00181, 05-07-90410, 05-01-00877, а также программы ОМН РАН Алгебраические и комбинаторные методы математической кибернетики.

Литература

[1] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. — 2004. — № 13. — С. 5–36.

[2] *Качалков А. В., Хачай М. Ю.* Квазар-Оффлайн. Распределенный вычислительный комплекс для решения задач распознавания образов // ММРО-13 (в настоящем сборнике). — 2007. — С. 591–594.

[3] *Asuncion A., Newman D. J.* UCI Machine Learning Repository — University of California, Irvine. — 2007. — www.ics.uci.edu/~mlearn/MLRepository.html.

[4] *Garg A., Roth D.* Margin distribution and learning algorithms // Int. Conf. on Machine Learning (ICML'03), Washington, DC USA. — 2003. — Pp. 210–217.

[5] *Webb G. I.* MultiBoosting: A technique for combining boosting and wagging // Machine Learning. — 2000. — Vol. 40, No. 2. — Pp. 159–196.

Сетевая геоинформационная технология комплексного анализа и прогнозирования

Гитис В. Г., Шогин А. Н.
gitis@iitp.ru, alex@viniti.ru
Москва, ИППИ РАН, ВИНИТИ РАН

В последние годы начаты работы по созданию распределенной геоинформационной среды. В частности, в 2006 году ведущей в области ГИС компанией ESRI запланирована разработка распределенной среды GeoWeb (ArcReview №1 (36) 2006). Это направление исследований в России инициировано в 2005 году программой Президиума РАН «Электронная Земля». В 2006 году впервые разработана базовая версия сетевой геоинформационной среды. Одним из важнейших элементов среды «Электронная Земля» являются распределенные сетевые ГИС ГеоПроцессор 2, КОМПАС V (`http://www.geo.iitp.ru`) и ГеоТайм II (β-версия, `http://www.geo.iitp.ru`). Функциональность этих систем ориентирована на решение двух типов задач:

1) просмотр многодисциплинарной географической информации (ГИ) и оценивание связей между ее компонентами;
2) нахождение многомерных зависимостей в ГИ, прогнозирование, обнаружение и распознавание целевых стационарных и динамических свойств изучаемой среды.

Технология сетевых ГИС ориентирована на пользователя, не являющегося специалистом в области ИТ. Поэтому для сетевых ГИС важно поддерживать свои аналитические возможности с помощью интуитивно понятных методов обработки данных. Перечислим еще несколько существенных требований к сетевым ГИС технологиям.

Первое требование состоит в поддержке комплексного геоинформацоного анализа разнотипных данных. Необходимость комплексного анализа обусловлена тремя факторами:

1) взаимодействием рассматриваемых процессов;
2) невозможностью прямых измерений их ключевых характеристик;
3) недостаточным объемом наблюдений и воздействием на результаты измерений шумов.

В этой ситуации для поиска устойчивых решений требуются методы, позволяющие комплексно использовать все доступные данные и экспертное знание. Второе требование относится к необходимости интегрировать ГИ из различных хранилищ данных. Обычно эти хранилища распределены в местах сбора и актуализации данных на серверах и на ПК пользователя (последнее обеспечивает конфиденциальность данных пользователя). Следующее требование обусловлено необходимостью гибкого изме-

нения функциональности ГИС пользователем с помощью подключения распределенных плагинов. Четвертое требование относится к обеспечению возможности сохранения результатов пользователя (ГИС слоев и ГИС-проекта). Пятое требование — обеспечение высокой интерактивности и высококачественной графики. Последнее требование — совместимость форматов данных с ГИС стандартами.

В современных ИТ перечисленные требования успешно реализуются либо с помощью сетевых приложений, либо с помощью апплетов. Сетевые ГИС ГеоПроцессор 2, КОМПАС V и ГеоТайм II написаны на языке Java 1.5. Это обеспечивает независимость программы от платформы пользователя, а также дает возможность ее выполнения в Java-машине либо как апплет в любом современном браузере, либо в качестве локальной программы.

Рассмотрим функциональность ГИС ГеоПроцессор 2, одной из сетевых ГИС среды «Электронная Земля». Эта система предназначена для пользователей с интересами от ознакомления on-line с комплексами геоданных до решения сложных геоинформационных проблем, таких как оценка природной опасности, прогноз природных ресурсов, оценка экологического состояния среды.

Система ГеоПроцессор 2 поддерживает следующие функции интерактивной обработки и комплексного анализа векторных (точки, линии, полигоны) и сеточных пространственных данных:

1. *Операции ввода/вывода:*
   - Загрузка данных распределенных на сетевых серверах и на ПК пользователя по XML файлу ГИС-проекта.
   - Динамическая загрузка векторных и сеточных (растровых) данных в форматах BIN (ГеоПроцессор), FLT (ASCII), SHP (ESRI), PTS (ASCII) и JPG.
   - Сохранение ГИС-проекта и полученных данных.
   - Выбор проекции карты, сохранение и вывод на печать.
2. *Визуальное исследование растровых и векторных данных:*
   - Композиция карты из растровых и векторных слоев.
   - Изменение размеров и масштаба карты с интерполяцией и без интерполяции сеточных слоев.
   - Динамическое изменение закраски, прозрачности, диапазона видимых значений, типа линий и размеров пиктограмм.
   - Картографическое измерение сеточных и векторных слоев.
   - Построение разрезов сеточных слоев карты по произвольному профилю и измерение значений по разрезу.
   - Моделирование освещенности.

- Формирование выборок прецедентов в виде совокупностей единичных точек и/или полигонов с помощью указания объектов на карте и с помощью автоматического выбора прецедентов по сеточному или точечному слою.
- Комплексный анализ по сходству: функция сходства формируется с помощью функции расстояния до прецедентов или как принадлежность к построенным по прецедентам полуинтервалам.
- Оценивание статистик одного или двух сеточных слоев (минимум, максимум, среднее, среднеквадратичное отклонение, гистограмма, корреляция и ошибка аппроксимации).
- Редактирование координатной сетки.

3. *Преобразование данных:*
   - Преобразование векторный слой → сеточный слой (вычисление сеточных полей расстояний до векторных объектов, близости, плотности и т.д.)
   - Преобразование сеточный слои → сеточный слой с помощью операций растровой фильтрации (вычисление градиента, сглаживание в произвольном скользящем окне, вычисление среднеквадратичного отклонения, выделение аномалий и т.д.) или с помощью вычисления конструируемых пользователем произвольных функций от нескольких сеточных слоев.
   - Преобразование сеточный слой и векторный слой → атрибуты векторного слоя с помощью вычисления статистик сеточного слоя в буферных зонах векторных объектов или с помощью конструируемых пользователем произвольных функций от атрибутов нескольких векторных слоев.
4. *Пространственный правдоподобный вывод:*
   - Оценивание функции сходства к выборке прецедентов.
   - Оценивание функции принадлежности к двум классам.
   - Оценивание непараметрической регрессии.
   - Оценивание функции распределения и нахождение логического выражения, объясняющего полученное решающее правило.

Работа выполнена при поддержке РФФИ, проекты № 00-07-90100, № 06-07-89139 и программы Президиума РАН «Электронная Земля».

Разработка и создание системы распознавания лиц с помощью объемных фотороботов на основе общедоступных установок виртуальной реальности

Дружинин А. А., Клименко С. В., Протасов В. И., Потапова З. Е.
adruzhinin@gmail.com
Москва

Развитие современных информационно-коммуникационных технологий привело к принципиально новым возможностям их использования для решения сложных трудно формализуемых задач. К таким задачам относятся, например, принятие решений коллективом экспертов в условиях, когда интеллектуальных ресурсов отдельного индивида не достаточно для полного и точного решения.

В настоящей работе исследуется возможность применения новой информационной технологии «генетического консилиума» [1, 3, 4] для восстановления субъективного портрета коллективом свидетелей с использованием объемного морфинга лица.

Имеющиеся системы создания субъективных портретов (фотороботов) основаны на плоских изображениях и не дают полного представления о лице, находящемся в розыске, с другой стороны, сам процесс создания фотороботов становится значительно эффективнее, если опознавателю предоставляется возможность рассмотрения восстанавливаемого лица с любого ракурса, не ограниченного положениями «анфас» и «профиль». Применение коллективных методов принятия решения к созданию фотороботов еще в большей степени приближает субъективный портрет к оригинальному изображению опознаваемого.

В настоящей работе были решены следующие задачи:

1. В качестве метода, координирующего совместную работу группы экспертов, был выбран и исследован метод генетического консилиума [1].
2. В качестве программы 3D-визуализации человеческой головы была выбрана и исследована программа FaceGen Modeller компании Singular Inversions [2].
3. Составлены правила взаимодействия и инструкции для экспертов в генетическом консилиуме при восстановлении субъективного портрета.
4. Проведено исследование качества решения поставленной задачи коллективом экспертов при совместном использовании методики генетического консилиума, программы 3D моделирования лица и технологии объемной визуализации Пульфрих-стерео. (Эффект Пульфриха — это оптическая иллюзия, которая базируется на том факте, что

мозг чуть дольше распознаёт тёмные оптические раздражители, чем светлые. Распознаваемый объект на мониторе непрерывно движется или вращается. В специальных очках одно стекло затемнено. Хотя оба глаза видят одну и ту же картинку, «затемнённый» глаз передаёт картинку в мозг чуть позже. Мозг, восстанавливает при этом псевдо-стереоизображение. Постоянное медленное вращение изображения обеспечивает программа FaceGen Modeller).

Для проверки эффективности и работоспособности разработанной комплексной методики был проведен ряд экспериментов по восстановлению субъективного портрета различными группами студентов. Каждый эксперт работал на отдельном компьютере и проходил предварительно тренинг по моделированию лица в программе FaceGen. Десятки проведенных экспериментов показали хорошую сходимость метода (от 3 до 7 итераций).

Для проверки устойчивости метода к некорректным данным в некоторых экспериментах одному из экспертов было дано задание намеренно вносить искажения в свой вариант, отдаляющие изображение лица от искомого. Для восстановления оригинала при этом потребовалось лишь большее количество итераций. Также были проведены эксперименты с использованием данной методики с одним свидетелем. Было показано, что в этом случае метод выступает своеобразным «усилителем интеллекта» одиночного эксперта при генерации фоторобота.

Работа выполнена при поддержке РФФИ, проект № 05-07-90346.

Литература

[1] *Протасов В. И., Панфилов Д. С., Здоровеющев Ю. Ю.* Генерация фоторобота с помощью сетевого человеко-машинного интеллекта. // Международная научно-техническая конференция «Интеллектуальные многопроцессорные системы ИМС-99», Таганрог, 1999. — С. 106–107.

[2] FaceGen Modeller 3.1. — `http://www.facegen.com/modeller.htm`.

[3] Генерация новых знаний сетевым человеко-машинным интеллектом. Постановка проблемы. // Нейрокомпьютеры. Разработка и применение. — 2001. — № 7–8.

[4] *Протасов В. И.* Метасистемный эффект самоорганизации интеллекта более высокого уровня из искусственных и естественных компонентов. // Сборник научных трудов IV Всероссийской научно-технической конференции «Нейроинформатика 2002», Москва, 2002. — С. 33–40.

Визуализация процессов обучения нейронных сетей

Емельянова Ю. Г., Малышевский А. А., Хачумов В. М.
vmh@vmh.botik.ru, Scorpio@Zhukovsky.net, tajra@mail.ru
Переславль-Залесский, Институт программных систем РАН

Рассматривается задача когнитивной графической поддержки процессов настройки искусственных нейронных сетей (ИНС). Визуализация позволяет непосредственно наблюдать за ходом обучения ИНС, предоставляя в наглядном виде информацию о степени готовности и возникающих аномалиях средствами машинной графики, что позволяет пользователю оперативно принимать решение [1, 2].

Постановка задачи

Рассмотрим постанову задачи для двух принципиально различных по структуре ИНС. Данному условию соответствуют, например, сеть с единственным выходом, реализующая логическую функцию XOR [2], и однослойный персептрон с десятью выходами, предназначенный для распознавания цифр. В первом случае используется генетический алгоритм настройки и активационная функция вида «сигмоид», а во втором — алгоритм Видроу–Хоффа и функция активации типа «линейный скачок». Требуется отобразить графически динамику процессов обучения нейронной сети. Для визуализации процессов обучения нейронной сети, прежде всего, необходима информация о весах синапсов — однонаправленных входных связей отдельных нейронов, которые подстраиваются в процессе обучения.

Визуализация настройки сети XOR

Поставим в соответствие сети (Рис. 1) вектор из семи прямоугольных полутоновых областей. Примем, что белый цвет обозначает ноль, оттенки синего — отрицательные значения, оттенки красного — положительные значения коэффициента. Чем больше по модулю значение весового коэффициента, тем ярче цвет соответствующего прямоугольника. Под областью вектора расположен зеленый квадрат, который «мигает» при каждой четной итерации. Это представление позволяет сопоставить скорость роста фитнес-функции со скоростью совершения итераций; помогает заметить ситуацию, при которой веса синапсов не меняются или меняются незначительно от итерации к итерации. Значение фитнес-функции отображается в виде оранжевой полосы. Чем ярче цвет и чем больше ее длина, тем ближе значения полученного и требуемого выходов. При удачном завершении настройки на экране появляется информация о количестве совершенных итераций, таблица входов и соответствующих им выходов (Рис. 2). Интерпретация получаемых на экране цветовых

картин дана в работе [1] и в настоящей статье модифицируется следующим образом.

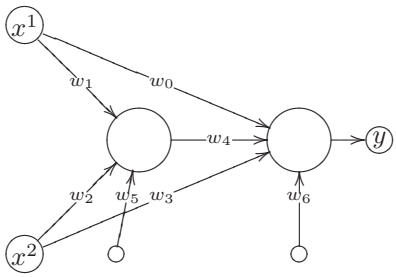


Рис. 1. ИНС XOR.

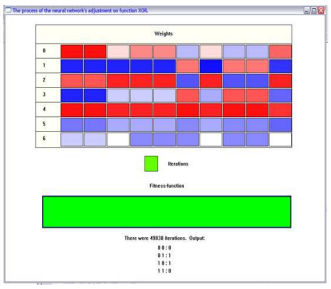


Рис. 2. Вид окна при успешной настройке сети.

1. Если значение фитнес-функции не изменяется от итерации к итерации и большинство ячеек ИНС окрашено ярким цветом, то сеть впала в «паралич».
2. Если значение фитнес-функции не меняется от итерации к итерации и большинство полос имеет некоторую промежуточную окраску, также не изменяющуюся во времени, то сеть вошла в локальный минимум.
3. Если когнитивный образ меняется от итерации к итерации, а фитнес-функция при этом практически не растет, то сеть не настраивается.

Визуализация настройки персептрона

Рассмотрим визуализацию настройки персептрона (Рис. 3, 4), предназначенного для распознавания цифр. Графический динамический образ входа нейрона строится в виде цветного прямоугольника (полосы) с изменяющейся толщиной. Для представления знака весового коэффициента применяется цвет: красный — для положительного, синий — для отрицательного. Толщина прямоугольника служит для представления динамики абсолютных значений весов: чем толще полоса, тем быстрее изменяются значения весов. С уменьшением толщины процесс стабилизируется, когда же полосы исчезают и более не появляются — веса нейрона практически настроены. Для представления выходов нейронов выбран подход с использованием полос синего и красного цветов, а также их оттенков. Оттенки прямоугольников меняются, отображая значения выходов. Каждый нейрон сети настраивается на эталонное изображение отдельно, поэтому, наблюдая настройку сети, мы видим сначала меняющееся изображение полосы первого нейрона, затем второго при статичном первом и т.д., как это показано на Рис. 4. При этом прямоугольник, отвечаю-

щий за первый нейрон, т.е. за тот, который уже настроен, приобретает зеленый цвет. Визуализация настройки персептрона, позволяет, как и в случае ИНС XOR, определять такие явления как паралич или локальный минимум. Ситуация, когда прямоугольник, отвечающий за выход с нейрона, окрашен в яркий цвет и не меняет свой оттенок длительное время, означает, что нейрон впал в паралич. Если оба прямоугольника, относящихся к одному нейрону, во время его настройки приняли статичное промежуточное положение, а нейрон все ещё является ненастроенным, то имеем локальный минимум.

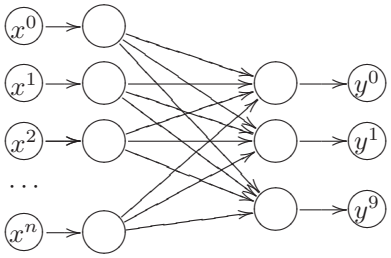


Рис. 3. Персептрон.

Рис. 4. Пример визуализации.

Заключение

В настоящей работе предложены подходы к визуализации процессов обучения некоторых нейронных сетей, имеющих различное назначение и архитектуру. Разработанные графические динамические представления весов нейронов ИНС и их выходов позволяют контролировать текущее состояние сети, оценивать качество и скорость обучения. Они дают общую картину внутреннего состояния ИНС не только в процессе обучения сети, но и в процессе ее непосредственного использования.

Работа выполнена при поддержке РФФИ, проект № 06-07-89083.

Литература

[1] *Таран А., Москаленко Ю.* Визуализация процесса обучения нейронной сети. — `http://old.festu.ru/ru/structure/library/Library/science/s127/article_9.htm`.

[2] *Хачумов В. М.* Логические элементы на нейронах // Интеллектуальные системы и компьютерные науки, Москва: Издательство механико-математического факультета МГУ, 2006. — С. 297–300.

Интеграция методов добычи данных и вывода по прецедентам в медицинской диагностике и выборе лечения

Карпов Л. Е., Юдин В. Н.

В Институте системного программирования РАН ведётся работа по применению собственных научных наработок в различных прикладных областях, требующих новых подходов к решению давно поставленных, но всё ещё актуальных проблем.

Проблема повышения качества и сокращения сроков медицинской диагностики решается с применением различных методов. Авторами поставлена задача рассмотрения действенности методов добычи данных и вывода по прецедентам в реализации систем поддержки принятия решений при диагностике и выборе лечения.

В данном контексте диагностика рассматривается как оценка состояния пациента и отнесение его к одному из возможных классов, а процесс лечения может рассматриваться как адаптивное управление и трактуется как последовательность управляющих воздействий на больного.

Организм как объект управления не может быть описан с помощью относительно простой математической модели, поэтому методы вывода по прецедентам в таких обстоятельствах могут рассматриваться как весьма перспективные.

Вывод по прецедентам — метод принятия решений, в котором используются знания о ранее возникавших ситуациях или случаях (прецедентах), в данном случае — об аналогичных ситуациях с другими пациентами. При рассмотрении новой проблемы (текущего случая) отыскивается похожий прецедент. Вместо того, чтобы каждый раз искать решение сначала, делается попытка использовать решение, принятое в сходной ситуации, возможно, адаптировав его к текущему случаю.

В основе всех систем вывода по прецедентам лежит тот или иной способ измерения степени близости прецедента и текущего случая. От этой меры зависит объем множества прецедентов, которые нужно обработать для достижения удовлетворительной классификации или прогноза. Основным недостатком таких систем является произвол, который допускают системы при выборе меры близости. Кроме того, безосновательным выглядит распространение общей меры близости на выборку данных в целом.

В ИСП РАН разработана программная система Спутник Врача, обеспечивающая поддержку врачебных решений в диагностике и выборе лечения. Система позволяет врачу использовать опыт экспертов и аккумулировать собственный опыт путём накопления и интерпретации данных

о пациентах в виде прецедентов, которые описываются в виде совокупности показателей, диагнозов, выбранных методов лечения и их исходов.

В системе применён подход к отбору прецедентов и адаптации решения, основанный на привлечении дополнительных знаний о предметной области, или фонового знания, методами добычи данных. Авторы предлагают уйти от распространения общей меры близости на базу прецедентов в целом, вычисляя эту меру каждый раз заново при рассмотрении очередного случая. Такие меры являются локальными, так как они привязываются к текущему случаю и отображают сходства текущего случая с другими объектами. Введенная мера определяется отношениями между текущим случаем и соседними объектами, в частности, полнотой описания текущего случая.

В базе прецедентов вводятся отношения эквивалентности, которые выражают принадлежность оцениваемых объектов к каким-либо классам. Классы представляют номинальную шкалу. Объекты, отнесённые к одному классу, считаются эквивалентными с точки зрения данной номинальной шкалы. Классы могут быть построены различными способами: с помощью экспертного знания, на основе обучающей выборки, или путем предварительной кластеризации базы прецедентов. Они используются как основа для предлагаемой меры близости прецедента и текущего случая.

Для реализации системы было выбрано решение, состоящее из двух функциональных блоков: Оболочки и Классификатора. Оболочка – уровень системы, реализующий интерфейс с конечным пользователем в терминах, с которыми привык работать врач: пациент, показатель, заболевание, лечение, исход. Классификатор (прикладной слой) – уровень, выполняющий основные операции анализа данных: классификацию, оценку, прогнозирование и выявление зависимостей в данных. Основными сущностями, с которыми оперирует Классификатор, являются класс, объект и признак объекта.

В совокупности эти два уровня представляют собой презентационный и прикладной слои программного обеспечения, предназначенного для системной поддержки выполняемых работ.

Врач часто вынужден принимать решение при неполном наборе показателей пациента, в условиях неоднозначной классификации или в случаях, когда заболевание выходит за пределы своей обычной симптоматики. Опытный врач может ставить диагноз по небольшому количеству симптомов, учитывая скрытые связи между показателями, при отсутствии специфических проявлений идентифицируемых заболеваний. Именно этому и призвана способствовать система. В отличие от эксперт-

ных систем, Спутник Врача не диктует врачу, какое решение принимать, а лишь помогает ему сделать выбор.

Первое испытание система Спутник Врача прошла в Московском областном клиническом институте (МОНИКИ) и в институте сердечно-сосудистой хирургии им. А. Н. Бакулева, где с её помощью удалось формализовать некоторые хорошо известные медикам примеры из их повседневной практики.

Работа выполнена при поддержке РФФИ, проекты № 06-07-89098 и № 06-01-00503.

Литература

[1] *Карпов Л. Е., Юдин В. Н.* Методы добычи данных при построении локальной метрики в системах вывода по прецедентам. — Москва: Институт Системного Программирования РАН, Препринт, 2006. — 42 с.

[2] *Карпов Л. Е., Юдин В. Н.* Адаптивное управление по прецедентам, основанное на классификации состояний управляемых объектов // Труды ИСП РАН, Москва: ИСП РАН, 2007. — С. 135–155.

[3] *Torgeir Dingsoyr* Integration of Data Mining and Case-Based Reasoning — 1998. — `www.idi.ntnu.no/~dingsoyr/diploma/`.

Квазар-Оффлайн: распределенный вычислительный комплекс для решения задач распознавания образов

Качалков А. В., Хачай М. Ю.

`kachalkov@gmail.com, mkhachay@imm.uran.ru`

Екатеринбург, Институт математики и механики УрО РАН

Работа посвящена анализу и моделированию предметной области в рамках унифицированного процесса, с использованием универсального языка моделирования UML, построению универсальной модели для описания задачи распознавания образов, описанию нескольких конкретных алгоритмов распознавания в терминах модели, и разработке распределенного вычислительного поратала «КВАЗАР–Офлайн» для решения задач распознавания образов и анализа эмпирических данных.

Основные результаты работы

1. Разработана новая методология создания распределенных вычислительных порталов для различных предметных областей. В рамках разработанной методологии построена универсальная модель подсистемы размещения и мониторинга заданий, с использованием унифицированного процесса разработки и языка моделирования UML.

2. В соответствии с предложенной методологией, реализован оригинальный распределенный вычислительный комплекс для решения задач распознавания образов, который включает в себя:
   (a) подсистему размещения и мониторинга заданий, реализованную на базе системы управления контентом `DotNetNuke`:
      — оригинальный специализированный модуль `Quasar`, реализующий доступ к системе;
      — база данных Quasar в соответствии с построенной концептуальной моделью, хранимые процедуры и функции;
      — компоненты доступа к базе данных Quasar.
   (b) вычислительную подсистему:
      — оригинальный базовый асинхронный вычислительный компонент, в рамках подхода, основанного на событиях, облегчающий доступ к инфраструктуре вычислительного портала;
      — вычислительная служба, отвечающая за инициализацию и запуск текущих заданий на счет с использованием подключаемой технологии загрузки вычислительных алгоритмов.
3. Разработаны новая концепция и рекомендации по реализации вычислительных алгоритмов, использующих вычислительную библиотеку численного анализа `dnAnalytics` на базе платформы .NET Framework 2.0[1].

Базовый асинхронный компонент `QuasarEngine`

Для того, чтобы вычислительная служба `Quasar` могла загружать и запускать вычислительный алгоритм на счет, необходимо, чтобы вычислительный компонент `.NET` был унаследован от служебного компонента `QuasarEngine` (см. рис. 1).

Компонент `QuasarEngine` предназначен для того, чтобы предоставить разработчикам вычислительных компонентов необходимую функциональность для взаимодействия с инфраструктурой системы `Quasar` и существенно упростить процедуру разработки.

Компонент `QuasarEngine` реализует шаблон проектирования асинхронных вычислений, основанный на событиях[2]. Реализация асинхронных вычислений необходима для того, чтобы иметь возможность передать запрос от пользователя об отмене вычислений непосредственно в вычислительный алгоритм. Также, данная реализация позволяет в перспективе реализовать поддержку сохранения промежуточного состояния

[1] `http://msdn.microsoft.com/netframework`
[2] `http://msdn2.microsoft.com/en-us/library/wewwczdw.aspx`

Рис. 1. Архитектура базового асинхронного компонента `QuasarEngine`.

вычислений с последующим продолжением вычислений с момента сохранения. Это может быть полезно для недетерминированного распределения вычислительного времени между заданиями и для приостановки вычислений от одного задания к другому.

Ключевым методом компонента является виртуальный метод `CalculateWorker`, в переопределенной версии которого разработчик и реализиует свой вычислительный алгоритм, используя библиотеку численного анализа `dnAnalytics`. Остальные методы, свойства и события компонента `QuasarEngine` предназначены для поддержки шаблона асинхронных вычислений, основанного на событиях.

Так, метод `CalculateAsync` создает новую асинхронную операцию с помощью метода `CreateOperation` объекта `AsyncOperationManager` для задания с идентификатором `taskID`, и вызывает метод `CalculateWorker` в отдельном потоке, передавая входные параметры — `TaskParameters`, идентификатор вновь созданной асинхронной операции `AsyncOperation`, а также делегат `SendOrPostCallback`, представляющий из себя метод обратного вызова для передачи сообщения контексту синхронизации, который будет вызван по окончанию асинхронной операции.

Метод `CancelAsync` предназначен для остановки текущих асинхронных вычислений с помощью метода `PostOperationCompleted` класса `AsyncOperation`.

Метод `ReportProgress` используется для публикации промежуточных и окончательных результатов в базе данных `Quasar`.

Работа выполнена на основе исследований, проводимых в отделе математического программирования Институте математики и механики УрО РАН по государственной бюджетной теме № 01.2.00102387, проектам РФФИ № 04-01-00108, № 04-01-96104 и президента РФ по поддержке ведущих научных школ, гранты № НШ-792.2003.1, № НШ-5595.2006.1.

Система QPSLab для анализа и распознавания числовых последовательностей с квазипериодической структурой

Кельманов А. В., Михайлова Л. В., Хамидуллин С. А.

`kelm@math.nsc.ru, okolnish@math.nsc.ru, kham@math.nsc.ru`

Новосибирск, Институт математики им. С. Л. Соболева СО РАН

Сиситема QPSLab (Quasi-Periodic Sequences Laboratory) предназначена для решения широкого спектра задач апостериорной (off-line) обработки (анализа и распознавания) массивов зашумленных данных (числовых последовательностей или дискретных сигналов) — результатов измерения характеристик изучаемых объектов различной природы, имеющих квазипериодическую структуру.

Под числовой последовательностью с квазипериодической структурой подразумевается всякая числовая последовательность, включающая такие фрагменты (подпоследовательности подряд расположенных членов), которые имеют характерные детерминированные или стохастические свойства, причем для всех пар следующих друг за другом фрагментов разность между номерами первых членов последующего и предыдущего фрагментов лежит в фиксированном интервале. Последовательность, имеющая квазипериодическую структуру, или квазипериодическая последовательность — это последовательность с квазипериодической сменой своих свойств. Например, в случае детерминированной последовательности под сменой свойств понимается изменение формулы общего члена последовательности, а в случае стохастической — изменение ее вероятностных характеристик.

Числовая (одномерная или многомерная) квазипериодическая последовательность как математический объект является мощным средством для адекватного описания широкого класса временных процессов, возникающих во многих приложениях. Спектр задач, на решение которых ориентирована система QPSLab, типичен, в частности, для электронной разведки, дистанционного зондирования, геофизики, биометрики, медицинской и технической диагностики, обработки речевых сигналов, радио-

локации, гидроакустики, телекоммуникации, криминалистики, поиска по мультимедийным базам данных, и др.

Отличительная особенность системы QPSLab состоит в том, что в ней реализован нетрадиционный подход к помехоустойчивому компьютерному анализу и распознаванию числовых последовательностей, сущность которого состоит в апостериорном способе обработки последовательности (т.е. обработки после накопления всех доступных для наблюдения данных) в сочетании с формализацией содержательной задачи как задачи принятия решения (проверки гипотез).

В основе трех традиционных широко распространенных подходов лежат последовательный (on-line) и апостериорный способы обработки последовательности в сочетании с формализацией содержательной задачи как задачи оценивания (оптимальной фильтрации), а также последовательный способ обработки в комбинации с формализацией задачи как задачи проверки гипотез. При реализации традиционных подходов проблемы комбинаторной оптимизации как правило не возникают.

В противоположность этому, в системе QPSLab реализованы технологии, которые существенным образом опираются на решение специфических задач комбинаторной оптимизации с целью выбора наилучшего из множества допустимых решений, мощность которого растет экспоненциально при увеличении длины обрабатываемой последовательности. При этом основу системы составляют оригинальные полиномиальные алгоритмы с гарантированными (априори доказуемыми) оценками точности решения (см. [1–4] и цитированые там работы).

Первоначальная версия открытой для пополнения системы содержит алгоритмы (технологии) помехоустойчивой обработки последовательностей, включающих квазипериодически чередующиеся ненулевые информационные фрагменты, имеющие одну и ту же размерность (число членов). При формализации содержательных задач предполагается, что последовательности, подлежащие обработке, искажены аддитивной некоррелированной гауссовской помехой. Для решения задач используется принцип максимального правдоподобия, который редуцируется к специфическим задачам комбинаторной оптимизации. К идентичным экстремальным задачам приводит формализация содержательных задач с опорой на критерий минимума суммы квадратов уклонений.

В настоящее время система QPSLab позволяет решать несколько десятков типовых задач, которые можно условно разбить на следующие классы: 1) обнаружение в числовой последовательности повторяющегося фрагмента; 2) распознавние последовательности, включающей повторяющийся фрагмент; 3) обнаружение и идентификация фрагментов; 4) распознавние последовательности, включающей фрагменты из алфа-

вита; 5) обнаружение фрагментов в последовательности и разбиение этой последовательности на серии идентичных фрагментов; 6) распознавние последовательности, включающей серии идентичных фрагментов; 7) обнаружение повторяющегося набора фрагментов; 8) распознавание последовательности, включающей повторяющийся набор фрагментов; 9) обнаружение и идентификация наборов фрагментов; 10) кластеризация последовательностей.

В перечисленных классах имеются как полиномиально разрешимые, так и NP-трудные задачи, для большинства из которых какие-либо эффективные алгоритмы с гарантированными оценками точности в настоящее время неизвестны. Пополнение системы QPSLab будет осуществляться за счет алгоритмов решения этих задач. В ближайшее время демонстрационная версия системы будет доступна через Интернет.

Работа поддержана РФФИ, проекты № 06-01-00058 и № 07-07-00022.

Литература

[1] *Кельманов А. В.* Апостериорный подход к решению типовых задач анализа и распознавания числовых квазипериодических последовательностей: обзор результатов // ММРО-12 — Москва: МаксПресс, 2005. — С. 125–128.

[2] *Кельманов А. В.* Проблемы оптимизации в типовых задачах помехоустойчивой апостериорной обработки числовых последовательностей с квазипериодической структурой // Докл. 3-й Всеросс. конф. «Проблемы оптимизации и экономические приложения». — Омск: ОмГТУ, 2006. — С. 37–41.

[3] *Кельманов А. В.* Полиномиально разрешимые и NP-трудные варианты задачи оптимального обнаружения в числовой последовательности повторяющегося фрагмента // Докл. Всеросс. конф. «Дискретная оптимизация и исследование операций». — Владивосток–Новосибирск: Ин-т математики СО РАН, 2007. — `http://math.nsc.ru/conference/door07/`.

[4] *Кельманов А. В.* О некоторых полиномиально разрешимых и NP-трудных задачах анализа и распознавания последовательностей с квазипериодической структурой // ММРО-13 (в наст. сборнике). — 2007. — С. 261–264.

Информационная технология количественной оценки состояния объектов природно-техногенной сферы по многоспектральным космическим изображениям

Кондранин Т. В., Козодеров В. В., Егоров В. Д.

`vkozod@mes.msu.ru`

(Москва)

Модернизация существующих и разработка новых технологий количественной оценки состояния объектов природно-техногенной сферы (совокупность лесных, водных и других экосистем, объединенных на кон-

кретном региональном уровне с объектами промышленного и сельскохозяйственного производства, дорожно-транспортной инфраструктуры и т.п.) являются важнейшей составной частью информационного обеспечения научных и прикладных задач космического мониторинга Земли. Научной основой решения таких задач является развитие математических моделей описания процессов формирования регистрируемого аппаратурой космического дистанционного зондирования (ДЗ) уходящего излучения и моделей распознавания образов наблюдаемых объектов по их многоспектральным изображениям.

Обоснование предлагаемой технологии на примере распознавания образов почвенно-растительных объектов и количественной оценки параметров состояния этих объектов изложены в работе [1]. Отличие технологии от традиционных подходов состоит в том, что с целью увязки используемых многоспектральных данных ДЗ и результатов моделирования взаимодействия излучения с объектами природно-техногенной сферы в терминах конкретных количественных характеристик, используются данные абсолютно калиброванных спутниковых систем. При этом при решении *прямой задачи* формируется исходная база спектральных образов указанных объектов при различных условиях освещенности, углах визирования аппаратуры ДЗ и разных условиях замутненности атмосферы. Для почвенно-растительного покрова возникает необходимость включения в соответствующую расчетную схему особенностей взаимодействия падающего солнечного излучения с отдельными фитоэлементами (листья/хвоя, ветки и др.) такой сложной системы с учетом их спектральной отражательной способности и условий их затенения при освещении Солнцем.

При постановке и решении *обратной задачи* в качестве исходной посылки используется то обстоятельство, что у специалистов-биологов, лесников и пр. существуют адекватные технологии определения объема фитомассы пробных площадок (в поле, в лесу и т.п.). Кроме того, для выбранных тестовых участков соответствующих наземных обследований существуют эмпирические связи между значениями зеленой фитомассы и общей биомассы древесины. Поэтому в предлагаемой технологии производится восстановление непосредственно именно этих количественных параметров растительности. Физико-математическое описание условий формирования спектральных образов наблюдаемых объектов и алгоритмы восстановления конкретных характеристик сводятся к процедуре обращения основного функционала многоспектральных данных ДЗ.

Указанный функционал оказывается зависящим от большого числа параметров, таких как типы лесной и межкроновой растительности, условия затенения фитоэлементов при их освещении прямым солнечным

и диффузным рассеянным излучением, оптическая толщина атмосферы и др., в том числе объем фитомассы. Исследуемый функционал представляет собой интеграл свертки суммарного падающего излучения с весовой функцией чувствительности аппаратуры ДЗ в пределах телесного угла визирования объекта при заданных зенитном угле визирования объекта и разности азимутальных углов визирования и Солнца с учетом спектральных отражательных способностей отдельных фитоэлементов.

Результаты решения *прямой задачи* для лесных экосистем можно образно представить как «книгу», каждый лист которой для разных типов лесных экосистем описывается координатами «плотность полога леса — ажурность крон деревьев». Плотность полога определяется взаимным расположением древостоев для выбранных классов породного состава леса (лиственные, хвойные, смешанные). Ажурность крон характеризует различия в распределении фитоэлементов на отдельных деревьях. В этих же координатах отображаются и значения объема фитомассы классифицируемых типов растительности. Для каждого «листа» такой «книги» рассчитываются спектральные интенсивности излучения, регистрируемого аппаратурой ДЗ. Решения обратной задачи находится путем поиска взаимных пересечений кривых спектральных интенсивностей: для каждого пиксела многоспектрального изображения производится поиск «минимального расстояния» между точками пересечения кривых интенсивностей и ближайшей изолинией величины фитомассы в указанной координатной системе.

Разработанная технология в настоящее время наряду с классом «растительность» обеспечивает выделение открытых водных поверхностей, облачности, а также объектов природно-техногенной сферы (городские территории и населенные пункты, дорожно-транспортная сеть, распаханные почвы и т.п.). Технология допускает также идентификацию не входящих в исходную базу данных модельных расчетов прямой задачи смешанных объектов, например, разных типов почвогрунтов. Однако, для элементов разрешения, относящихся к таким объектам, значительно возрастает время поиска решений обратной задачи, что делает неизбежным использование высокопроизводительных кластерных систем.

В работе демонстрируются примеры реализации предлагаемой технологии на примере объектов природно-техногенной сферы и лесных экосистем европейской территории России. В качестве исходных использовались данные аппаратуры ДЗ MODIS (Moderate-Resolution Imaging Spectroradiometer — видеоспектрорадиометр среднего разрешения) и ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer — усовершенствованный радиометр измерений из космоса теплового излучения и отражения) спутника Terra, а также аппаратуры ETM+

(Enhanced Thematic Mapper — усовершенствованный тематический картограф) спутника Landsat-7. При обработке использовались исходные абсолютно калиброванные данные указанной аппаратуры в 6–7 спектральных каналах видимой и ближней инфракрасной области спектра.

Исследования проводятся в рамках проектов РФФИ №05-05-64199 и №05-07-90176, а также проекта 4809 программы Федерального агентства по образованию «Развитие научного потенциала высшей школы (2006–2008 годы)».

Литература

[1] *Козодеров В. В., Кондранин Т. В., Косолапов В. С., Головко В. А., Дмитриев Е. В.* Восстановление объема фитомассы и других параметров состояния почвенно-растительного покрова по результатам обработки многоспектральных спутниковых изображений. — Исследование Земли из космоса, 2007. — №1. — С. 57–65.

Модели и методы построения и поддержки функционирования интеллектуальных адаптивных систем защиты информации

Котенко И. В.
ivkote@iias.spb.su
Санкт-Петербург, Институт информатики и автоматизации РАН

Используемым в настоящее время подходам к защите информации в распределенных компьютерных системах присущ целый ряд недостатков, и системы защиты информации (СЗИ) оказываются не в состоянии эффективно решать задачу управления защищенностью в режиме реального времени. Эти недостатки обусловлены, главным образом, узкой специализацией отдельных средств обеспечения безопасности, неразвитыми механизмами верификации защиты на этапах создания и поддержки, неадекватными механизмами определения уязвимостей, анализа рисков и определения уровня защищенности, мониторинга состояния сетей и адаптации к изменению условий функционирования [1]. В докладе предлагается подход к разработке и использованию СЗИ, основанный на использовании интеллектуальной надстройки над традиционными механизмами защиты и построении единой унифицированной среды для создания и поддержки функционирования систем защиты.

Интеллектуализация механизмов защиты

В соответствии с современными представлениями перспективная СЗИ распределенных компьютерных систем должна представлять собой взаимоувязанную, многоэшелонированную и непрерывно контролируемую

систему защиты используемых информационных, программных и аппаратных ресурсов, способную оперативно реагировать на удаленные и локальные компьютерные атаки и несанкционированные действия (НСД), накапливать знания о способах противодействия, обнаружения и реагирования на атаки и НСД и использовать их для усиления защиты.

Такая СЗИ должна предоставлять, по крайней мере, три уровня защиты. Первый уровень защиты составляют «традиционные» средства защиты, реализующие функции идентификации и аутентификации, криптографической защиты, разграничения доступа, контроля целостности, регистрации и учета, межсетевого экранирования. Второй уровень включает в себя средства проактивной защиты, обеспечивающие сбор необходимой информации, анализ защищенности, мониторинг состояния сети, обнаружение атак, противодействие их реализации, введение злоумышленника в заблуждение, и т. п. Третий уровень соответствует средствам управления защитой, которые осуществляют интегральную оценку состояния сети, управление защитой и адаптацию политик безопасности и компонентов СЗИ.

Первый уровень достаточно широко представлен в существующих исследованиях. Разработка механизмов защиты, относящихся ко второму и особенно третьему уровню, реализующих по существу интеллектуальную надстройку над традиционными механизмами защиты, составляет в настоящее время приоритетную задачу в области теоретических и прикладных исследований по построению информационно-безопасных распределенных вычислительных систем.

В рамках решения этой задачи в работе предлагается комплекс формальных методов, моделей, алгоритмов и построенных на их основе программных прототипов, реализующих следующие интеллектуальные механизмы защиты [1]:

1) сбор информации о состоянии информационной системы и ее анализ за счет механизмов обработки и слияния информации из различных источников;
2) проактивное предупреждение атак и препятствование их выполнению;
3) обнаружение аномальной активности и явных атак, а также нелегитимных действий и отклонений работы пользователей от политики безопасности, предсказание намерений и возможных действий нарушителей;
4) активное реагирование на попытки реализации действий нарушителей путем автоматической реконфигурации компонентов защиты для отражения действий нарушителей в реальном масштабе времени;

5) дезинформацию злоумышленника, сокрытие и камуфляж важных ресурсов и процессов, «заманивание» злоумышленника на ложные (обманные) компоненты с целью раскрытия и уточнения его целей, рефлексивное управление поведением злоумышленника;
6) мониторинг функционирования сети и контроль корректности текущей политики безопасности и конфигурации сети;
7) поддержку принятия решений по управлению политиками безопасности, в том числе по адаптации к последующим вторжениям и усилению критических механизмов защиты.

Поддержка жизненного цикла систем защиты

В процессе использования различных механизмов защиты необходимо осуществлять поддержку защищенной информационной среды на различных этапах жизненного цикла, включая этапы их проектирования, конфигурирования, развертывания, функционирования и модификации. Поэтому, кроме создания отдельных перспективных механизмов защиты, необходимо решать задачу разработки моделей и методов построения единой унифицированной системы (среды), осуществляющей поддержку всего жизненного цикла СЗИ, включая адаптивное управление политиками безопасности [1].

В работе предлагается подход к осуществлению непрерывной цепочки различных этапов жизненного цикла распределенных защищенных компьютерных систем (с множеством прямых и обратных связей от одного этапа к другому). Данный подход предполагает реализацию следующих механизмов:

1) спецификацию политик безопасности и архитектуры (или конфигурации) защищаемой системы;
2) трансформацию политик безопасности с целью их уточнения (детализации) с учетом описания защищаемой системы;
3) верификацию политик безопасности (проверку правильности и устранение конфликтов);
4) определение уровня безопасности и анализ рисков;
5) моделирование поведения системы защиты в различных условиях функционирования;
6) изменение политик в соответствии с требуемым уровнем безопасности и возможностями по использованию различных ресурсов и выделению финансовых средств и на защиту информации;
7) реализацию политик безопасности в системе, в том числе трансляции сформированных правил безопасности в параметры конфигурации и настройки программно-аппаратного обеспечения;

8) проактивный мониторинг выполнения политик безопасности, в том числе обнаружение отклонений работы пользователей от политики безопасности, обнаружение вторжений и анализ уязвимостей;
9) адаптацию поведения распределенных защищенных компьютерных систем и реализованных политик безопасности в соответствии с условиями функционирования.

В докладе приводятся примеры задач классификации, прогнозирования и анализа данных, используемые в предлагаемых механизмах защиты (в частности, при сборе информации о состоянии и ее анализе, обнаружении аномальной активности, дезинформации злоумышленника и мониторинге сети) и механизмах поддержки различных этапов жизненного цикла системы защиты, в том числе при адаптации ее поведения.

Заключение

В работе предложен подход к разработке и использованию интеллектуальных адаптивных систем защиты информации распределенных компьютерных систем. Подход основан на реализации интеллектуальных механизмов управления защитой и построении единой унифицированной среды для создания и поддержки функционирования систем защиты на всем их жизненном цикле, включая адаптивное управление политиками безопасности.

Работа выполняется при поддержке РФФИ (проект № 07-01-00547), программы фундаментальных исследований ОИТВС РАН (контракт № 3.2/03) и Фонда содействия отечественной науке.

Литература

[1] *Котенко И. В., Юсупов Р. М.* Технологии компьютерной безопасности // Вестник РАН. — 2007. — Т. 77. — № 4.

Исследование методов извлечения информации из текстов с использованием автоматического обучения и реализация исследовательского прототипа системы извлечения информации

Куршев Е. П., Кормалев Д. А., Сулейманова Е. А., Трофимов И. В.
epk@epk.botik.ru, dk@conrad.botik.ru, yes@helen.botik.ru, igor@warlock-98.botik.ru
Переславль-Залесский, ИПС РАН

Задача извлечения информации из текста [8] заключается в автоматической обработке набора документов с целью выделения релевантных

данных и представления их в структурированной форме. По глубине анализа текста и степени перехода от текста к модели предметной области технология извлечения информации (ТИИ) занимает промежуточное место между информационным поиском и гипотетической технологией понимания текстов. Извлечение информации может осуществляться в «слабом» или «сильном» смысле. Круг задач, решаемых системами первого типа, относится к так называемым «малым» задачам семантического анализа текстов [5], так как для их решения достаточен локальный контекст и ограниченный, локальный, синтаксический анализ. Результаты извлечения информации в «слабом» смысле и характер их представления несколько ограничивают возможности дальнейшего использования добытых из текста данных. Извлечением информации в «сильном» смысле мы назвали бы переход от базы текстовых фактов к такому их представлению, которое можно было бы использовать как интеллектуальный информационный ресурс, своего рода базу текстовых знаний.

Наши исследования были направлены на усовершенствование методов и расширение возможностей ТИИ, что позволило бы подойти вплотную к решению задачи извлечения информации в «сильном» смысле.

Средства технологии извлечения информации

Для выражения знаний о предметной области в задачах извлечения информации используется два основных вида средств: правила, описывающие текстовые ситуации (контексты), и ресурсы знаний. ТИИ обычно использует модель текста, основанную на аннотациях [9], отличающуюся простотой и высокой степенью универсальности. Описание текстовых ситуаций выполняется при помощи правил на специальном языке. При использовании модели аннотаций используются языки, основанные на языке CPSL [7]. Правило CPSL состоит из двух частей: образец и действия, выполняемые при успешном сопоставлении образцу. Образец для сопоставления — обобщение регулярного выражения, где в роли символов выступают аннотации.

Развитие языка правил. Нами был разработан расширенный диалект языка CPSL. Предлагаемые нами расширения [3] преследуют две цели: 1) обеспечить возможность описания более сложных контекстов и 2) снизить объем рутинной работы при создании системы правил за счет более компактного описания контекста. В число предложенных расширений языка правил входят развитая система типов данных для значений атрибутов аннотаций, логические метасимволы, метасимволы перехода, списки значений («микрословари»), опережающая и ретроспективная проверка, дополнительные квантификаторы [3].

Усовершенствование ресурсов знаний. Интерес представляют не столько системы собственно извлечения информации, сколько системы, обеспечивающие возможности аналитической обработки накопленной информации. Для достижения такого результата недостаточно лишь таксономии понятий предметной области, обогащенной атрибутивными или меронимическими связями (т. е. тезауруса), — требуется набор ресурсов знаний, всесторонне описывающих предметную область. В ходе изучения подходов к построению ресурсов знаний сформировалась многомерная классификация ресурсов по их характеру и назначению: «предметные–лингвистические», «априорные–полученные из текстов», «описание концептов–описание экземпляров» [6].

Применение машинного обучения

Трудоемкость построения контекстных правил вручную высока, поэтому для облегчения разработки и настройки приложений желательно использовать средства автоматизированного создания правил. Желательно также было бы в автоматическом или автоматизированном режиме пополнять ресурсы знаний.

Автоматическое построение правил. По своей сути задача построения набора правил близка к задаче наполнения высокоточных специализированных словарей моделей управления. Получаемые правила должны обеспечить извлечение информации в «слабом» смысле: распознавание текстовой ситуации, выделение целевых объектов и их свойств, а также, возможно, идентификацию некоторых отношений. Мы выбрали индуктивный подход к построению правил. Обучение идет на основе размеченных примеров, при этом используется две основных операции: обобщение и специализация [2]. Для повышения точности правил и обеспечения достаточной степени обобщения используется двухфазный сценарий обучения [4]. Предварительные эксперименты показали перспективность такого подхода.

Автоматизированное пополнение ресурсов знаний. Вопрос интеграции произвольных извлеченных текстовых фактов в ресурсы знаний пока открыт. Более насущные задачи ставит перед нами необходимость оперативного «запоминания» и использования извлеченной из некоторого текста информации для анализа этого же текста. Минимальная интеллектуальность системы предполагает, что текстовое выражение однажды распознанного объекта в дальнейшем будет рапознаваться и без «диагностического» контекста. По мере анализа текста идет пополнение динамического словаря текста [1], в дальнейшем эта информация подтверждается или опровергается экспертом. Для полноценного использования динамического словаря нужно решить ряд проблем: об-

работка неизвестных морфологическому анализатору склоняемых слов, учет синтаксической модели, описывающей словоизменение распознанной конструкции.

Работа выполнена при поддержке РФФИ, проект № 05-01-00442а.

Литература

[1] *Александровский Д. А., Кормалев Д. А., Кормалева М. С., Куршев Е. П., Сулейманова Е. А., Трофимов И. В.* Развитие средств аналитической обработки текста в системе ИСИДА-Т // КИИ-2006. Труды конференции. — Т. 2. — М.: Физматлит, 2006. — С. 555–563.

[2] *Кормалев Д. А.* Автоматическое построение правил извлечения информации из текста // 1-я межд. конф. «Системный анализ и информационные технологии» САИТ–2005. — Т. 1. — М.: КомКнига, 2005. — С. 205–209.

[3] *Кормалев Д. А., Куршев Е. П.* Развитие языка правил извлечения информации в системе ИСИДА-Т // Межд. конф. «Программные системы: теория и приложения», ИПС РАН, Переславль-Залесский, октябрь 2006 г. — Т. 1. — М.: Физматлит, 2006. — С. 365–377.

[4] *Кормалев Д. А.* Обобщение и специализация при построении правил извлечения информации // Конф. КИИ-2006. — Т. 2.— М.: Физматлит, 2006. — С. 572–579.

[5] *Леонтьева Н. Н., Семенова С. Ю.* Инструменты построения фрейма «ПЕРСОНА» // НТИ, Сер. 2. Информ. процессы и системы. — 2001. — № 8.

[6] *Сулейманова Е. А.* Классификация ресурсов знаний в системе извлечения информации из текста // ММРО-13 (наст. сб.). — 2007. — С. 625–628.

[7] *Appelt D. E.* The Common Pattern Specification Language: Technical report. — SRI International, Artificial Intelligence Center, 1996.

[8] *Appelt D. E., Israel D. J.* Introduction to Information Extraction. Tutorial // 16th Int'l. Joint Conf. on Artificial Intelligence IJCAI'99, Sweden, 1999.

[9] *Grishman R.* TIPSTER Text Architecture Design. Version 3.1. — New York: NYU, 1998.

Модель эволюции нуклеотидной последовательности

Любецкий В. А., Жижина Е. А., Горбунов К. Ю., Селиверстов А. В.

lyubetsk@iitp.ru

Москва, Институт проблем передачи информации РАН

Задача эволюции предковой последовательности вдоль данного эволюционного дерева G широко изучается и относится к числу важнейших проблем биоинформатики. Эволюционирующая последовательность может быть, например, нуклеотидной, т. е. в алфавите из четырех букв $\{A, C, T, G\}$. Предполагается, что эти буквы заменяются друг на друга,

и, кроме того, возможны еще два события: в случайное место последовательности *вставляется* некоторый участок в том же алфавите или *удаляется* такой участок. Таким образом, длина n эволюционирующей последовательности переменная.

В дереве G каждому j-му ребру приписана его длина t_j, которая интерпретируется как время эволюции вдоль этого ребра. Каждой позиции i от 1 до n сопоставляется скорость эволюции r_i в позиции i. Значение r_i обычно находится из гамма-распределения с последующим усреднением результата моделирования. Из биологии известны несколько вариантов конкретной матрицы R, которая управляет заменой одних букв на другие. Тогда эволюция последовательности букв задается простым правилом: буква в любой i-й позиции последовательности σ_j, сопоставленной началу j-го ребра, преобразуется в букву в той же позиции последовательности σ'_j в конце j-го ребра вероятностно как $\exp(R \cdot r_i t_j)$, где R — известная матрица соответствующего размера. Сама эволюционирующая последовательность называется еще *первичной структурой*, и в ней образуется так называемая *вторичная структура* — множество спиралей, где каждая спираль — это некоторое спаривание нуклеотидов по правилу G с C и T с A, см. [1]–[4].

Фундаментальная проблема биоинформатики состоит в моделировании эволюции вдоль данного дерева последовательности вместе с вторичной структурой в ней. Ниже предложена модель для описания такой эволюции. Как приложение этой модели, мы рассматриваем задачу построения *множественного выравнивания последовательностей с учетом их вторичной структуры*. А именно, в обозначениях, которые будут даны ниже, вторичная структура в концевых последовательностях $\{\sigma_m\}$, индуцированная минимальной конфигурацией σ^*, позволяет для исходных данных $\{\theta_m\}$, в которых вторичная структура заранее не была определена, получить множественное выравнивание с учетом этой индуцированной эволюционным процессом вторичной структуры. Соответствующий алгоритм будет изложен.

Описание модели эволюции последовательности вместе с ее вторичной структурой

Дано дерево G эволюции нуклеотидной последовательности с заданными длинами ребер, и концевым вершинам дерева приписаны *современные последовательности* $\{\theta_m\}$. *Конфигурацией* называется некоторое произвольное сопоставление всем вершинам, включая концевые, последовательностей переменной длины n. Последовательности из данной конфигурации, которые сопоставляются концевым вершинам дерева, назовем *концевыми последовательностями*, и обозначим их $\{\sigma_m\}$. Ниже предлагается функционал $H(\sigma)$, аргументом которого является конфи-

гурация σ, и минимум которого при значении аргумента σ^* соответствует эволюции предковой последовательности вдоль дерева вместе с вторичной структурой в ней.

Функционал выражает три условия на искомую конфигурацию σ:

1) для каждой последовательности σ_j и в каждой позиции $i = 1, \ldots, n$ любой последовательности из конфигурации происходит независимая замена букв вдоль любого ребра j согласно матрице замен R, как указано выше, и также вставка/удаление участков — слагаемое $H_1(\sigma)$ ниже;
2) значения концевых последовательностей конфигурации близки к соответствующим современным последовательностям — слагаемое $H_2(\sigma)$;
3) последовательности из конфигурации по возможности сохраняют вторичную структуру от начала ребра к его концу и вдоль целого пути в дереве, т. е. в течение многих поколений; при этом функционал меньше, если такие пути длиннее и их больше — слагаемое $H_3(\sigma)$.

Уточним, что *путем* φ называется путь в дереве G, вдоль ребер которого сохраняется высокая близость вторичных структур (как везде «близость» понимается в смысле некоторого фиксированного порога). *Длиной* $l(\varphi)$ такого пути φ назовем число ребер в нем. *Временем* $t(\varphi)$ пути φ назовем сумму длин t_j ребер вдоль φ. Далее φ везде *обозначает* путь в этом смысле.

Поскольку последовательности σ_j и σ'_j, приписанные соответственно началу и концу ребра j, имеют, вообще говоря, разные длины, то дальше предполагается, что для каждой конфигурации σ и каждого ее ребра j выполнено стандартное выравнивание последовательностей σ_j и σ'_j. Выровненные последовательности, которые теперь включают знак пробела, будем обозначать соответственно δ_j и δ'_j.

Нами предложен такой функционал $H(\sigma)$ и стохастический алгоритм для поиска его глобального минимума, основанный на идее аннилинга. А именно, положим $H(\sigma) = (H_1(\sigma) + H_3(\sigma)) + H_2(\sigma)$, где

$$H_1(\sigma) = \sum_j - \ln P(\sigma_j, \sigma'_j, t_j),$$

j пробегает все ребра дерева G; t_j — время, приписанное ребру j;

$$P(\sigma_j, \sigma'_j, t_j) = \prod_i \exp(R \cdot r_i t_j)(\sigma_j, \sigma'_j) \cdot \prod_k \exp(-\varepsilon \ln(l_k + 1)),$$

где i пробегает все столбцы, в которых буква соответствует при выравнивании букве, k пробегает все максимальной длины участки в этом вы-

равнивании типа, содержащие с одной из сторон только пробелы, и тогда l_k — длина такого участка, ε — параметр, отвечающий за значимость пробелов по сравнению с заменами букв. Далее

$$H_2(\sigma,\theta) = -\lambda \sum_{\sigma_m} \delta(\sigma_m,\theta),$$

где λ — параметр, отвечающий за соотношение двух слагаемых $(H_1(\sigma) + +H_3(\sigma))$ и $H_2(\sigma)$; σ_m пробегает все концевые последовательности конфигурации σ, а θ — современные последовательности; функция δ начисляет некоторый штраф за расхождение по каждой позиции у последовательностей σ_m и θ. Заметим, что вторичная структура в θ не предполагается известной. Наиболее трудным является вопрос о том, как правильно записать слагаемое $H_3(\sigma)$. Представим его в виде суммы

$$H_3(\sigma) = -\Big(H_0(\sigma) + k\sum_j U\big(Q(\sigma_j,\sigma'_j),t_j\big)\Big).$$

Здесь первое слагаемое $H_0(\sigma)$ — сумма по всем ребрам j сумм энергий всех спиралей в σ_j со значениями ниже некоторого порога. Это слагаемое отражает тот факт, что в искомой конфигурации σ, описывающей эволюцию, многие σ_j предполагаются со спиралями, имеющими низкую энергию. Второе слагаемое отражает сохранение вторичной структуры и, как следствие, длину и количество путей φ, о которых говорилось выше. Предполагается, что минимизация функционала с указанным выше $H_3(\sigma)$ влечет минимизацию того же функционала с более сложным слагаемым

$$H'_3(\sigma) = -\Big(H_0(\sigma) + k\sum_\varphi l(\varphi)\sum_{j\in\varphi} U\big(Q(\sigma_j,\sigma'_j),t_j\big)\Big).$$

Сохранение вторичной структуры при переходе от σ_j к σ'_j, т. е. величина Q, описывается в [5]. Принимается $U\big(Q(\sigma,\sigma'),t\big) = Q - \ln(1+t/\mu)$. В качестве алгоритма поиска глобального минимума была принята схема аннилинга для алгоритма Метрополиса.

Обоснование алгоритма

В качестве алгоритма поиска глобального минимума $\sigma*$ указанного функционала $H(\sigma)$ мы рассматриваем схему аннилинга на основе алгоритма Метрополиса-Хастингса, которая представляет собоймарковскую цепь на пространстве всех конфигураций модели. Марковская цепь имеет стационарную гиббсовскую меру

$$P_m(\sigma) = \frac{\exp(-\beta_m H(\sigma))}{\sum\limits_\sigma \exp(-\beta_m H(\sigma))}$$

при каждом фиксированном β_m.

Теорема 1. *Если для параметра β_m выполняется условие $\lim\limits_{m\to\infty} \frac{\log m}{\beta_m} > \text{const}$, то описанный выше алгоритм обладает следующим свойством: $\lim\limits_{m\to\infty} \mathsf{P}\{\sigma(m) \in E_{\min}\} = 1$, где $E_{\min}$ — множество глобальных минимумов нашего функционала, $\sigma(m)$ — конфигурация, возникшая к m-й итерации, $\mathsf{P}\{\cdot\}$ — распределение этой марковской цепи в момент времени m.*

Работа поддержана РФФИ, проект № 07-01-00445, и МНТЦ 2766.

Литература

[1] *Seliverstov A. V., Lyubetsky V. A.* Translation regulation of intron containing genes in chloroplasts // Journal of Bioinformatics and Computational Biology. — 2006. — V. 4, № 4. — P. 783–793.

[2] *Lyubetsky V., Pirogov S., Rubanov L., Seliverstov A.* Modeling classic attenuation regulation of gene expression in bacteria // Journal of Bioinformatics and Computational Biology. — 2007. — V. 5, № 1. — P. 155–180.

[3] *Vitreschak A. G., Mironov A. A., Lyubetsky V. A., Gelfand M. S.* Functional and evolutionary analysis of the T-box regulon in bacteria // Genome Biology. — 2007. — in print.

[4] *Горбунов К. Ю., Любецкий В. А.* Эволюция предковых регуляторных сигналов вдоль дерева эволюции фактора транскрипции // Молекулярная биология. — 2007. — Т. 41, в печати.

[5] *Горбунов К. Ю., Миронов А. А., Любецкий В. А.* Поиск консервативных вторичных структур РНК // Молекулярная биология. — 2003. — Т. 37, № 5. — С. 850–860.

Средства OLAP-моделирования и их применение в задачах здравоохранения

Ноженкова Л. Ф.

expert@icm.krasn.ru

Красноярск, Институт вычислительного моделирования СО РАН

Технология оперативной аналитической обработки многомерных данных OLAP (On-line Analytical Processing) считается одним из разделов интеллектуального анализа данных. Аналитические OLAP-модули все чаще появляются в составе отечественных и зарубежных продуктов и финансово-производственных приложений. Существо аналитической обработки сводится к автоматизированной поддержке формирования аналитических запросов, агрегированию данных, операциям над многомерным кубом данных с использованием плоских представлений — кросс-таблиц, кросс-диаграмм, картограмм. Наибольшее применение технология OLAP

получила в бизнес-среде, где, как правило, решение конкретной аналитической задачи укладывается в рамки одного многомерного куба. При этом классические OLAP-решения мало пригодны к использованию в прикладных областях, где необходим комплексный анализ данных, связанный с реализацией сложных аналитических алгоритмов. Примерами прикладных областей, в которых указанные проблемы не позволяют эффективно применять традиционные средства OLAP-технологии, являются здравоохранение, образование, социальная защита населения и множество других. Методы расчета аналитических показателей и решения задач планирования в этих прикладных областях представляют собой сложные многошаговые процессы анализа многомерных данных. Возникает необходимость поэтапной обработки данных. Например, задачи планирования медицинской помощи представляют собой сложные многошаговые процессы с большим количеством расчетов, многообразием входной и выходной информации, сложными внутренними взаимосвязями.

Коллективом отдела прикладной информатики Института вычислительного моделирования СО РАН предложен новый подход к решению разнообразных задач с применением OLAP-технологии, основанный на построении комплексов так называемых «OLAP-моделей» [1]. OLAP-модель строится пользователем и несет в себе описательную информацию о решении некоторой аналитической задачи. Структурно OLAP-модель состоит из информации, описывающей исходные данные и их взаимосвязи, измерения и показатели информационного куба, операции над кубом, способы представления результатов вычисления и способы сохранения результатов для последующего использования. Введение в модель операций сохранения результатов расчета в источник (хранилище) данных позволило реализовать поэтапный анализ данных путем создания комплексов OLAP-моделей.

Комплекс представляет собой совокупность OLAP-моделей, связанных по данным. В рамках одного расчета модели образуют последовательно выполняемую цепочку операций, при этом данные, рассчитанные одной моделью, в дальнейшем используются другими моделями.

Ядром системы оперативной аналитической обработки данных является OLAP-машина, которая представляет собой механизм выполнения запросов пользователя на выборку многомерной информации и изменения ее представления. От архитектуры OLAP-машины зависит и уровень решаемых аналитическим инструментом задач, и степень свободы пользователя при решении этих задач. Предложена и реализована оригинальная архитектура OLAP-машины, отличительными особенностями которой являются выполнение нерегламентированных запросов пользо-

вателя, использование встроенного языка программирования для расчета значений аналитических объектов и возможность использования составных иерархий в качестве измерений. Составная иерархия позволяет упорядочить информационные объекты одновременно по нескольким измерениям. Использование иерархий в качестве измерений многомерного куба потребовало применения специальных структур данных и алгоритмов, направленных в первую очередь на уменьшение временных затрат. Разработаны средства автоматизации создания специализированных OLAP-приложений: инструментальное ядро в виде набора компонент, связанных с OLAP-машиной, среда проектирования экранных форм пользовательского интерфейса, мастер быстрого создания приложений.

Расширен функциональный состав хранилищ данных [2]. Введены новые конструктивные элементы, выполняющие функции поддержки связных многошаговых аналитических расчетов: OLAP-модель, сложное иерархическое измерение, таблица расчетных значений (агрегатов) и группа отчётных форм. Применение оригинальных технологических компонентов позволило выполнять сложные многошаговые аналитические расчёты. На первом шаге на вход аналитического инструмента поступают исходные обрабатываемые данные. Далее идёт последовательное выполнение шагов расчёта, параметры которых описаны в репозитарии хранилища в виде OLAP-моделей. Взаимодействие моделей между собой происходит путем передачи через хранилище информации в виде таблиц агрегатов и данных репозитария. Выполнение многошагового расчета сопровождается так называемым интерактивным аналитическим экспериментом — возможно вмешательство пользователя в выполнение расчёта для модификации параметров и настройки модели. Процесс формирования каждой из аналитических моделей сопровождается взаимодействием пользователя со средствами управления хранилищем. На любом из этапов построения модели возможен возврат к более ранним этапам.

Разработаны средства создания OLAP-приложений с адаптированным для специалистов предметной области интерфейсом. Инструментарий разработки адаптированных интерфейсов позволяет создавать ориентированные на конкретную задачу OLAP-приложения на основе инструментального ядра, полностью ограждая пользователя от сложной внутренней организации системы и сохраняя при этом весь функционал. Важным требованием к разрабатываемым приложениям является отражение специфики и традиций конкретной предметной области, в том числе, создание и применение словаря терминов соответствующей прикладной области.

Разработанные средства организации хранения и оперативной аналитической обработки данных послужили основой для создания и использования централизованного хранилища медицинских данных регионального уровня [3, 4, 5]. Фактологические данные в хранилище стекаются из множества информационных систем, работающих в учреждениях здравоохранения и обязательного медицинского страхования. Состав информации в хранилище данных направлен на анализ эффективности деятельности медицинских учреждений в системе здравоохранения региона и оказывает существенную помощь в решении задач оперативного управления и планирования. Медицинская информация разнородна и включает статистические (обобщенные), персонифицированные, пространственно-распределенные, и другие типы данных. Для увеличения наглядности представления результатов оперативного аналитического моделирования разработаны средства отображения результатов OLAP-анализа на электронной карте. Полученные результаты применены для информационно-аналитической поддержки задач национального проекта «Здоровье» по реструктуризации сети медицинских учреждений.

Работа выполнена при поддержке гранта Президента для ведущих научных школ № НШ-3428.2006.9, гранта РФФИ № 05-07-90244 и гранта по проекту № 7 Программы фундаментальных исследований Президиума РАН № 14.

Литература

[1] *Дудина Ю. В., Ишенин П. П., Ноженкова Л. Ф.* Технология реализации аналитических моделей средствами системы «Аналитик» для решения задач планирования / Труды Всероссийской конференции «Информационно-аналитические системы и технологии в здравоохранении и ОМС». — Красноярск: КМИАЦ, 2002. — С. 246–254.

[2] *Жучков Д. В.* Автоматизация обработки больших массивов данных // «Открытое образование». Приложение — Красноярск: ООО «Экспресс-Офсет», 2006. — С. 56–62.

[3] *Евсюков А. А., Ноженкова Л. Ф.* Оперативное геомоделирование сети медицинских учреждений // Вестник КрасГАУ. — 2006. — № 13. — С. 114–118.

[4] *Исаева О. С.* Построение информационных моделей для OLAP-анализа медико-демографических данных // Журнал «Открытое образование». Приложение. — Красноярск, 2006. — С. 65–73.

[5] *Ноженков А. И., Коробко А. В., Никитина М. И.* Информационно-аналитическая поддержка формирования территориальной программы бесплатной медицинской помощи // Вестник КрасГАУ, 2006. — № 13. — С. 108–113.

Создание системы распределенного отказоустойчивого хранения цветных крупноформатных изображений

Попов С. Б.
spop@smr.ru
Самара, Институт систем обработки изображений РАН

В настоящий момент растет интерес к системам параллельной или распределенной обработки изображений. В первую очередь это связано с тем, что появилась насущная потребность в обработке крупноформатных изображений. Наблюдается устойчивая тенденция к увеличению размеров формируемых изображений во многих областях деятельности. Однако увеличение размера изображения порождает большие проблемы при их обработке, хранении и передаче данных.

Обработка изображений с использованием параллельных или распределенных систем помогает преодолеть «проклятие размерности» в процессе вычислений, но для большинства таких систем обработки изображений узким местом являются предварительный этап рассылки данных исходных изображений по компьютерам распределенной вычислительной системы и завершающий этап сбора обработанных фрагментов в единое изображение. Попытки распараллелить или существенно сократить эти необходимые, но непроизводительные этапы распределенной обработки изображений приводят к идее распределенного изображения.

Распределенное изображение — это структура данных, определяющая способ и параметры разбиения изображения на фрагменты, список компьютеров, где находятся эти фрагменты, место их размещения и формат хранения. В данном случае фрагменты обрабатываемых изображений хранятся непосредственно на компьютерах, выполняющих параллельную обработку. Каждая задача параллельной программы обрабатывает тот фрагмент изображения, который расположен на компьютере, где выполняется данная задача, результат обработки сохраняется здесь же, как часть нового распределенного изображения, полученного в результате работы всех задач, участвующих в распределенной обработке.

Несмотря на очевидность данной идеи, она не получила распространения потому, что исследователи не смогли предложить удовлетворительных решений возникающих при этом проблем.

— Как выполнить декомпозицию (разбиение), близкую к оптимальной, для априори неизвестной последующей задачи обработки?
— Как решить проблемы сбалансированности загрузки компьютеров, участвующих в обработке при заранее выполненной декомпозиции данных?

— Как обеспечить достаточный уровень отказоустойчивости распределенного хранения фрагментов изображений?
— Как обеспечить виртуальную целостность распределенного изображения?
— Как обеспечить приемлемый уровень интерактивности системы при визуализации распределенных изображений?

В данной работе предлагается новый подход к организации хранения данных в виде распределенных изображений при параллельной обработке на многопроцессорных системах различной архитектуры.

Анализируя варианты декомпозиции изображений при параллельном выполнении различных операций обработки, следует отметить, что наиболее целесообразным для распределенного изображения представляется декомпозиция в виде перекрывающихся фрагментов. Особенность предлагаемого подхода заключается в том, что размер необходимого перекрытия фрагментов определяется не параметрами последующей задачи обработки, поскольку она априори неизвестна, а необходимостью решения проблем сбалансированности загрузки компьютеров и обеспечения достаточного уровня отказоустойчивости распределенного хранения фрагментов изображений.

Распределенное изображение определяется в виде набора *перекрывающихся* фрагментов изображения. Для каждого из M компьютеров фрагмент формируется следующим образом. Все строки изображения делятся на $2M$ блоков одинакового размера. Фрагмент распределенного изображения на m-м компьютере содержит два основных блока с номерами $2m-1$ и $2m$, а также два так называемых теневых блока с номерами $2m-2$ и $2m+1$, которые являются основными для двух соседних узлов, соответственно для $(m-1)$-го и $(m+1)$-го. Два основных блока соответствуют варианту декомпозиции изображения на непересекающиеся фрагменты, причем младший основной блок хранится в качестве одного из теневых блоков на компьютере с меньшим номером, а старший — на компьютере с большим номером. Таким образом, изображение разбивается на непересекающиеся фрагменты, а затем к фрагменту добавляются с каждой стороны по половине примыкающего фрагмента, хранящегося на соседнем компьютере. В распределенном изображении за каждым узлом хранения закреплен определенный фрагмент изображения (основные данные), части соседних фрагментов (теневые данные) хранятся здесь дополнительно. Именно данные основных блоков формируются на узле в процессе распределенной обработки изображения. По окончании процесса обработки выполняется обмен теневыми данными.

Такое заведомо избыточное разбиение решает проблему отказоустойчивости, то есть позволяет восстановить изображение при отказе одного из узлов хранения.

Одновременно с этим, предложенный принцип декомпозиции распределенного изображения позволяет реализовать оригинальный алгоритм динамического распределения нагрузки при выполнении операций поэлементной обработки или локальной обработки скользящим окном. Суть его заключается в следующем. Каждый вычислительный узел, участвующий в параллельной обработке, начинает вычисления с первой строки старшего основного блока, затем формируется последняя строка младшего блока. Далее попеременно формируются строки старшего блока в порядке возрастания номера и строки младшего блока в порядке убывания. Когда общее количество сформированных на узле строк достигнет некоторого определенного значения, например половины количества строк в блоке основных данных, узлам, обрабатывающим соседние фрагменты, рассылаются сообщения о том, что текущий узел на четверть завершил свою работу. Если к этому времени соответствующих сообщений от соседей не получено, то это означает, что текущий компьютер должен запланировать себе формирование и тех строк результирующего изображения, которые являются для него теневыми.

Рассматривая взаимодействие двух соседних вычислительных узлов, можно заметить, что они совместно формируют ту часть результирующего изображения, которая содержится между серединами их основных данных, причем каждый из них имеет всю необходимую информацию, чтобы сделать эту работу самостоятельно. В процессе работы смежные узлы двигаются навстречу друг другу, сообщая о скорости своего процесса вычислений при достижении заранее определенных моментов, например четверти всей работы, половины, трех четвертей и, наконец, полного завершения. При этом прогнозируемый номер строки изображения, на которой эти процессы встретятся, постоянно корректируется в зависимости от текущей загрузки вычислительных узлов. Таким образом, все процессы завершат свою работу практически одновременно. Разница при этом составит на больше, чем время обработки одной строки. В результате будет сформировано результирующее изображение в полном объеме, но размещение его данных по компьютерам, содержащим новое распределенное изображение, будет неравномерным. Однако пользователь может получить результат своего запроса сразу по завершении процесса обработки. Далее в фоновом режиме узлы хранения только что сформированного распределенного изображения обменяются своими данными для того, чтобы привести структуру распределенного изображения к необходимому виду.

При реализации функции визуализации крупноформатных распределенных изображений приемлемый уровень интерактивности системы обеспечивает наличие так называемого эскиза изображения на каждом или некоторых узлах распределенной системы хранения. В этом случае визуализирующее приложение, обращаясь к узлу с наиболее быстрым доступом, получает данные, необходимые для предварительного просмотра изображения. Количество узлов хранения, которые содержат эскизы, может выбираться исходя из конфигурации локальной сети рабочей группы пользователей, совместно использующих распределенную систему хранения, или допустимого уровня избыточной информации, которая может храниться в такой системе. При просмотре изображения в полноразмерном режиме пользователю отображается только выбранный им фрагмент. При прокрутке изображения в окне происходит подкачка необходимых данных. Для обеспечения комфортного уровня интерактивности подкачка выполняется в режиме чтения с упреждением (prepaging). Наличие перекрытий хранящихся фрагментов делает этот процесс более плавным.

Таким образом, предложенный подход к организации данных в распределенных изображениях снимает большинство из обозначенных выше проблем.

Работа выполнена при поддержке РФФИ, проект № 07-07-00210.

Использование текстового индекса при работе с документами в универсальной базе данных

Пржиялковский В. В.
prz@ccas.ru
Москва, Вычислительный центр РАН

Современные универсальные базы данных изначально создавались как «фактографические», т. е. допускающие хранение скалярных данных, организованных в таблицы. Хранение и обработка «сложно устроенных» данных традиционно выполнялась специализированными системами. Это порождало проблемы интеграции и синхронизации логически единых, но разнородных данных, вынуждено разнесённых по разным системам, а также препятствовало использованию таких возможностей универсальных баз данных, как хранение особо больших объёмов, восстановимость после потерь, поддержание целостности данных, защита. Однако, начиная с первой половины 90-х годов, эти проблемы стали постепенно разрешаться вследствие того, что основные производители универсальных СУБД начали добавлять в свои системы возможности работы со сложно устроенными данными.

Характерной в этом отношении является СУБД Oracle. Сейчас она дает возможность хранить в базе данных и обрабатывать (в разной степени) следующие категории нескалярных данных:

1. «Большие неструктурированные объекты» (LOB, large objects). БД относится к ним как к строкам байтов или текстовых символов с разрешённой длиною до 128 терабайт. СУБД обеспечивает только хранение (обеспечивается степень сжатия больше, чем в файловых системах), а обработка и интерпретация возложена на прикладные программы.
2. Структурированные объекты, построенные по принципам объектно-ориентированного подхода. СУБД полностью знает структуру хранимых объектов и использует её в запросах и в хранении.
3. Частично структурированные объекты. Объекты этого рода как правило не имеют жёсткой структуры и устроены по-разному соответственно разным предметным областям.

Для работы с объектами последней категории Oracle использует так называемые «предметные индексы» (domain indexes). Предоставляются средства программирования такого рода индексов, но одновременно имеется и несколько встроенных видов, готовых к употреблению. Например, пространственный индекс позволяет хранить в базе данных и обрабатывать с помощью СУБД пространственные данные; текстовый индекс, вместе со встроенной в СУБД поисковой текстовой машиной (Oracle Text), позволяет расширить возможности базы данных традиционным инструментарием информационно-поисковых систем (ИПС).

Замечательно, что запросы к базе данных могут быть комбинированными, например, в одном запросе может содержаться обращение как к скалярным таблично-организованным данным, так и к текстовым документам.

Возможности текстового индекса

В Oracle Text имеется три разновидности текстового индекса применительно к трем случаям текстовой обработки:

— `CTXSYS.CONTEXT` — для выполнения полнотекстового поиска по текстовым документам как внутреннего хранения, так и внешнего (файловая система, интернет);
— `CTXSYS.CTXCAT` — для выполнения упрощенного и ускоренного поиска в каталогах с краткими описаниями, например в лентах новостей;
— `CTXSYS.CTXRULE` — для построения классификаций, или рубрикаций, документов при том, что признаки классифицирования описываются набором характерных запросов.

Документы для индексирования могут находиться в базе данных, в интернете, в файлах или же в произвольном месте, обращение к которому можно оформить программой. Форматы документов могут быть самых разных видов, включая простой текст, PDF, RTF, MS Word, XML, HTML и прочие.

Наибольшие возможности имеет полнотекстовый индекс. Содержательно он хранит тройки ⟨документ, словоместо, индексируемое слово или несколько слов⟩, и позволяет по предъявленному поисковому слову получить список пар ⟨документ, словоместо⟩. В сочетании с имеющейся программной логикой он позволяет осуществлять поиск по документам со следующими свойствами:

— Точный поиск. В документе ищется в точности указанное слово, например `'Java'`.
— Позиционный поиск группы слов. Может выполняться поиск комбинации слов, например `'Java Development Kit'`. Может искаться близкое расположение слов, например слово `'Java'`, расположенное на расстоянии не более чем N слов от слова `'development'`. Может искаться встреча двух слов, например, `'Java'` и `'development'`, в одном предложении или же параграфе.
— Нечеткий поиск. Могут искаться слова, похожие по звучанию или похожие по написанию (последствия опечаток). Могут искаться слова одного корня. Могут искаться слова с указанными подстроками. Могут искаться слова с привлечением тезауруса, построенного в соответствие со стандартом ISO-2788.

Результаты поиска ранжируются по встроенному правилу, или по запрограммированному самостоятельно.

Продвинутые возможности Oracle Text позволяют отойти от поиска по словам и обеспечивают следующее:

— Тематический поиск. Поиск документов по темам, а не отдельным словам.
— Рубрицирование предъявленного документа по темам.
— Автоматическое формирование резюме документа.
— Автоматическое формирование набора классификационных правил на основе «обучения» предъявленными документами.
— Кластеризация. Группирование документов по близости содержания.

Некоторые возможности текстового индекса в Oracle Text не имеют готовых реализаций для русского языка. Например, это касается поиска по словам, близким по произношению, или же морфологического поиска. Однако имеется программный инструментарий для восполнения некоторых подобного рода пробелов.

Количественные характеристики текстового индекса

Практичность информационной системы часто напрямую зависит от её технических характеристик. Ниже приводятся характерные затраты ресурсов компьютера при работе Oracle Text, доступные для простой самостоятельной проверки.

В документации Oracle имеется два файла по Oracle Text в формате PDF: `b14217.pdf` и `b14218.pdf`. Их общий объем примерно 4,6 Мб или 742 страницы. Полученные характеристики индексирования: время построения индекса — 77 сек.; объем всех выше перечисленных структур индекса — 8 Мб; количество лексем в индексе — 40546; характерное время выполнения запроса сочетания `'oracle text'` по индексу — 1,28 сек.

Замеры приблизительные и соответствуют процессору Celeron с тактовой частотой 1 Ггц и оперативной памяти 512 Мб. Формат PDF — не самый экономный; индексирование документа формата HTML даст меньшее количество лексем, меньший объём индекса, более быстрое его построение и более быстрые ответы на запросы.

Более мощные вычислительные возможности способны дать лучшие характеристики. В отечественной практике есть случай использования Oracle Text для индексации и семантического анализа в базе данных электронных писем размером в несколько терабайт (без учёта вспомогательных структур).

Работа выполнена при поддержке РФФИ, проект №07-07-00181.

Литература

[1] Text Application Developer's Guide, 10g Release 2 (10.2). — Oracle Corp., Part Number B14217–01.

[2] Text Reference, 10g Release 2 (10.2). — Oracle Corp., Part Number B14218–01.

[3] *Плешко В. В.* Поиск с учетом словоформ русского языка. — Oracle Magazine, июнь/июль 2003. `www.oracle.com/ru/oramag/june2003/index.html?russia\_rco3.html`.

Распознавание подобных изображений в больших базах данных

Протасов В. И., Потапова З. Е., Сулейменов О. М., Сыроешкин Р. В., Челинцева Е. В.

protonus@yandex.ru

Московский государственный горный университет

В настоящей работе рассматриваются простые геометрические методы распознавания для двух видов объектов: контуров предметов и растровых изображений.

Каждому объекту методы ставят в соответствие некоторый двоичный дескриптор, однозначно описывающий этот объект. Методы построены таким образом, что если предлагаемый к распознаванию объект может быть получен из любого объекта базы данных с помощью поворота на произвольный угол, изменения масштаба и кадрирования (а в случае растрового изображения и изменения освещенностей в разумных пределах), то он должен найтись в базе данных по совпадению частей дескрипторов. Общая часть искомого изображения и изображения из базы данных, подобного ему, должна составлять при этом более половины изображения.

Метод распознавания объектов по контурам подробно описан в [1]. Там предложен метод нахождения характерных точек контура, которые могут быть получены при использовании единственного настроечного параметра. Вначале на контуре находятся две крайних характерных точки, являющихся концами главного, наибольшего диаметра контура, а следующие характерные точки находятся с использованием простого геометрического метода. Суть его заключается в следующем: между двумя характерными точками при обходе контура можно поставить еще одну, удовлетворяющую двум условиям:

— сумма расстояний от первой характерной точки до испытуемой, и от испытуемой до второй характерной точки должна быть максимальной;
— отношение высоты, опущенной из испытуемой точки на основание образованного тремя точками треугольника, к длине основания превышает некую наперед заданную величину (обычно 0.05).

Далее для произвольной характерной точки строится вспомогательный треугольник, образованный ею и двумя соседними характерными точками. Из треугольника берутся две относительные величины:

— отношение длины наименьшей стороны, прилегающей к данной характерной точке, к длине наибольшей стороны;
— отношение длины противолежащей стороны к сумме длин прилегающих.

В единичном квадрате характерной точке ставится в соответствие изображающая точка, стоящая на пересечении двух таким образом полученных относительных координат. Единичный квадрат разбивается на 32×32 клетки, и эта точка становится единицей в 1024-разрядном дескрипторе изображения. Число таких единиц, расположенных в разных частях единичного квадрата, может быть меньше числа характерных точек контура, поскольку в одну клетку единичного квадрата может попасть несколько характерных точек. На качестве распознавания этот эф-

фект не сказывается, поскольку он проявляется одинаковым образом как на искомом изображении, так и на подобном ему. Дескрипторы контура заносятся в базу данных. В дальнейшем, если контур подвергнут аффинным преобразованиям и кадрированию, то количество совпадающих единиц дескриптора элемента базы данных и искомого будет использовано для целей распознавания.

Простой геометрический метод распознавания был применен и во втором случае, когда рассматриваются растровые изображения.

Разработка метода совмещения изображений в рамках представленной работы осуществлялась на основе анализа задач исследований Земли из космоса, при решении которых необходимо совмещать разные изображения, и основных требований к решению задачи совмещения. Общая постановка проблемы была сформулирована следующим образом.

Имеется два снимка участка земной поверхности, снятых цифровой камерой в разное время с разных высот (отличающихся, по крайней мере, не более, чем в два раза), при различном освещении (отношение средних освещенностей двух снимков не должно отличаться более чем в два раза). Ось съёмки перпендикулярна поверхности в пределах ± 5 градусов, снимки могут быть повернуты относительно оси съёмки на произвольный угол. Требуется определить, какие участки поверхности, находящиеся в области пересечения обоих снимков, подверглись изменениям (фиксация появления и исчезновения объектов, отличающихся на двух снимках). Дополнительным требованием является существование возможности определять относительные высоты объектов.

Исходя из постановки задачи, более простым решением является поиск совпадающих участков рельефа и расположенных на нём объектов, с последующим вычленением их из обоих снимков для нахождения изменений.

Для определения идентичности двух точек левого и правого снимка вокруг них проводится ряд из N окружностей. В пределах каждого кольца (первое кольцо расположено между окружностями с радиусом, равным нулю, и первым радиусом) вычисляется средняя величина относительной освещенности. Суммарная освещенность в пределах кольца при этом делится на площадь кольца и на величину средней освещённости круга, охватывающего все кольца. Полученная таким образом относительная освещенность нормируется к целому числу (в нашей работе максимум относительной освещенности равен 255). Два полученных таким образом N-мерных вектора сравниваются друг с другом, и по величине суммарной квадратичной невязки компонент делается вывод о степени их идентичности. Если снимки получены в разных масштабах, то необходимо варьировать также внешний радиус кольцевого фильтра,

оставляя неизменными их относительные радиусы. При таком подходе алгоритм распознавания идентичных точек на левом и правом снимках становится нечувствительным к повороту снимков, изменению масштаба и абсолютным величинам освещенностей.

Естественно, при таком подходе метод обладает «неподъемными» затратами вычислительного времени. Для ускорения поиска подобных изображений дескрипторы растровых изображений вычисляются следующим образом. При $N = 8$ в восьмимерном гиперкубе с длиной ребра в 256 единиц мы образуем массив из 256 реперных точек, с равномерной плотностью заполняющих его пространство. Для каждого внутреннего пикселя (отступаем от каждого края растрового изображения на величину, равную внешнему радиусу кольцевого фильтра) вычисляем координаты 8-мерного вектора по алгоритму, описанному выше. Далее находим в пространстве гиперкуба ближайшую реперную точку, и записываем по адресу этого репера относительное расстояние между этими точками, если там еще не было записи или наше относительное расстояний меньше записанного там прежде. Просканировав все внутренние пиксели растрового изображения, мы получаем его дескриптор.

Как показали вычислительные эксперименты на больших базах данных, поиск подобных изображений по дескрипторам занимает небольшое время — для ПК с тактовой частотой 2 Ггц порядка нескольких секунд. Расчет самого дескриптора составляет для растрового изображения с 256 градациями серого цвета размером более миллиона пикселей порядка десятка секунд. При работе с разномасштабными изображениями это время вырастает на порядок. Вероятность распознавания подобного изображения при параметрах, приведенных выше, составила 0.99.

Литература

[1] *I. Kalajev, V. Protasov, V. Shapoval* An Efficient Method for the Recognition of Three-Dimensional Objects from a Contour Segment // Pattern Recognition and Image Analysis.— 1998. — Vol.8. No 2— P. 196–197.

Информационная система, поддерживающая процесс построения моделей прогнозирования свойств химических соединений

Сенкова Т. Н., Кумсков М. И., Миловидов А. Н., Свитанько И. В.

qsar_msu@mail.ru, kumskov@mail.ru

Москва, Кафедра Вычислительной математики, Мехмат МГУ

Накоплены значительные массивы экспериментальных данных по физико-химическим свойствам, биологической активности, а также боль-

шое число QSPR- и QSAR-моделей (Quantitative Structure-Property Relationship и Quantitative Structure-Activity Relationship). Актуальной является задача унификации представления таких моделей в рамках информационной системы (ИС), способной единообразно хранить как данные о дескрипторах молекул, так и параметры этапов построения моделей прогнозирования, использующих разнообразные методы классификации и распознавания образов.

Задача системы — информационная поддержка процесса построения и сравнения QSRP/QSAR-зависимостей (различающихся как по качеству, так и по сложности моделей прогнозирования) за счет единообразного их хранения на основе реляционной СУБД.

Система ориентирована на представление ранее найденных QSAR-зависимостей, основанных на различных (по детализации) формах хранения молекул, которые могут иметь вид графа или молекулярной поверхности, «раскрашенной» локальным физико-химическим свойством (ЛФХС), например, с использованием потенциала или способности отдавать или принимать электрон (донорно-акцепторные факторы) и т. п. При формировании признаков на молекулярной поверхности вычисляются экстремумы ЛФХС — особые точки. В результате структурным объектом, подлежащим анализу и классификации, становится маркированный граф, вершины которого располагаются в особых точках на молекулярной поверхности.

Для анализа конкретного свойства неизвестно, на каком уровне следует описывать молекулы — на топологическом, на планарном или на пространственном. Выбор уровня представления молекул и адаптация признаков для конкретного свойства проводится математиком-исследователем динамически в процессе построения QSAR-моделей.

Пользователь-химик должен суметь запустить прогноз новых молекул без присутствия математика спустя месяцы и годы после построения модели. Проблема состоит в том, что последовательность преобразования молекул может быть построена на основе модулей и программ, имеющих различные алгоритмы и различные значения входных параметров. Так, существует десяток программ, позволяющих построить пространственную укладку атомов молекулы, т. е. вычислить ее «3D-конформацию». Эти программы используют различные принципы, начиная с методов молекулярной механики и кончая квантово-химическими расчетами. При проведении прогнозирования свойств новых молекул химиком, эти новые структуры должны пройти тот же путь преобразований теми же самыми программными модулями, что прошли молекулы обучающей выборки при построении модели математиком. Таким образом, все версии программ, участвующих в цепочке преобразований молекул, и их параметры

должны единообразно сохраняться в репозитории, чтобы была возможность полной повторяемости результатов модели при прогнозировании.

Молекулярные структуры после построения QSAR модели не должны «стираться», поскольку, во-первых, они определяют «область допустимых значений» (ОДЗ) модели и, во-вторых, могут быть повторно использованы (при появлении новых данных о механизмах действия) для QSAR моделирования.

ИС должна поддерживать три типа пользователей:

1. *Пользователь, ответственный за контент ИС*, выполняет загрузку структурных массивов в базу данных, отвечает за проверку правильности данных о свойствах, за формирование фрагментного представления молекул и его описания и за загрузку пространственного представления молекул и описания его характеристик; пополняет каталог выборок, представленных в системе, и их описаний.
2. *Пользователь-исследователь* выполняет построение и верификацию моделей — отвечает за создание и тестирование программ в среде MATLAB, реализующих модели, выбранные для поддержки в ИС, за сохранение параметров моделей с лучшими показателями по качеству прогноза, за сохранение протоколов тестирования моделей и их аналитического сравнения, а также за определение «области действия» модели.
3. *Конечный пользователь* проводит расчеты по прогнозированию свойств новых соединений, используя модели, поддерживаемые в ИС; сохраняет протоколы расчетов, делает запросы на изменение, включая необходимую корректировку моделей или области их действия.

В моделях системы будут реализованы ранее разработанные методы [1–3], основанные на использования структурных фрагментов молекул, а также:

1) метод идентификации и классификации информативных особых точек (ОТ) на пространственной молекулярной структуре, описывающих физико-химические экстремумы локального свойства на молекулярной поверхности;
2) метод формирования матрицы «молекула–признак» на основе символьных дескрипторов (ЭО — элементов описания), представляющих 3D конфигурации двоек, троек и четверок ОТ на молекулярных поверхностях;
3) метод идентификации ЭО — определение отношения эквивалентности ЭО с использованием нечеткого (fuzzy) кодирования взаимных расстояний между ОТ в ЭО;
4) методы расчета молекулярного подобия по матрице «молекула–элемент описания».

Работоспособный прототип системы предполагается создать на основе реляционной СУБД в двухзвенной архитектуре клиент-сервер. На первом этапе в репозиторий планируется загрузить пять выборок со структурами молекул и их свойствами, а для вычисления «топологического сходства» молекул на фрагментах будут построены индексные файлы, описывающие состав молекул в виде цепочек из 2-х, 3-х и четырех атомов. Для вычисления «пространственного сходства» планируется загрузить 3D конформации молекул, рассчитанные потенциалы на молекулярных поверхностях и сформировать соответствующие индексные файлы по дескрипторам, описывающим 3D отношения между активными центрами.

Работа поддержана грантом РФФИ № 07-07-00282.

Литература

[1] *Svitanko I. V., Kumskov M. I., Tcheboukov D. E., Dolmat M. S., Zakharov A. M., Ponomareva L. A., Grigor'eva S. S., Chichua V. T.* QSAR modeling on the base of electrostatic molecular surface (amber fragrances) // 16-th Eur. Symp. on Quantum Structure-Activity Relationships and Molecular Modelling, Italy: EuroQSAR-2007. —

[2] *Захаров А. М., Кумсков М. И., Пономарева Л. А.* Эволюционный алгоритм построения моделей «структура-свойство» для молекулярных поверхностей с использованием аппарата нечеткой логики // ММРО-12, Москва, 2005. — С. 109–112.

[3] *Захаров А. М., Свитанько И. В., Григорьева С. С., Чичуа В. Т.* Поиск особых точек на молекулярных поверхностях с использованием нечеткого кластер-анализа // ММРО-12, Москва, 2005. — С. 333–335.

Классификация ресурсов знаний в системе извлечения информации из текста

Сулейманова Е. А.

yes@helen.botik.ru

Переславль-Залесский, ИПС РАН

Ядром интеллектуальной системы анализа текстов является сплав знаний различной природы. Соотношение между компонентами системы, которые ведают разными знаниями, зачастую далеко не очевидно [3, 4].

В статье описан подход к систематизации знаний системы интеллектуального анализа текста в контексте задачи извлечения информации [2].

Основание классификации

Знания систематизируются по трем измерениям, которым поставлены в соответствие бинарные дифференциальные признаки со значения-

ми: «предметные» — «лингвистические», «о классах» — «об индивидах», «априорные» — «из текстов».

Признаки делят пространство знаний на 8 секторов, каждому из которых соответствует свой набор значений признаков, Рис. 1. Рассмотрим, что представляют собой эти сектора (нумерация секторов произвольная), и какой компонент ресурса знаний соответствует каждому из них.

Типы знаний и компоненты системы

1. *«предметные», «о классах», «априорные»*
   Этому набору признаков отвечает *онтология* — общие знания системы об устройстве мира и предметной области в терминах концептов (классов сущностей) и их свойств. Концепты, атрибуты, признаки, отношения для удобства будем называть *элементами онтологии*. Элементы онтологии организованы в иерархические структуры.
2. *«предметные», «об индивидах», «априорные»*
   Наряду с общими знаниями о концептах интеллектуальная система должна располагать сведениями и о свойствах некоторых конкретных индивидов — экземпляров концептов. Эти сведения содержатся в *базе априорных фактов*.
3. *«предметные», «об индивидах», «из текстов»*
   Это целевые знания — знания о свойствах конкретных объектов, извлекаемые системой из текстов. Из них формируется *база текстовых фактов*[1].
4. *«предметные», «о классах», «из текстов»*
   В контексте перспективной задачи автоматизации пополнения онтологии на основе текстов в этот сектор должны попасть обнаруженные системой в текстах *знания о ранее не известных* (не отраженных в онтологии) *концептах и/или новых свойствах известных концептов*.
5. *«лингвистические», «о классах», «априорные»*
   Сюда можно отнести все лексикографические источники (за исключением упомянутых в последующих пунктах) и лингвистические модели. Отдельно стоит сказать о *словаре базовой предметной лексики системы*. Он организован как дескрипторный словарь: дескриптор представляет множество синонимических выражений. В отличие от тезауруса, дескрипторы в словаре не связаны друг с другом никакими парадигматическими отношениями, а содержат лишь ссылки на элементы онтологии. Таким образом, лингвистические знания от-

[1] В соответствии с предложенным ранее [1] разделением подходов к извлечению информации на извлечение информации в «слабом» и «сильном» смысле, упомянутые здесь текстовые факты — это факты в «сильном» смысле, то есть результат переработки первичных текстовых фактов (извлеченных в «слабом» смысле) в знания.

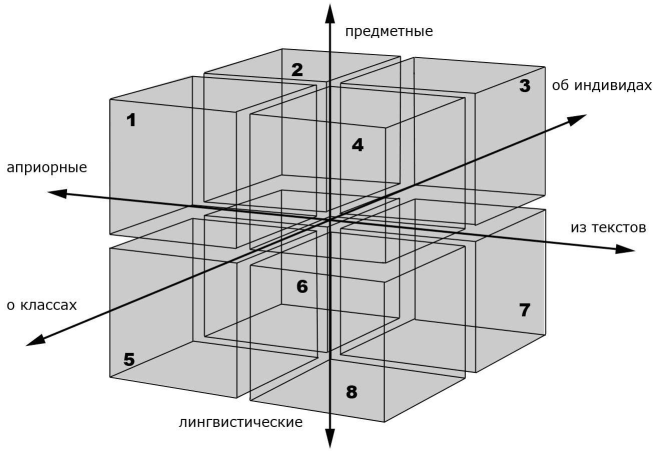


Рис. 1. Куб знаний.

делены от экстралингвистических и отпадает необходимость в тезаурусе как некоторой переходной форме.

6. *«лингвистические», «об индивидах», «априорные»*
   В системе это *словарь собственных имен*. Его словарные входы тоже содержат ссылки на предметные знания, но это ссылки на конкретные объекты (экземпляры концептов) из базы априорных фактов. Кроме того, словарным входам приписаны довольно общие категории типа «ФИО», «название организации» (такие категориальные метки удобно использовать на этапе извлечения первичных текстовых фактов).
7. *«лингвистические», «об индивидах», «из текстов»*
   В процессе анализа текста и извлечения фактов строится динамический *словарь новых собственных имен* — в него включаются имена обнаруженных в тексте объектов. После проверки динамический словарь может служить источником пополнения словаря собственных имен системы.
8. *«лингвистические», «о классах», «из текстов»*
   Последний сектор возвращает нас к теме автоматизированного пополнения онтологии: здесь речь идет о соответствующем *пополнении словаря базовой предметной лексики системы*.

Заключение

(Пополняемая) онтология и базы априорных и текстовых фактов (верхние сектора) образуют *базу предметных знаний системы*. Нижние сектора — это база *лингвистических знаний*. Отметим, что границу

между знаниями о классах и экземплярах (передняя и задняя половины «куба знаний») можно считать незыблемой, тогда как грань между знаниями априорными и извлеченными из текстов (левая и правая половины) периодически стирается.

Работа выполнена при поддержке РФФИ, проект № 05-01-00442а.

Литература

[1] *Куршев Е. П., Сулейманова Е. А.* Представление предметных знаний в системах интеллектуального анализа текста // Междунар. конф. «Программные системы: теория и приложения», ИПС РАН, Переславль-Залесский, октябрь 2006 г. — Т. 1 — М.: Физматлит, 2006. — С. 379–390.

[2] *Куршев Е. П., Кормалев Д. А., Сулейманова Е. А., Трофимов И. В.* Исследование методов извлечения информации из текстов с использованием автоматического обучения и реализация исследовательского прототипа системы извлечения информации // ММРО-13 (наст. сб.) — 2007. — С. 602–605.

[3] *Нариньяни А. С.* Кентавр по имени ТЕОН: тезаурус+онтология // Междунар. семинар Диалог'2001 по компьютерной лингвистике и ее приложениям. — Т. 1. — Аксаково, 2001. — С. 184–188.

[4] *Нариньяни А. С.* ТЕОН-2: от тезауруса к онтологии и обратно // Междунар. семинар Диалог'2002 «Компьютерная лингвистика и интеллектуальные технологии». — Т. 1. — М.: Наука, 2002. — С. 307–313.

Комплексная методика электрокардиографической диагностики на основе эвристических и количественных подходов дипольной электрокардиотопографии

Титомир Л. И., Трунов В. Г., Айду Э. А. И.

titomir@iitp.ru

Москва, Институт проблем передачи информации им. А. А. Харкевича РАН

Исследования последних десятилетий показали, что точность кардиологической диагностики может быть существенно повышена при использовании оптимальных систем отведений и компьютерных технологий, обеспечивающих образную визуализацию данных и определение наиболее информативных параметров электрофизиологического состояния сердца. Наиболее информативным методом неинвазивной электрокардиографии является многоэлектродное картирование электрического поля сердца с использованием нескольких десятков синхронных отведений [1]. Основной недостаток этого метода — слишком сложная измерительная процедура, неприемлемая во многих случаях диагностического исследования.

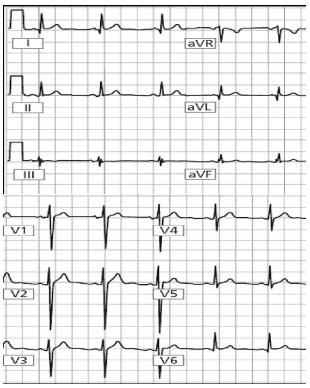


Рис. 1. Стандартные электрокардиограммы.

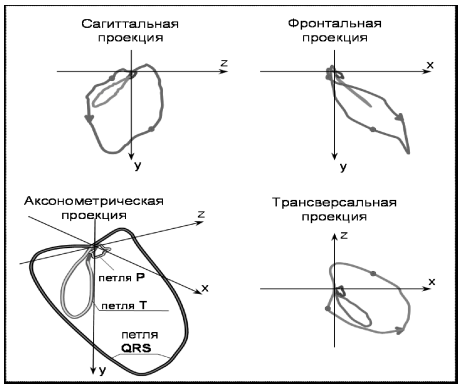


Рис. 2. Векторкардиограммы.

Стандартная электрокардиография с 12 отведениями характеризуется простой измерительной процедурой, однако в ней используются диагностические параметры без достаточного физико-физиологического и анатомического обоснования, а также без содержательно образного (топографического) представления этих параметров. Результаты измерений изображаются в виде скалярных кривых, Рис. 1. В векторкардиографии используются три компоненты вектора сердца, измеряемые ортогональной системой отведений (в частности, системой Франка или системой Макфи-Парунгао), и изображения траектории конца этого вектора (векторкардиографические петли), дающие наглядное представление основных пространственных направлений и динамики изменения величины и ориентации эквивалентного электрического диполя сердца, Рис. 2. Еще бóльшая наглядность достигается при использовании предложенного упрощенного метода электрокардиографического картирования — дипольной электрокардиотопографии (ДЭКАРТО), в котором исходными данными, как и в векторкардиографии, являются компоненты вектора сердца, Рис. 3 [2]. Из-за ограниченности исходной информации здесь используются наиболее простые модели электрогенной зоны, в частности, односвязный фронт деполяризации. Распределение электрофизиологических состояний желудочков в заданный момент времени в проекции на сферу отображения, окружающую сердце, представляется в форме моментных дэкартограмм деполяризации и реполяризации. Определяются также суммарные карты возбуждения — карты прихода активации, длительности активации и ускорения реполяризации. Для топографического представления характеристик предсердий используется карта активации предсердий с распределением относительного параметра разме-

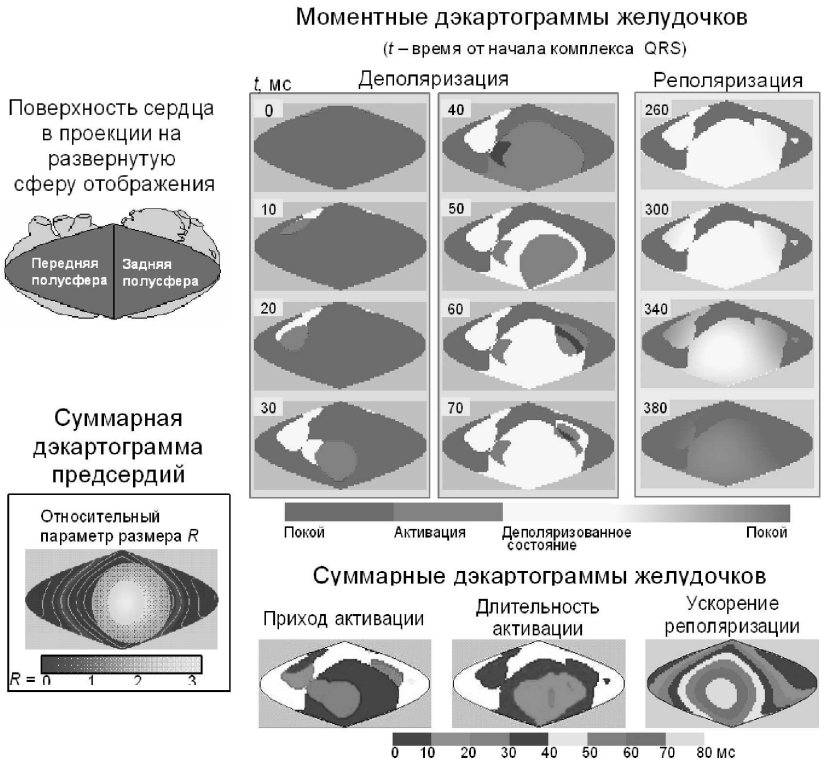


Рис. 3. Дэкартограммы (дипольные электрокардиотопограммы).

ра предсердия на сфере отображения. В методе ДЭКАРТО достигается визуальная наглядность характеристик с привязкой к анатомическим элементам поверхности сердца и возможность получения высокоинформативных количественных параметров на основе достаточно простой измерительной процедуры. В частности, для целей дифференциальной диагностики гипертрофии сердца по дэкартограммам предложены количественные индексы ILVH и IRVH (для гипертрофии левого и правого желудочков, соответственно); они вычисляются как интегралы длительности активации по областям дэкартограммы, на которые проецируются соответствующие анатомические области сердца. Сопоставление значений указанных индексов с заданным порогом обеспечивает распознавание гипертрофии [3, 4]. Получены также количественные параметры для оценки интенсивности и локализации ишемических и инфарктных очагов в желудочках, дифференциальной диагностики увеличения предсердий и выявления других патологических состояний сердца.

Экспериментально-клиническими исследованиями было показано, что совместный анализ векторкардиограмм и дэкартограмм с использованием как эвристических подходов к оценке визуальных изображений, так и методов распознавания образов по количественным параметрам, позволяет существенно повысить точность диагностики гипертрофии сердца по сравнению со стандартной электрокардиографией.

Работа выполнена при поддержке РФФИ проект № 07-01-00025-а.

Литература

[1] *Titomir L. I., Kneppo P.* Bioelectric and Biomagnetic Fields. Theory and Application in Electrocardiology. — Boca Raton etc.: CRC Press, 1994. — 336 p.

[2] *Титомир Л.И., Трунов В. Г., Айду Э. А. И.* Неинвазивная электрокардиотопография — Москва: Наука, 2003. — 198 с.

[3] *Блинова Е. В., Сахнова Т. А., Сергакова Л. М., Атауллаханова Д. М., Ощепкова Е. В., Лазарева Н. В., Айду Э. А. И., Трунов В.Г., Титомир Л. И.* Новые подходы к диагностике гипертрофии левого желудочка методом дипольной электрокардиотопографии. — Терапевт. архив, 2005. — Т. 77, № 4. — С. 8–10.

[4] *Titomir L. I., Trunov V. G., Aidu E. A. I., Sakhnova T. A., Blinova E. V.* Recognition of Right Ventricular Hypertrophy Using the `DECARTO` Images of the Cardioelectric Characteristics. — 2007 — `www.measurement.sk/2007/S2/p2.html`.

Многоагентная среда для проведения экспериментов по защите компьютерных сетей

Уланов А. В., Котенко И. В.

`ulanov@iias.spb.su, ivkote@iias.sbp.su`

Санкт-Петербург, Институт информатики и автоматизации РАН

В работе предлагается подход и реализованная среда многоагентного моделирования для анализа существующих и перспективных методов защиты от атак «распределенный отказ в обслуживании». Подход базируется на представлении систем, реализующих компьютерные атаки и защиту от них, в виде команд интеллектуальных агентов. В работе реализован ряд методов кооперативной защиты и проведено исследование их эффективности.

Защита от атак DDoS

Один из наиболее опасных классов атак в Интернете — это атака «распределенный отказ в обслуживании» (Distributed Denial of Service, DDoS). Предполагается, что перспективная система защиты от DDoS должна работать на основе кооперации различных систем, сетей и глобальных механизмов защиты, расположенных как в отдельных подсетях, так и в Интернете.

К распределенным кооперативным механизмам DDoS защиты [1] относятся: перенос ресурсов (Server Roaming), изменение количества ресурсов, дифференциация ресурсов (Market-based Service Quality Differentiation (MbSQD), Transport-aware IP router architecture), аутентификация (Secure Overlay Services (SOS)), а также механизмы, реализующие отслеживание (Gateway-based mechanism) с помощью разметки пакетов, хранения сигнатур или генерации служебных пакетов.

Цель данной работы заключается в разработке среды моделирования для исследования атак и механизмов защиты (на примере DDoS) и формулировании обоснованных рекомендаций по выбору эффективных механизмов защиты. В докладе рассматриваются предложенные подход, разработанная среда моделирования и проведенные эксперименты по исследованию интеллектуальных кооперативных механизмов защиты.

Подход и модели защиты

Предлагаемый подход к моделированию заключается в следующем. Исследуемые процессы рассматриваются как взаимодействие команд программных агентов в динамической среде, заданной посредством модели сети Интернет [2]. Поведение системы проявляется в локальных взаимодействиях отдельных агентов.

Существует, по крайней мере, три класса команд агентов: злоумышленники, команда защиты и агенты-пользователи. Агенты различных команд могут быть в состоянии безразличия, кооперироваться или соперничать.

Выделено два класса агентов команды атаки: «демон» — исполнитель атаки, «мастер» — координатор команды. В соответствии с общим подходом к защите от атак DDoS [1, 2], выделены следующие классы команды защиты: «сэмплер» — обработка данных трафика, «детектор» — обнаружение атаки, «фильтр» — фильтрация трафика, «ограничитель» — ограничение трафика, агент «расследования». Команда защиты совместно реализует определенный механизм защиты. Различные команды защиты могут взаимодействовать по разным схемам.

В работе продемонстрировано функционирование следующих механизмов защиты, осуществляющих классификацию вредоносного трафика на основе построенной модели нормального трафика [1]: Count Filtering (HCF), Source IP address monitoring (SIPM), Bit per Second analysis (BPS). В HCF используется предположение, что пакеты из одной подсети проходят одинаковое количество скачков (хопов) от отправителя до получателя. SIPM построен на том, что в начале атаки появляется много пакетов от новых для системы отправителей. BPS определяет атаку по превышению заданного порога трафика.

Для построения модели нормального сетевого трафика используется следующий способ. Легитимные клиенты обращаются к защищаемому серверу, а он обрабатывает их запросы, таким образом создавая выборку нормального трафика. Аналогичным образом создается выборка вредоносного трафика, при этом задействуются атакующие агенты команды атаки. При обучении настраиваются внутренние параметры методов защиты, такие как интервал получения данных о трафике и сдвиг этого интервала.

Основное внимание в кооперативных механизмах защиты уделяется методам распределенной фильтрации и ограничения трафика. Проведено моделирование следующих кооперативных механизмов: DefCOM, COSSACK и пяти предложенных классов. Предложены такие схемы кооперации: без кооперации, на уровне фильтров, на уровне сэмплеров, слабая кооперация и полная кооперация.

Среда моделирования и эксперименты

Предложенный подход предполагает разработку среды моделирования, архитектура которой включает следующие компоненты [2]: базовая среда моделирования, пакет имитации сети Интернет, среда многоагентного моделирования, библиотека предметной области. Первый компонент является базовым для остальных.

Данная архитектура была реализована для многоагентного моделирования распределенных механизмов защиты с использованием базовой системы моделирования дискретных событий OMNeT++, симулятора сетей INET Framework и программных моделей, разработанных на C++.

Модель противоборства в среде моделирования задается следующими параметрами: топологией и конфигурацией сети, конфигурацией команд атаки, параметрами атак DDoS, конфигурацией команд защиты, параметрами защиты от DDoS, параметрами кооперации команд. В качестве исследуемых выходных параметров механизмов защиты рассматриваются: количество ложных срабатываний; количество пропусков атак; процент трафика атаки и нормального трафика в исследуемой сети; время реакции и др.

Исследуемые механизмы защиты основаны на реализации режимов обучения и собственно защиты с обновлением данных. В режиме обучения производится сбор данных по заведомо легитимному трафику. В режиме защиты, на основе сравнения текущих данных с модельными, выполняется обработка сетевого трафика. Несоответствие считается аномалией или атакой, и принимаются контрмеры. Если аномалий не обнаружено или они малы, то данные заносятся в модель, то есть происходит ее обновление.

Заключение

С использованием разработанной среды моделирования проведено множество экспериментов по исследованию различных типов атак, нахождению оптимальных параметров механизмов защиты, сравнению различных механизмов защиты, режимов кооперации и механизмов адаптации команд агентов. В дальнейшем планируется совершенствование среды моделирования, расширение библиотеки предметной области и системы многоагентного моделирования. Работа выполняется при финансовой поддержке РФФИ (проект № 07-01-00547), программы фундаментальных исследований ОИТВС РАН (контракт № 3.2/03) и Фонда содействия отечественной науке.

Литература

[1] *Уланов А. В., Котенко И. В.* Защита от DDoS-атак: механизмы предупреждения, обнаружения, отслеживания источника и противодействия // Защита информации. Инсайд. — 2007. — № 1 — 3.

[2] *Котенко И. В., Уланов А. В.* Команды агентов в кибер-пространстве: моделирование процессов защиты информации в глобальном Интернете // Сборник Института системного анализа РАН. — М.: URSS, 2006.

Модели и алгоритмы комплексного планирования модернизации и функционирования катастрофоустойчивой информационной системы

Юсупов Р. М., Соколов Б. В., Охтилев М. Ю., Потрясаев С. А.

yusupov@iias.spb.su, sokol@iias.spb.su

Санкт-Петербург, Институт информатики и автоматизации РАН

Анализ существующих и прогнозируемых кризисных и чрезвычайных ситуаций, повсеместно возникающих в настоящее время в различных предметных областях, показывает, что они перестают быть отраслевыми, а перерастают в аварии и катастрофы, имеющие уже межотраслевой характер [1]. В этих условиях исследовать и решать проблемы повышения катастрофоустойчивости как конкретных прикладных процессов (бизнес-процессов), так и информационной систем (ИС), обеспечивающих их выполнение, необходимо уже в рамках междисциплинарного подхода, интерпретируя соответствующие задачи как задачи управления структурной динамикой (УСД) указанными процессами и системами [2].

Главная особенность задачи планирования модернизации ИС состоит в том, что переход от «старой» (не являющейся катастрофоустоучивой) ИС к «новой» (модернизированной, катастрофоустойчивой) ИС не может быть проведен мгновенно. На практике это приводит к тому,

что на достаточно длительном интервале времени (периоде модернизации ИС) происходит совместная эксплуатация элементов и подсистем «старой» и «новой» ИС. Так, например, создание и ввод в эксплуатацию дублирующих и резервных ИС, располагающихся на значительном расстоянии от основной ИС, предполагает поэтапный ввод элементов и подсистем соответствующих комплексов автоматизации. Однако в этих условиях показатели качества и эффективности бизнес-процессов, поддерживаемых данными ИС, не должны ухудшаться. Таким образом, всякое изменение и развитие той или иной подсистемы (структуры) ИС объективно осуществляется одновременно с решением оперативных (текущих) задач, стоящих перед соответствующей бизнес-системой (БС). Поэтому и возникает необходимость совместной постановки задач комплексного планирования модернизации и функционирования ИС. При этом планирование должно осуществляться комплексно и затрагивать все основные элементы и подсистемы существующей и создаваемой ИС [2, 3, 4]. Совместное решение задач комплексного планирования модернизации и функционирования ИС предполагает: построение соответствующего полимодельного комплекса, описывающего все основные аспекты исследуемых процессов; разработку комбинированных методов, алгоритмов и методик многокритериального полимодельного синтеза программ управления модернизацией и функционированием существующей и внедряемой ИС.

В ходе проведенных исследований были предложены следующие основные фазы и этапы решения задачи комплексного планирования модернизации и функционирования (задачи выбора оптимальных программ управления структурной динамикой КАИС).

На первой фазе должно осуществляться формирование допустимых вариантов многоструктурных макросостояний КАИС или, говоря другими словами, должен проводиться структурно-функциональный синтез нового облика модернизируемой ИС, соответствующего складывающейся (прогнозируемой) обстановке. В указанной ситуации задачи, решаемые на данной фазе, сводятся к задачам структурно-функционального синтеза КАИС.

На второй фазе проводится выбор конкретного варианта многоструктурного макросостояния КАИС с одновременным синтезом (построением) адаптивных планов (программ) управления переходом КАИС из текущего в требуемое (выбранное) макросостояние. При этом рассматриваемые планы должны обеспечивать такое эволюционное развитие КАИС, при котором наряду с реализацией программ перехода из соответствующих макросостояний предусматривается одновременно и реализация программ устойчивого управления КАИС в промежуточных макрососто-

яниях. На данной фазе приходится решать совокупность частных задач многоуровневой и многоэтапной оптимизации. Обобщенный алгоритм решения данных задач должен включать следующие этапы (шаги) [2].

Шаг 1. В интерактивном режиме осуществляется автоматизированная подготовка, контроль, анализ и ввод всей исходной информации, необходимой для решения задачи управления структурной динамикой КАИС.

Шаг 2. Планирование проведения комплексного моделирования процессов адаптивного управления функционированием и развитием КАИС в текущей и прогнозируемой обстановке, планирование проведения вычислительных экспериментов в имитационной системе (ИмС), определение состава и структуры моделей, методов и алгоритмов решения частных задач моделирования, расчет времени, необходимого для решения указанных задач.

Шаг 3. Генерирование, на основе проведения комплексного моделирования, допустимых вариантов функционирования КАИС в исходном, промежуточных и требуемых многоструктурных макросостояниях, вывод результатов моделирования ЛПР, предварительный интерактивный структурно-функциональный анализ указанных результатов моделирования; формирование классов эквивалентных многоструктурных макросостояний КАИС.

Шаг 4. Автоматизированный ввод допустимых вариантов функционирования КАИС, проверка корректности заданной системы ограничений, окончательный выбор необходимого уровня агрегирования при описании моделей УСД КАИС, вычислительной схемы и плана вычислительных экспериментов по поиску оптимальных программ УСД КАИС.

Шаг 5. Поиск оптимальных программ управления структурной динамикой КАИС, при которых обеспечивался переход из заданного в синтезируемое многоструктурное макросостояние КАИС, устойчивое управление функционированием КАИС в промежуточных многоструктурных макросостояниях.

Шаг 6. Имитация условий реализации оптимального плана управления переходом КАИС из текущего в требуемое (выбранное) макросостояние при наличии возмущающих воздействий и с учетом различных вариантов их компенсации на основе методов и алгоритмов оперативного управления.

Шаг 7. Структурная и параметрическая адаптация плана, СПМО и информационного обеспечения ИмС к возможным (прогнозируемым на имитационных моделях) состояниям объекта управления (ОУ), управляющей подсистемы (УП), внешней среды. В ходе указанной адаптации, кроме того, вводится необходимый уровень структурной избыточ-

ности КАИС, обеспечивающий на этапе реализации плана компенсацию не предусмотренных в плане возмущающих воздействий.

После проведения требуемого числа вычислительных экспериментов осуществляется оценивание устойчивости сформированного адаптивного плана УСД КАИС.

Шаг 8. Вывод полученных результатов комплексного адаптивного планирования применения КАИС, их интерпретация и коррекция ЛПР.

Одно из главных достоинств предлагаемого метода поиска оптимальных программ УСД КАИС состоит в том, что в ходе формирования вектора программных управлений в финальный момент времени, наряду с оптимальным планом, одновременно получаем и то искомое многоструктурное макросостояние, находясь в котором КАИС сможет выполнять поставленные перед ней задачи в складывающейся (прогнозируемой) обстановке с требуемой степенью устойчивости.

В настоящее время разработаны комбинированные методы и алгоритмы решения задач выбора оптимальных программ УСД КАИС в централизованном и децентрализованном режимах ее функционирования. В качестве базового метода предложено использовать сочетание метода ветвей и границ и метода последовательных приближений. Теоретическое обоснование данного метода основано на доказанной теореме о свойствах релаксированной задачи выбора оптимальной программы УСД КАИС. Особенности реализации предлагаемого комбинированного метода исследованы при решении различных прикладных задач [2, 3, 4].

При исследовании различных классов задач УСД КАИС были предложены алгоритмы параметрической и структурной адаптации соответствующих моделей, основанные на методах нечеткой кластеризации и анализа иерархий, методах аналитико-имитационного моделирования. Кроме того, для проверки конструктивности использования предлагаемого подхода к решению рассматриваемых задач осуществлялась разработка прототипа программного обеспечения процессов поиска оптимальных вариантов УСД КАИС различного целевого назначения [2].

Работоспособность программного комплекса была проверена на примере решения задач планирования модернизации и функционирования центра управления полетами навигационными космическими аппаратами (ЦУП НКА). Разработаны и исследованы несколько прототипов программ, реализующих решения перечисленных задач. В докладе приводятся результаты исследования устойчивости планов совместного функционирования основного и дублирующего ЦУП НКА при различных сценариях воздействия внешней среды. Для описания моделей планирования в разработанном программном комплексе используется широко распространенный сегодня язык XML. Применение XML позволяет в одном

файле хранить как исходные данные модели (процессы, ресурсы и их характеристики), так и результаты выполнения алгоритма построения оптимального плана, а также обеспечивает простоту сопряжения с другими программными комплексами, с помощью которых решаются задачи оптимального распределения ресурсов, отображения полученных результатов, интерактивного взаимодействия с лицом, принимающим решения.

Исследования выполнены при поддержке РФФИ (гранты №07-07-00169, №06-07-89242, №05-08-18111), ОИТВС РАН (проект №2.5), СПб Научного Центра РАН (проект №112).

Литература

[1] *Юсупов Р. М.* Наука и национальная безопасность. — М: Наука, 2006. — 290 с.

[2] *Охтилев М. Ю., Соколов Б. В., Юсупов Р. М.* Интеллектуальные технологии мониторинга и управления структурной динамикой сложных технических объектов. — М: Наука, 2006. — 410 с.

[3] *Калинин Б. Н., Соколов Б. В.* Многомодельный подход к описанию процессов управления космическими средствами // Теория и системы управления. — 1995. — №1. — С. 56–61.

[4] *Соколов Б. В., Юсупов Р. М.* Комплексное моделирование функционирования автоматизированной системы управления космическими аппаратами // Проблемы управления информатики. — 2002. — №5. — С. 103–117.

Развитие технологии решения связанных по экспертному заключению задач, основанной на логических тестах и средствах когнитивной графики

Янковская А. Е.

yank@tsuab.ru

Томск, Томский архитектурно-строительный университет

В докладе излагается развитие технологии решения нового класса задач высокого уровня сложности, связанных по экспертному заключению (ЭЗ) [1], в плане оценки одного и того же объекта (явления, ситуации, состояния и др.) по каждой из альтернативных групп признаков (АГП), для одной из которых возможно экспертное заключение, соотнесенное с каждой из АГП, по которой эксперт не может дать ЭЗ, и выявления закономерностей в данных и знаниях по каждой АГП и перекрестных связей. Под ЭЗ понимается заключение эксперта (высококвалифицированного специалиста в соответствующей проблемной области) по описанию объекта, например, в медицине — постановка диагноза, прогноз. В докладе рассматриваются:

— матричная модель представления данных и знаний [2];

— теория логических тестов — безусловных, смешанных (представляющих собой оптимальное сочетание безусловных и условных составляющих) диагностических тестов (БДТ, СДТ) [2, 3];
— алгоритмы построения всех минимальных, всех или части БДТ [4] и выбора оптимального для принятия достоверных решений подмножества БДТ, СДТ на основе глубоких оптимизирующих логико-комбинаторных (л-к), а также генетических преобразований [5];
— методы оценки весовых коэффициентов признаков и тестов [6];
— методы ортогонализация ДНФ булевых функций;
— методы логического вывода на основе построенных тестов;
— методы коллективного принятия решений [6];
— графические средства, в том числе когнитивные, ориентированные на пользователей различной квалификации [2].

При логическом (л-к, логико-вероятностном, логико-комбинаторно-вероятностном) выводе будут учтены дополнительные факторы и различные направления принятия решений, связанные воедино моделью представления данных и знаний, адекватной широкому кругу конкретных и междисциплинарных областей при комплексном решении задач диагностического, классификационного, прогностического, организационно-управленческого характера.

Описание одного и того же объекта в ряде проблемных областей (медицина [1], экология, экобиомедицина, геоэкология, психология и др.) может быть осуществлено в разных признаковых пространствах, по одному из которых эксперт может дать заключение, а по другим — нет [1]. В связи с этим автором было введено понятие «задач, связанных по ЭЗ» и предложена идея решения таких задач. Развитие технологии решения задач, связанных по ЭЗ, основано на трансформации результатов решения одной (1-й) задачи на основе экспертных знаний в решение других задач, взаимосвязанных с результатом решения 1-й задачи по ЭЗ. Развиваемая технология позволит выявить закономерности в данных и знаниях, определить взаимосвязь между компонентами проблемной среды (включающей информационную составляющую каждой из задач) и реализовать возможность принятия решений по информационной составляющей каждой задачи и по оптимальной АГП, сформированной из характеристических признаков всех задач.

Предложенная автором доклада идея трансформации решения задач связана с практической необходимостью соотнесения различных параметров, полученных на основе тех или иных измерений с традиционными (общепринятыми), определяемыми экспертами по другим характеристикам, имеющей место в области медицины, управления, психологии,

образовании и др. Развиваемая технология будет реализована в интеллектуальном инструментальном средстве (ИИС) ИМСЛОГ [7].

Знания по 1-й задаче представляются в виде матрицы описаний объектов (ситуаций, состояний, явлений) и матриц различений трех типов, каждый из которых задает совокупность различных механизмов классификации (зависимых, независимых и реализующих последовательности действий) объектов, разбивающих их на классы эквивалентности. Данные по каждой из других задач представляются матрицей описаний тех же самых объектов по соответствующей АГП, и матрицами различений, используемых при решении 1-й задачи.

Создаваемые алгоритмы параллельного адаптивного перекодирования и декодирования признаков расширят класс решаемых задач [8].

Введены новые виды закономерностей: подмножества сигнальных признаков 1-го и 2-го рода, указывающих на возможность переходов объектов из одного образа в другой и из состояния исследуемого объекта, принадлежащего одному образу, в другой образ, соответственно, а также АГП по решаемым задачам, перекрестные связи, оптимальная группа признаков.

Разработаны алгоритмы выявления новых видов закономерностей. Введены новые критерии оптимизации выбора оптимального подмножества безусловных БДТ и СДТ и намечены пути развития алгоритмов (л-к, на основе метода анализа иерархий, генетический) их построения.

Развитие графических, в том числе когнитивных, средств осуществляется в плане отражения вероятностных характеристик и учета нового класса задач и связанных с ним новых видов закономерностей. Развитие когнитивных управляющих элементов существенно расширит возможности пользовательского интерфейса.

Воплощение вышеизложенных подходов и методов в технологию решения связанных по ЭЗ задач, реализуемой на основе ИИС ИМСЛОГ в прикладных интеллектуальных системах, существенно расширит границы их применимости при комплексном решении диагностических, прогностических, организационно-управленческих задач большой размерности для широкого круга проблемных и междисциплинарных областей (медицина, психология, проектирование, экология, биология, социология, геоэкология, экобиомедицина, управление, генетика, управление и др.).

Работа выполнена при поддержке РФФИ, проект № 07-01-00452а.

Литература

[1] *Yankovskaya A. E., Ametov R. V., Muratova E. A., Chernogoryuk G. E., Mandel I. A.* Information technology for solving of problems connected on expert conclusion and construction of medical intelligent system on basis of

this technology // Proc. Of The Second IASTED Intern. Multi-Conf. ACIT-ACA, June 20-24, 2005, Novosibirsk, Russia. — Pp. 187–192.

[2] *Янковская А. Е.* Логические тесты и средства когнитивной графики в интеллектуальной системе // Новые информационные технологии в исследовании дискретных структур: Докл. 3-ей Всерос. конф. с межд. участ., Томск: Изд-во СО РАН, 2000. — С. 163–168.

[3] *Янковская А. Е.* Синтез смешанных логических тестов на основе ускоренных шагово-циклических алгоритмов спуска // Математические методы распознавания образов, ММРО–11, Москва, 2003. — С. 224–226.

[4] *Гедике А. И., Янковская А. Е.* Построение всех безызбыточных безусловных диагностических тестов в интеллектуальном инструментальном средстве ИМСЛОГ // Интеллектуальные системы, Интеллектуальные САПР. Тр. Межд. науч.-тех. конф. Т. 1. — М.: Физматлит, 2005. — С. 209–214.

[5] *Колесникова С. И., Можейко В. И., Цой Ю. Р., Янковская А. Е.* Алгоритмы выбора оптимального множества безызбыточных диагностических тестов в интеллектуальных системах поддержки принятия решений // 1-я межд. конф. САИТ-2005 — Т. 1. — М.: КомКнига, 2005. — С. 256–262.

[6] *Yankovskaya A., Kolesnikova S.* An Approach to Calculation of Feature Weight Coefficients on the Base of Multisets Formalism in Intelligent Systems // Knowledge-Based Software Engineering. Proc. of the 6th Joint Conf. Vol. 108. IOS Press, 2004. — Pp. 159–168.

[7] *Yankovskaya A. E., Gedike A. I., Ametov R. V., Bleikher A. M.* IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition // Pattern Recognition and Image Analysis. — 2003. — Vol. 13, № 4. — Pp. 650–657.

[8] *Yankovskaya A.E., Ametov R.V., Muratova E.A.* Transformation of the Quntitative Feature in the IMSLOG Intelligent Software Tool // «Intelligent Systems», «Intelligent CAD's» Proc. of the Intern. Sc. Conf. — Vol. 3. — Moscow: Physmatlit, 2005. — Pp. 146.

Алгоритм параллельного адаптивного перекодирования признаков с учетом мнения экспертов

Янковская А. Е., Муратова Е. А.

yank@tsuab.ru, muratova@tpu.ru

Томск, Томский архитектурно-строительный университет,
Томский политехнический университет

Для решения задач, связанных по экспертному заключению, предложен 4-х этапный эвристический алгоритм адаптивного перекодирования количественных признаков в серию бинарных с учетом мнения экспертов для случая большого числа классов (образов).

Введение

В рамках технологии решения задач, связанных по экспертному заключению [1], весьма актуальна задача перекодирования количественных признаков в серию бинарных, позволяющего выявить наибольшее количество логических закономерностей в данных и знаниях [2] по каждой из альтернативных групп признаков, для одной из которых возможно экспертное заключение, соотнесенное с каждой из альтернативных групп признаков, по которой эксперт не может дать заключение.

Трудоемкость решения задачи перекодирования существенно возрастает с увеличением числа классов (образов) и необходимостью учета мнения эксперта (экспертов), дающего экспертное заключение (экспертные заключения) по результатам перекодирования.

Качество принимаемых решений зависит от количества выявленных логических закономерностей по каждому из видов закономерностей (константные, устойчивые, неинформативные, альтернативные, зависимые, существенные, сигнальные, все минимальные и безызбыточные различающие подмножества признаков, оптимальная группа признаков, перекрестные связи и весовые коэффициенты [3]) по каждой из заданной альтернативной группе признаков.

В развитие технологии решения задач, связанных по экспертному заключению, для матричной модели представления данных и знаний и введенного понятия закономерностей предлагается 4-х этапный алгоритм параллельного перекодирования признаков с учетом мнения экспертов. Отметим, что параллельность перекодирования возможна при пересечении признаковых пространств по альтернативным группам признаков.

Адаптивное перекодирование признаков в бинарные

На основе разработанных алгоритмов адаптивного перекодирования признаков [4, 5], рассмотренных для случая разделения двух классов (образов) при решении классической задачи распознавания [6], и экспериментальных исследований этих алгоритмов на базе интеллектуального инструментального средства (ИИС) ИМСЛОГ [7] в рамках развиваемой технологии предлагается 4-х этапный алгоритм перекодирования.

На 1-м этапе алгоритма (рассматривается случай для большого числа классов (образов)) разбиваемый количественный признак делится первоначально на несколько равномерных интервалов, число которых может определяться математически в зависимости от объема обучающей выборки или подбираться экспериментально в интерактивном режиме работы алгоритма, или на основе знаний экспертов исследуемой проблемной области. Это связано с тем, что наличие большого числа классов (образов) предполагает и достаточно большой размах значений исследуемого при-

знака. При этом значения признаков для отдельных классов (образов) могут быть достаточно близки, а значения для других классов (образов) могут значительно отличаться от «основной группы» значений и получаемое разбиение на интервалы может оказаться информативным только для небольшого числа классов (образов). Другие классы (образы) будут близки, вследствие чего не разделены. На 2-м этапе полученное разбиение фиксируется, а интервал (или интервалы), имеющий большую «плотность попадания», разбивается далее по одному или всем имеющимся способам (равномерное, неравномерное, комбинированное, экспертное [4, 5]). На 3-м этапе для каждой из альтернативных групп перекодированных признаков с учетом и без учета мнения эксперта (экспертов) выявляются вышеупомянутые логические закономерности. На 4-м этапе формируются информативные интервалы для перекодируемых признаков, соответствующие более высокому уровню выявленных логических закономерностей.

Для реализации эвристического алгоритма адаптивного перекодирования количественных признаков в серию бинарных был создан в ИИС ИМСЛОГ динамически подключаемый плагин.

Заключение

С целью расширения границ применения ИИС ИМСЛОГ для решения задач, связанных по экспертному заключению, создан оригинальный алгоритм параллельного адаптивного перекодирования признаков с учетом мнений экспертов для случая большого числа классов (образов). Данный алгоритм вписывается в концепцию построения архитектуры ИИС ИМСЛОГ и позволяет решать задачи в большом признаковом пространстве, особенно характерном для междисциплинарных областей, каковыми являются биомедицина, экобиомедицина, геоэкология, и ряд других.

Дальнейшие исследования будут направлены на сравнение получаемых результатов с применением различных подходов преобразования количественных признаков в серию бинарных.

Работа выполнена при поддержке РФФИ, проект № 07-01-00452.

Литература

[1] *Yankovskaya A. E., Ametov R. V., Muratova E. A., Chernogoryuk G. E., Mandel I. A.* Information technology for solving of problems connected on expert conclusion and construction of medical intelligent system on basis of this technology // Proc. Of The II IASTED Intern. Multi-Conf. Automation, Control, And Inf. Techn. (ACIT-ACA), June 20-24, 2005, Novosibirsk, Russia. — P. 187–192.

[2] *Янковская А. Е.* Логические тесты и средства когнитивной графики в интеллектуальной системе // Новые информационные технологии в иссле-

довании дискретных структур: Докл. 3-ей Всерос. конф. с межд. участ., Томск: Изд-во СО РАН, 2000. — С. 163–168.

[3] *Янковская А. Е.* Выявление закономерностей в альтернативных группах признаков, связанных по экспертному заключению // Науч. сессия МИФИ-2004. Сб. науч. тр. Т. 3. М., 2004. — С. 126–127.

[4] *Берестнева О. Г., Муратова А. Е., Янковская А. Е.* Эффективный алгоритм адаптивного кодирования разнотипной информации // Искусственный интеллект в XXI веке. Тр. Международного конгресса. Т. 1. М.: Физматлит, 2001. — С. 155–166.

[5] *Янковская А. Е., Муратова А. Е., Аметов Р. В.* Преобразование количественных признаков в интеллектуальном инструментальном средстве ИМСЛОГ // Интеллектуальные системы (AIS'05), Интеллектуальные САПР (CAD-2005). Труды Международных научно-технических конференций. Т. 1. — М.: Физматлит, 2005. — С. 282–287.

[6] *Журавлев Ю. И., Гуревич И. Б.* Распознавание образов и анализ изображений // Искусственный интеллект: В 3-х кн. Кн.2. Модели и методы: справ. / Под ред. Д. А. Поспелова. — М.: Радио и связь, 1990. — С. 149–191.

[7] *Аметов Р. В., Гедике А. И., Янковская А. Е.* Интеллектуальное инструментальное средство ИМСЛОГ (версия 2004 года) // Интеллектуальные системы (AIS'05), Интеллектуальные САПР (CAD-2005). IX Национальная конф. по искусственному интеллекту с межд. участием (КИИ-2004). Сб. науч. тр. Т. 2. М.: Физматлит, 2004.

Содержание

Алфавитный указатель

Е

Ж

З

И

К

Л

М

Н

О

П

Р

С

Т

Научное издание

Математические методы
распознавания образов

Сборник докладов
13-й Всероссийской конференции,
посвящённой 15-летию РФФИ

Напечатано с готового оригинал-макета

Подписано в печать 29.08.2007 г.
Формат 60×901/16. Печать офсетная.
Бумага офсетная № 1. Печ. л. 41,75. Тираж 300 экз. Заказ № 6658.
Издательство ООО «МАКС Пресс»
Лицензия ИД №00510 от 01.12.1999

119992, ГСП-2, Москва, Ленинские горы, МГУ им. М. В. Ломоносова,
2-й учебный корпус, 627 к.
Тел. 939-3890, 393-3891, Тел./Факс. 939-3891.