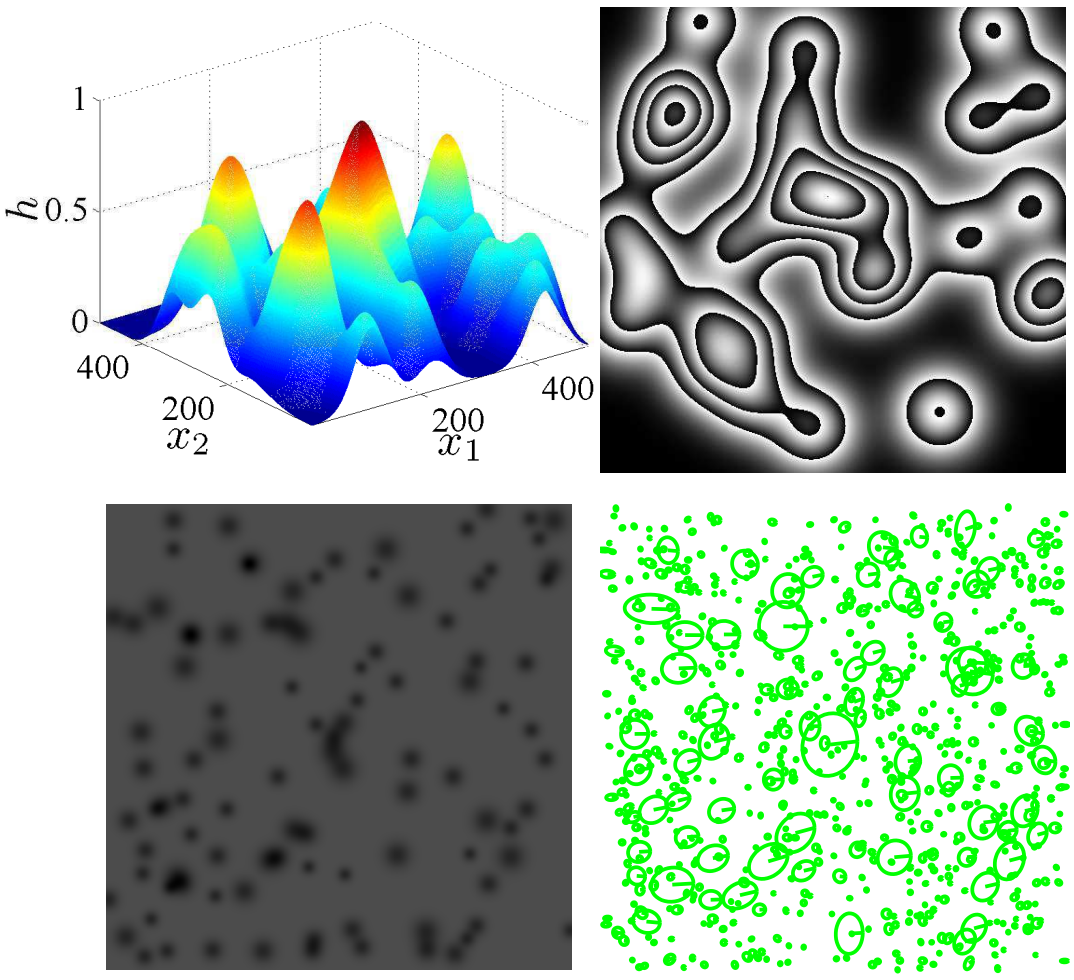
ISSN 2223-3792

Машинное обучение и анализ данных

Май — Сентябрь 2013 **Том 1, номер 5**



Машинное обучение и анализ данных

Journal of Machine Learning and Data Analysis

Журнал «Машинное обучение и анализ данных» публикует новые теоретические и обзорные статьи с результатами научных исследований в области теоретических основ информатики и её приложений. Цель журнала — развитие теории машинного обучения, интеллектуального анализа данных и методов проведения вычислительных экспериментов. Принимаются статьи на английском и русском языках.

Журнал включен в российский индекс научного цитирования РИНЦ. Информация о цитировании статей находится на сайте Российского индекса научного цитирования. ISSN 2223-3792, номер свидетельства о регистрации ЭЛ № ФС 77-55486.

- Архив журнала http://www.ccas.ru/jmlda/
- Новостной сайт http://jmlda.org/
- Электронная система подачи статей http://jmlda.org/papers/

Тематика журнала:
— классификация, кластеризация, регрессионный анализ,
— алгебраический подход к проблеме синтеза корректных алгоритмов,
— многомерный статистический анализ,
— выбор моделей и сложность,
— предсказательное моделирование,
— статистическая теория обучения,
— методы прогнозирования временных рядов,
— методы обработки и распознавания сигналов,
— методы оптимизации в задачах машинного обучения и анализа данных,
— методы визуализации данных,
— обработка и распознавание речи и изображений,
— анализ и понимание текста,
— информационный поиск,
— прикладные задачи анализа данных.

Вычислительный центр Российской академии наук
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Москва, 2013

Содержание

Обнаружение движения устойчивых отражателей по серии спутниковых радиолокационных снимков земной поверхности*

Василейский А. С.[1], ***Карацуба Е. А.***[2], ***Карелов А. И.***[1], ***Кузнецов М. П.***[3], ***Рейер И. А.***[2]

A.Vasileisky@gismps.ru, karacuba@mi.ras.ru, A.Karelov@gismps.ru, mikhail.kuznecov@phystech.edu, reyer@forecsys.ru

1 — Центр внедрения космических технологий ОАО «НИИАС»

2 — Вычислительный центр им. Дородницына РАН

3 — Московский физико-технический институт

Набор устойчивых отражателей радиолокационного сигнала описан метрической конфигурацией их взаимного расположения и вектором их условных высот. По серии метрических конфигураций требуется определить движение некоторой части устойчивых отражателей относительно всего набора. В работе предложен алгоритм построения серии метрических конфигураций по зашумленным данным со спутниковых снимков земной поверхности и выявления связей между устойчивыми отражателями. Предложен метод обнаружения движения отражателей, исследованы его свойства. Метод проиллюстрирован синтетическими и реальными данными.

Ключевые слова: *радиолокация, синтезированная апертура, SAR-интерферометрия, устойчивые отражатели, метрическая конфигурация.*

The algorithm of persistent scatterers movement detection on the satellite radar images of the Earth surface*

Vasileisky A. S.[1], ***Karatsuba E. A.***[2], ***Karelov A. I.***[1], ***Kuznetsov M. P.***[3], ***Reyer I. A.***[2]

1 — Space Technology Application Center of NIIAS

2 — Dorodnitsyn Computing Centre

3 — Moscow Institute of Physics and Technology

A set of radar signal persistent scatterers is described with a metrics configuration of their mutual location and with their heights vector. Using the set of metrics configuration, the movement of subset of the persistent scatterers concerning the whole set should be detected. An algorithm of the set of metrics configuration construction using noisy satellite images of the Earth surface and an algorithm of the persistent scatterers relations detection have been suggested. A method of the persistent scatterers movement detection has been proposed and its properties have been analyzed. The method is illustrated by the synthetic and real data.

Keywords: *radiolocation, synthetic aperture radar, SAR interferometry, persistent scatterers, metric configuration.*

Введение

Радиолокационная интерферометрия является одним из наиболее эффективных способов дистанционного выявления смещений земной поверхности, вызванных оползнями,

Работа выполнена при финансовой поддержке РФФИ, проект №11-07-13160.

землетрясениями, извержением вулканов, криогенными и карстовыми процессами. Этот метод находит все большее применение для мониторинга состояния протяженных инфраструктурных объектов топливно-энергетического и транспортного комплекса, например, объектов железнодорожной инфраструктуры, расположенных в зонах потенциально опасного воздействия геодинамических процессов.

Данная работа посвящена описанию и развитию методов, позволяющих оценивать смещение объектов с течением времени с использованием двух и более радиолокационных спутниковых снимков одного и того же участка земной поверхности.

Радиолокаторы с синтезированной апертурой (synthetic aperture radar — SAR) являются разновидностью активной аппаратуры дистанционного зондирования Земли [1, 2, 3, 4, 5], позволяющей осуществлять детальную съемку земной поверхности (с пространственным разрешением до 1 м) [6], с использованием относительно небольших антенн на борту космических аппаратов. Детальность изображений достигается путем искусственного увеличения эффективного размера бортовой антенны спутника при его орбитальном движении.

Аппаратура SAR-систем регистрирует амплитуду и фазу отраженного земной поверхностью радиолокационного сигнала. Принцип интерферометрии заключается в восстановлении цифровой модели поверхности путем анализа фазовых компонент двух или более SAR-снимков, сделанных из различных близко расположенных точек орбиты. Дифференциальная интерферометрия применяется для получения информации о малых изменениях ландшафта.

Ранее в работах [2, 3, 4, 7, 8, 9] были исследованы методы обработки фазовых компонент радиолокационных изображений для решения задачи интерферометрии. В основном земная поверхность представляет собой диэлектрическую среду со значительной вариацией характеристик, которая сильно зашумляет фазу сигнала при отражении, и простое сравнение фазовых компонент двух радиолокационных изображений является неэффективным методом для оценки смещений поверхности.

Для решения этой проблемы было разработано понятие устойчивых отражателей [2, 3] — объектов земной поверхности, для которых амплитудно-фазовые характеристики радиолокационного сигнала при отражении незначительно меняются со временем. Такими объектами могут являться металлические конструкции, здания и сооружения, а также открытые участки местности со слабо выраженным растительным покровом. На каждом из SAR-снимков одного и того же участка земной поверхности выделяется набор устойчивых отражателей и исследуется изменение фазы отраженного сигнала для каждого отражателя. По этому изменению фазы определяется сдвиг соответствующего участка земной поверхности.

Были предложены методы выделения устойчивых отражателей, основывающиеся на вероятностной модели отраженной волны [8, 10]. Вероятностная модель основывается на том, что волна, принятая радаром, описывается функцией, которая состоит из действительной и мнимой части. Каждая из этих частей зашумлена гауссианой с определенным значением дисперсии и скрыта для наблюдения. В каждой точке радар регистрирует амплитуду и фазу волны. Амплитуда принятой волны описывается нелинейной комбинацией действительной и мнимой части принятой волны (а именно, корнем из суммы их квадратов) и имеет распределения Релея [10]. Фаза также описывается нелинейной комбинацией действительной и мнимой части, и ее распределение записывается в виде интеграла ошибок. Таким образом, амплитуда и фаза зарегистрированного радиолокационного изобра-

жения в каждом пикселе являются зависимыми случайными величинами, зависящими от дисперсии шума.

Важным свойством данной вероятностной модели является то, что дисперсия фазы приблизительно совпадает с отношением дисперсии амплитуды к среднему. На этом факте основывается метод выделения устойчивых отражателей, описанный в [8]. Имея последовательность SAR-снимков одного и того же участка земной поверхности, для каждого пиксела изображения определяется последовательность значений амплитуды и фазы, т. е., выборка значений амплитуды и фазы. Устойчивыми отражателями являются точки изображения с большим средним и маленькой дисперсией амплитуды и небольшим разбросом фазы. Причем разброс фазы является функцией от дисперсии амплитуды, и не возникает необходимости работать с сильно зашумленным фазовым изображением для подсчета дисперсии фазы. Для определения устойчивого отражателя в каждом пикселе подсчитывается дисперсия амплитуды и осуществляется пороговое отсечение.

Также был предложен метод, основывающийся на описанной вероятностной модели и на определении устойчивого отражателя как пиксела с наибольшим значением SCR (signal-to-clutter ratio) [10]. По определению, SCR является отношением энергии волны в пикселе изображения к суммарной энергии соседних пикселов. Одним из методов оценки SCR является метод максимального правдоподобия, в котором функция распределения фазы зависит от параметра SCR. Правдоподобие максимизируется по параметру SCR для выборки значений фазы каждого пиксела изображения. В качестве устойчивых отражателей выбираются пикселы изображения с наибольшим значением SCR.

Перечисленные методы имеют ряд недостатков. Один из них заключается в том, что необходимо выполнять предобработку данных для устранения артефактов, вызванных влиянием атмосферы или перемещениями земной поверхности. После определения положения устойчивых отражателей для оценки их смещений необходимо учитывать относительные высоты отражателей. Кроме того, предложенные методы более устойчиво работают в местах сгущения отражателей на изображении и менее эффективно выделяют изолированные отражатели.

Для устранения перечисленных недостатков были предложены методы выявления устойчивых отражателей на последовательностях радиолокационных снимков на основе анализа графов [3]. Главным отличием этих методов от вышеперечисленных является то, что для пары снимков рассматриваются не отдельные пикселы или участки изображений, а их пары. Вводится понятие оценки когерентности каждой пары пикселов или отражателей. Каждой паре отражателей ставится в соответствие дуга, имеющая определенный вес в смысле оценки когерентности. Из всех дуг выбираются те, которые имеют наибольший вес. Пары отражателей, образующие дугу с наибольшим весом, являются устойчивыми отражателями. Причем оценка когерентности рассчитывается путем максимизации параметров, характеризующих относительную высоту и скорость смещения отражателей. При решении глобальной оптимизационной задачи в качестве решения вычисляются положения устойчивых отражателей и их смещения. Заметим, что этот метод является более вычислительно сложным за счет перебора всех пар пикселов. Предложены методы сокращения перебора, основанные на совместном анализе пикселов только в пределах заданной окрестности.

В данной работе предлагается метод построения системы устойчивых отражателей, дополняющий вышеуказанные методы и устраняющий некоторые их недостатки. Предлагается следующий трехшаговый алгоритм. На первом шаге устойчивые отражатели выделяются отдельно на амплитудных компонентах каждого из SAR-изображений последова-

тельности. При этом в качестве устойчивых отражателей выделяются однородные светлые области (в дальнейшем — блобы) на амплитудной составляющей. Для поиска блобов предложено использовать LOG-детектор [11, 12].

На втором шаге алгоритм сопоставляет найденные на разных снимках последовательности системы отражателей. Второй шаг логически разделяется на два этапа. На первом этапе для каждой последовательной пары снимков сопоставляются друг с другом соответствующие им системы отражателей. При этом используются идеи известного на практике подхода SIFT [13]. Рассматривается многомерное пространство отражателей одной картинки (отражатель задается градиентным признаковым описанием), и для каждого отражателя второй картинки отыскивается ближайший к нему отражатель из построенного пространства. При этом алгоритм является неустойчивым к выбросам, то есть находит много избыточных связей между отражателями. Для решения этой проблемы на втором этапе выполняется построение метрических конфигураций систем отражателей [14, 15], и из построенных связей между отражателями убираются те, которые наихудшим образом влияют на качество сопоставления метрических конфигураций. Данный подход позволяет решить проблему возможного относительного сдвига отражателей и смещения изображений.

На третьем шаге рассчитываются относительные скорости смещения устойчивых отражателей. При этом используются идеи, описанные в [3]. Таким образом, отличие предлагаемого подхода от приведенного выше графового алгоритма состоит в поиске системы отражателей на первом и втором шаге. Работоспособность алгоритма проиллюстрирована на наборе синтетических данных.

Постановка задачи обнаружения устойчивых отражателей

Математическая модель набора SAR-изображений. Введем необходимые обозначения. Рассмотрим изображения — матрицы размера $m \times n$ яркостей (соответствующих амплитудной компоненте) $\mathbf{Z}$, высот $\mathbf{H}$ и фаз $\mathbf{\Phi}$. Элементы z_{ij}, h_{ij}, φ_{ij} этих матриц принадлежат множеству $\mathbb{R}^1_+$.

Целочисленные величины — индексы i, j матриц поставлены в соответствие точкам x, y отрезков $\mathcal{X}, \mathcal{Y}$:

$$x = x(i), \quad y = y(j), \quad i \in \{1, ..., m\}, \quad j \in \{1, ..., n\}.$$

Соответствие задано таким образом, что величины x, y измеряются в метрах и интерпретируются в рамках данной задачи как координаты (x, y) некоторой картографической проекции, соответствующей изображениям $\mathbf{Z}, \mathbf{H}, \mathbf{\Phi}$.

Изображению, задаваемому матрицами $\mathbf{Z}, \mathbf{H}, \mathbf{\Phi}$, поставлены в соответствие функции яркости, высоты и фазы $z, h, \varphi : \mathbb{R}^2 \to \mathbb{R}^1_+$.

При этом в виде измерений задан набор матриц $\{\mathbf{\Phi}_i\}_{i=1}^k$, $\{\mathbf{Z}_i\}_{i=1}^k$, где i-я матрица соответствует радиолокационному снимку номер i. Набор матриц $\{\mathbf{H}_i\}_{i=1}^k$ скрыт для наблюдения.

Задача поиска устойчивых отражателей. Рассмотрим матрицу $\mathbf{\Sigma} = \|\sigma^2(i, j)\|$ размеров $m \times n$, описывающую расположение устойчивых отражателей на изображениях. Этой матрице поставлена в соответствие функция $\sigma^2 : \mathbb{R}^2 \to \mathbb{R}^1_+$. Эта матрица является единственной для всего набора снимков. Она является скрытой для наблюдения. Элемент $\mathbf{\Sigma}(i, j)$ в каждой точке является дисперсией отраженной волны. Будем считать, что меньшее значение дисперсии $\sigma^2(x, y)$ означает меньшее воздействие шумов различного вида на

волну, отраженную участком поверхности, соответствующим точке (x, y) и, как следствие, большую достоверность нахождения устойчивого отражателя в точке (x, y).

Предполагается, что функции $\{\varphi_i(x, y), z_i(x, y)\}_{i=1}^k$, задающие набор изображений, зависят от функции дисперсий $\sigma^2(x, y)$ следующим образом. Если в некоторой точке (x', y') значение дисперсии мало:

$$\sigma^2(x', y') < \varepsilon,$$

то разброс величин $\{\varphi_i(x', y')\}_{i=1}^k$ и $\{z_i(x', y')\}_{i=1}^k$ в точке (x', y') является небольшим по i, а среднее значение яркости $E_i z_i(x', y') > \delta$ в точке (x', y') является большим некоторой наперед заданной величины δ. Точку (x', y') такого вида будем считать устойчивым отражателем.

Исходя из этого, зададим множество устойчивых отражателей R_ε в качестве точек с маленьким значением дисперсии отраженной волны:

$$R_\varepsilon = \{(x, y) | \sigma(x, y) < \varepsilon\}.$$

Задача поиска устойчивых отражателей ставится следующим образом. Будем исходить из того, что отражатели являются точками с высоким значением амплитуды $E_i z_i(x', y') > \delta$, и искать их в виде светлых пятен (односвязных областей, в которые можно вписать круг радиусом ξ). Имея набор матриц $\{\mathbf{Z}_i\}_{i=1}^k$, требуется найти оценку матрицы $\mathbf{\Sigma}$ и описать множество устойчивых отражателей R_ε.

Задача восстановления сдвига земной поверхности. Предполагается, что на скрытый набор матриц высот $\{\mathbf{H}_i\}_{i=1}^k$ действует некоторое отображение η, задающее сдвиг земной поверхности. Будем считать, что это отображение задано на дискретном множестве точек:

$$\eta(\mathbf{H}_{i-1}) = \mathbf{H}_i, \quad i = 1, ..., k.$$

По множеству найденных отражателей R_ε и набору матриц $\{\mathbf{\Phi}_i\}_{i=1}^k$ требуется сделать оценку отображения $\hat{\eta}$ в точках, соответствующих устойчивым отражателям, оценив таким образом относительный сдвиг земной поверхности.

Базовый алгоритм обнаружения устойчивых отражателей

Алгоритм обнаружения устойчивых отражателей на основе оценок когерентности между парами точек на двух и более изображениях был предложен в [3, 5]. Отметим, что алгоритм использует матрицы амплитуд $\{\mathbf{Z}_i\}_{i=1}^k$ только для начального приближения, оперируя, в основном, только матрицами фаз $\{\mathbf{\Phi}_i\}_{i=1}^k$.

Рассматривается k SAR-изображений одного и того же участка земной поверхности. Среди этих изображений выбирается мастер-изображение. Эффективные методы выбора мастер-изображения описаны в [16, 17]. Для каждого i-го изображения и мастер-изображения рассматриваются все пары точек-«кандидатов» в устойчивые отражатели. Для i-го изображения рассчитываются величины разности фаз $\delta\varphi_{a,i}$ для всех пар точек между мастером и i-м изображением, где a обозначает дугу, связывающую соответствующие пары точек (таким образом, всего дуг — $N(N-1)$ для $N = m \times n$ точек на изображении).

Величина разности фаз $\delta\varphi_{a,i}$, зависящая от пары точек a, моделируется следующим образом:

$$\delta\varphi_{a,i} = \frac{4\pi}{\lambda}[T_i \delta\nu_a + \alpha B_i \delta h_a + \varepsilon_{a,i}]_{2\pi}.$$

В этой формуле известными параметрами являются длина волны λ, B_i и T_i — пространственная и временная база интерферометрической пары радиолокационных снимков

[1, 8, 7], которые соответствуют расстоянию и времени между двумя положениями радара в разные моменты съемки; $\varepsilon_{a,i}$ — шумовая составляющая. Неизвестные параметры δh_a и $\delta\nu_a$ отвечают за относительную высоту и скорость движения двух отражателей, связанных дугой a.

Вводится понятие когерентности γ_a, ассоциированной с дугой a:

$$\gamma_a = \max_{\delta h_a, \delta\nu_a} |\sum_{i=1}^{k} w_{a,i} e^{j\varepsilon_{a,i}}|,$$

где j — комплексная единица. Чем больше значение параметра γ_a, тем выше значение когерентности между двумя точками на изображении, и для больших значений γ_a пара точек, связанных дугой a, объявляется парой устойчивых отражателей. При этом максимизация величины γ_a производится по параметрам δh_a и $\delta\nu_a$, таким образом, при решении задачи оптимизации сразу вычисляется оценка всех относительных высот и скоростей движения отражателей.

Предлагается итеративный алгоритм поиска множества P_L устойчивых отражателей. На первом шаге формируется множество P_0 — начальное множество отражателей, являющихся точками с максимальной амплитудой.

За P_l обозначается множество, состоящее из точек и соединяющих их дуг, выбранных на итерации l. На следующей итерации $l+1$ строится множество P_{l+1}, являющееся объединением множества P_l с множеством дуг

$$\gamma_a > \gamma_T,$$

где γ_T — пороговое значение, а оценка когерентности γ_a

$$\gamma_a = \max_{\delta h_a, \delta\nu_a} \Big|\sum_{i=1}^{k} w_{a,i} \exp\left(j\left[\delta\varphi_{a,i} - \frac{4\pi}{\lambda}(T_i\delta\nu_a + \alpha B_i \delta h_a)\right]\right)\Big|.$$

Итеративный процесс останавливается, когда на шаге L не может быть выбрано новых дуг со значениями оценки когерентности больше порогового значения. При этом множество P_L представляет собой множество точек устойчивых отражателей и соединяющих их дуг, причем для каждой дуги вычислены оценки относительной высоты δh_a и относительной скорости движения $\delta\nu_a$ отражателей, использующиеся для определения величины сдвига земной поверхности. По найденным величинам δh_a и $\delta\nu_a$ производится оценка отображения $\hat{\eta}$.

Утверждается [8], что предложенный метод устойчив к атмосферным и другим шумам измерения фазовой составляющей, а также устойчив к неоднородной структуре расположения устойчивых отражателей.

Алгоритм обнаружения устойчивых отражателей в качестве высокоамплитудных пятен и сопоставления систем отражателей на основе анализа метрических конфигураций

Предлагаемый алгоритм состоит из трех последовательных этапов. На первом этапе каждому SAR-изображению последовательности ставится в соответствие система отражателей, найденных в качестве однородных светлых пятен (в дальнейшем — блобов [18]). На втором этапе найденные системы отражателей сопоставляются, и для последовательности снимков строится единая система устойчивых отражателей. Таким образом, на первых

двух этапах решается задача поиска устойчивых отражателей и оценки матрицы $\mathbf{\Sigma}$. На третьем этапе производится вычисление относительных высот и скоростей движения отражателей с использованием идей из [8], и решается задача восстановления отображения $\hat{\eta}$ сдвига земной поверхности.

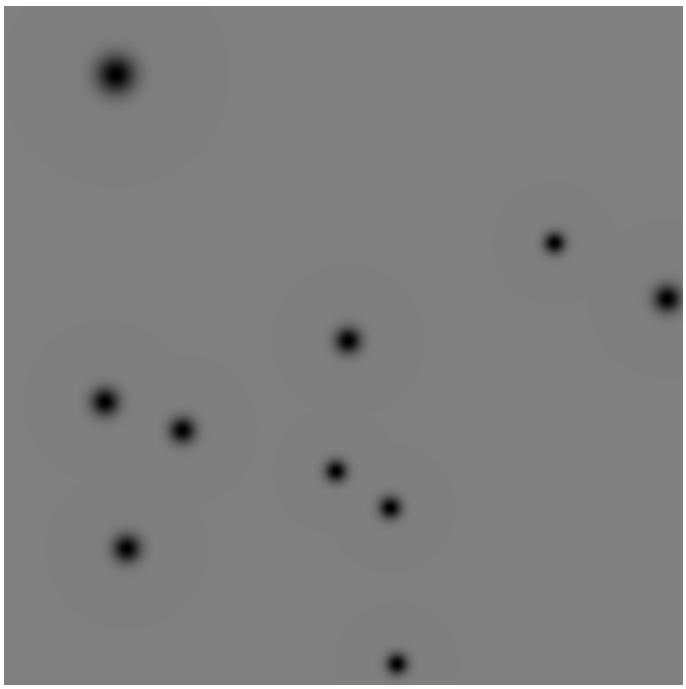
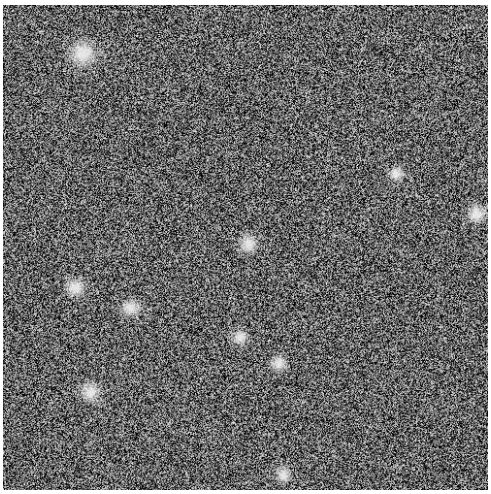


Рис. 1: Модельные данные: поверхность дисперсий

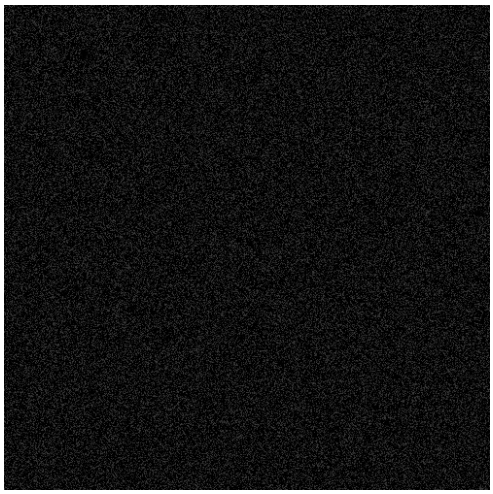
Алгоритм поиска блобов для одного изображения. На первом этапе для каждого изображения i найдем систему отражателей по амплитудной составляющей $\{\mathbf{Z}_i\}$. Будем строить систему отражателей с помощью метода, описанного в [18]. Используется метод нахождения устойчивых отражателей, основанный на принципе LoG-детектора. Этот принцип представляет собой свертку изображения с лапласианом гауссиан разных масштабов и поиск максимумов в полученной серии сглаженных изображений. Производится также процедура уточнения эллипсоидальной формы блобов, основанная на сингулярном разложении матрицы, состоящей из взвешенных значений производной функции интенсивности.

В качестве результата, на этом этапе вычисляются множества $\{R_{\varepsilon,i}\}_{i=1}^{k}$ центров отражателей для каждого изображения i. В дальнейшем описании для сокращения обозначений опустим параметр ε множества отражателей, являющийся некоторой пороговой константой для алгоритма на данном этапе.

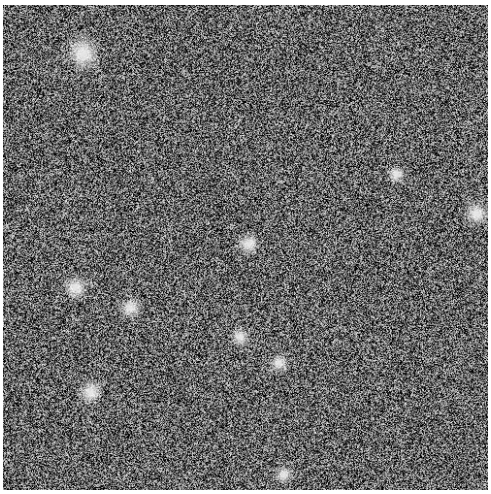
Алгоритм сопоставления систем отражателей. На втором этапе строится система отражателей R, единая для всех изображений последовательности. Идея метода заключается в попарном сравнении систем отражателей на последовательных снимках и сопоставлении найденных центров отражателей.



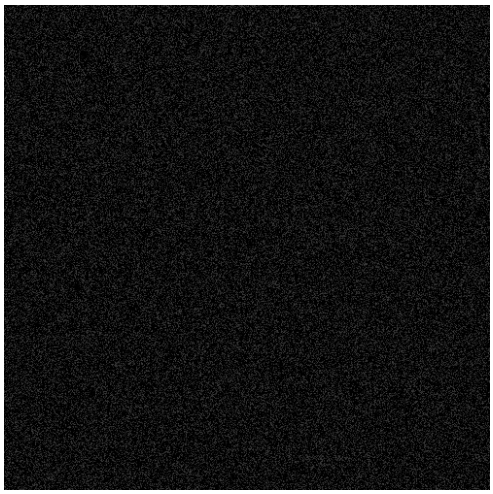
(а) Амплитудное изображение



(б) Фазовое изображение без учета высот



(в) Амплитудное изображение



(г) Фазовое изображение без учета высот

Рис. 2: Модельные данные.

Алгоритм заключается в следующем. Дан набор амплитудных изображений $\{\mathbf{Z}_i\}_{i=1}^k$ и соответствующие им системы отражателей $\{R_i\}_{i=1}^k$. На первом шаге для изображений $\mathbf{Z}_1, \mathbf{Z}_2$ строится система отражателей $R_{\{1,2\}}$, состоящая из тех отражателей, которые прошли процедуру сопоставления систем R_1 и R_2. На шаге i для изображений $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_{i+1}$ строится система отражателей $R_{\{1,...,i+1\}}$, состоящая из отражателей, которые прошли процедуру сопоставления систем $R_{\{1,...,i\}}$ и R_{i+1}. Таким образом, на шаге $k-1$ мы построим систему отражателей $R_{\{1,...,k\}}$, единую для всех изображений $\mathbf{Z}_1, ..., \mathbf{Z}_k$.

Для сопоставления двух систем отражателей R_1 и R_2 предлагается следующий двухшаговый метод. Первый шаг является распространенным в обработке изображений методом сопоставления SIFT дескрипторов [13]. Он состоит в том, что каждому найденному на первом шаге отражателю ставится в соответствие гистограмма градиентов яркости по разным

направлениям. Два отражателя на разных изображениях считаются одной точкой, если в многомерном пространстве гистограмм между этими отражателями мало евклидово расстояние. При выполнении этой процедуры рассчитываются две равномощные системы отражателей R_1^0, R_2^0, где отражатели одного и другого изображения соответствуют друг другу в порядке индексации. При этом в системах R_1^0, R_2^0 может присутствовать большое количество ложных пар отражателей. Это связано с тем, что спутниковый снимок поверхности может содержать большое количество областей, являющихся похожими друг на друга по направлениям градиентов.

На втором шаге выполняется построение множества $R_{\{1,2\}}$ путем удаления лишних пар отражателей из множеств R_1^0, R_2^0. Для этого строятся метрические конфигурации точек $\|r_{ij}^1\|, \|r_{ij}^2\|$ [14, 15] из множеств R_1^0, R_2^0. Для количественной ошибки оценочного связывания [19] вводится переменная δ_{ij},

$$\delta_{ij} = |r_{ij}^1 - r_{ij}^2|.$$

Вектор расстояний, ассоциированный с парой номер i, есть

$$\delta_i = \{\delta_{1,i}, ..., \delta_{i-1,i}, \delta_{i,i+1}, ..., \delta_{i,N}\},$$

где N — мощность множества отражателей.

Пара сопряженных точек принимается, если $\|\delta_i\| < \Delta$ и отклоняется в противоположном случае. Здесь $\|\delta_i\|$ является усредненным значением всех компонент.

Приведенный метод позволяет оставить в множестве только те отражатели, которые являются стабильными в смысле метрической конфигурации, то есть реальные объекты на изображении, инвариантные к преобразованиям сдвига и поворота.

Восстановление сдвига земной поверхности. Располагая множеством отражателей R, рассчитаем относительные высоты и скорости движения отражателей, решив задачу оптимизации базового алгоритма для всех пар устойчивых отражателей:

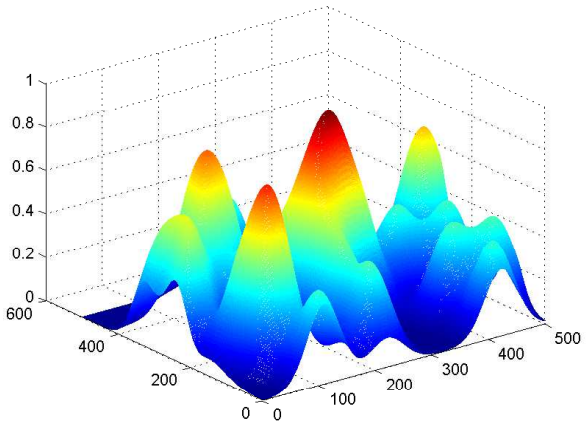
$$\gamma_a = \max_{\delta h_a, \delta \nu_a} \Big| \sum_{i=1}^{k} w_{a,i} \exp\big(j\big[\delta\varphi_{a,i} - \frac{4\pi}{\lambda}(T_i \delta\nu_a + \alpha B_i \delta h_a)\big]\big)\Big|.$$

Отметим, что главным отличием от базового алгоритма является то, что мы используем уже построенные системы устойчивых отражателей, то есть дуги a являются известными. Найденные δh_a и $\delta\nu_a$ относительных высот и скоростей движения отражателей однозначно характеризуют отображение $\hat{\eta}$.

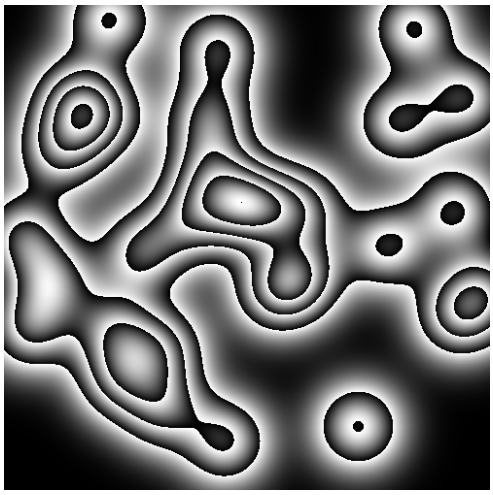
Моделирование данных для SAR-интерферометрии

Для иллюстрации свойств предлагаемого алгоритма выполнена процедура построения набора синтетических данных для SAR-интерферометрии. Нашей задачей является определить функции амплитуды $z(x, y)$, фазы $\varphi(x, y)$ и высоты $h(x, y)$ на конечном наборе точек, задаваемых матрицами $\mathbf{Z}$, $\mathbf{\Phi}$ и $\mathbf{H}$. Важным допущением при генерации данных является то, что функция амплитуды $z(x, y)$ не зависит от высоты $h(x, y)$ соответствующего участка местности. Другими словами, яркость изображения не зависит от высоты съемки. Заметим, что в реальности эти две функции могут быть зависимыми. Мы будем предполагать лишь зависимость фазы $\varphi(x, y)$ от высоты $h(x, y)$; именно эта зависимость и интересует нас при решении задачи интерферометрии.

Разделим алгоритм построения модельных данных на два этапа. На первом этапе будем моделировать набор матриц $\{\mathbf{Z}_i\}$ и $\{\mathbf{\Phi}_i\}$ независимо от высоты, причем каждая матрица

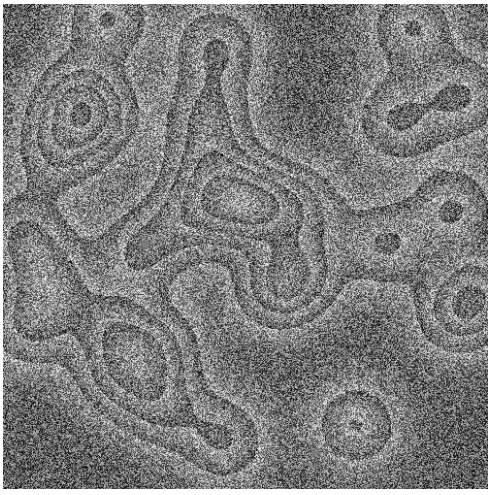


(а) Высота

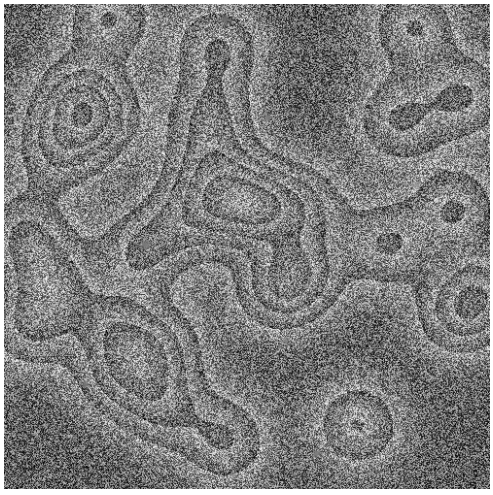


(б) Фазовая составляющая без учета шума

Рис. 3: Модельные данные



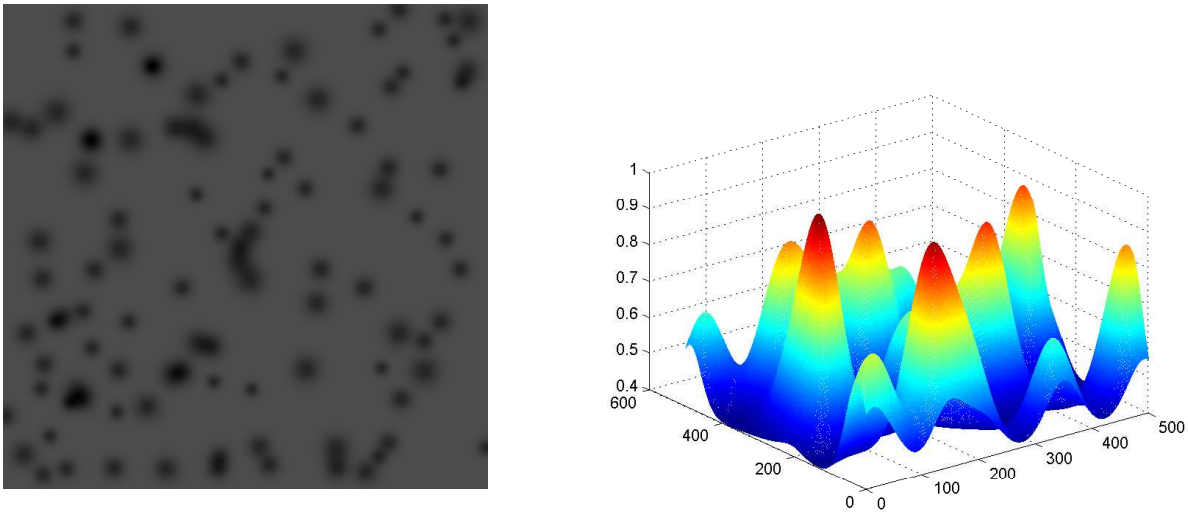
(а) Фазовая составляющая



(б) Фазовая составляющая при сдвинутых высотах

Рис. 4: Модельные данные

i в наборе будет соответствовать i-му SAR-снимку последовательности. На втором этапе сгенерируем набор скрытых матриц высот $\{\mathbf{H}_i\}$, элементы в которых меняются медленно от матрицы к матрице, и подкорректируем соответствующие им матрицы фаз $\{\mathbf{\Phi}_i\}$. Нашей задачей будет являться восстановление изменений матриц высот $\{\mathbf{H}_i\}$ при наблюдении матриц амплитуды $\{\mathbf{Z}_i\}$ и фазы $\{\mathbf{\Phi}_i\}$.



(а) Поверхность дисперсий (б) Поверхность высот

Рис. 5: Экспериментальные данные

Моделирование амплитудной и фазовой составляющей. Для того чтобы смоделировать отраженную волну, воспользуемся вероятностной моделью сигнала [10]. Будем исходить из следующих вероятностных предположений. Действительная $R(i,j)$ и мнимая $I(i,j)$ части функции отраженной волны являются случайными величинами, распределенными по нормальному закону:

$$R(i,j) \sim \mathcal{N}(\mu_R(i,j), \sigma^2(i,j)), \quad I(i,j) \sim \mathcal{N}(\mu_I(i,j), \sigma^2(i,j)).$$

Отметим, что в данной вероятностной модели величина дисперсии $\sigma^2(i,j)$ одинакова для действительной и мнимой части.

Амплитуда отраженной волны описывается функцией от действительной и мнимой части:

$$z(i,j) = \sqrt{R^2(i,j) + I^2(i,j)}$$

и имеет распределение Райса

$$f(z(i,j)|\nu(i,j), \sigma(i,j)) = \frac{x}{\sigma^2(i,j)} \exp\big(-\frac{(x^2+\nu^2(i,j))}{2\sigma^2}\big) J_0\big(\frac{x\nu(i,j)}{\sigma^2(i,j)}\big),$$

где

$$\nu(i,j) = \sqrt{\mu_R^2(i,j) + \mu_I^2(i,j)},$$

а J_0 — функция Бесселя. Распределение фазы $\varphi(i,j)$ может быть описано в терминах интеграла ошибок [10] и опущено из-за громоздкости.

Отметим, что величина $\sigma(i,j)$ отвечает за отражающее свойство земной поверхности в точке (i,j): при маленьких значениях $\sigma(i,j)$ на последовательности SAR-снимков в точке (i,j) наблюдается постоянство фазы и амплитуды, и точка (i,j) считается устойчивым отражателем. Исходя из этого, на первом шаге необходимо смоделировать величину дисперсии $\sigma(i,j)$ во всех точках $m \times n$, то есть построить матрицу $\mathbf{\Sigma} = \|\sigma(i,j)\|$. Будем делать это следующим образом: во всех точках (i,j) примем значение $\sigma(i,j) = \text{const}$ кроме разреженного множества точек, являющихся устойчивыми отражателями. В точках устойчивых отражателей примем $\sigma(i,j) = 0$ и сделаем эти точки центрами гауссианов, чтобы получить сглаженное изображение отражателей.

Пример такого построения показан на рис. 1. Черными пятнами обозначены устойчивые отражатели: значение $\sigma(i,j)$ в центрах этих пятен равно нулю и гладко размывается гауссианой до значения $\sigma(i,j) = 0.5$, что соответствует цвету фона рисунка моделируемого распределения дисперсий. Смоделируем матрицу высот $\mathbf{H}$ набором гауссиан. Пример такой поверхности показан на рис. 3а.

Исходя из имеющегося распределения дисперсий, сгенерируем действительную и мнимую части отраженной волны (в простейшем случае, сделаем их константными), зашумив их гауссианами с нулевым средним и соответствующей дисперсией $\sigma^2(i,j)$. Полученное модельное распределение дисперсий должно использоваться при генерации действительной и мнимой частей отраженной волны для всех SAR-снимков последовательности.

Пример двух таких реализаций приведен на рис. 2. Слева показаны сгенерированные амплитудные компоненты изображений, на которых светлые пятна являются устойчивыми отражателями; справа — фазовые компоненты. Отметим, что на этом шаге никак не учитываются высоты поверхности, существенно влияющие на фазовые компоненты изображений. Моделирование высот и корректировку фазы рассмотрим в следующем параграфе.

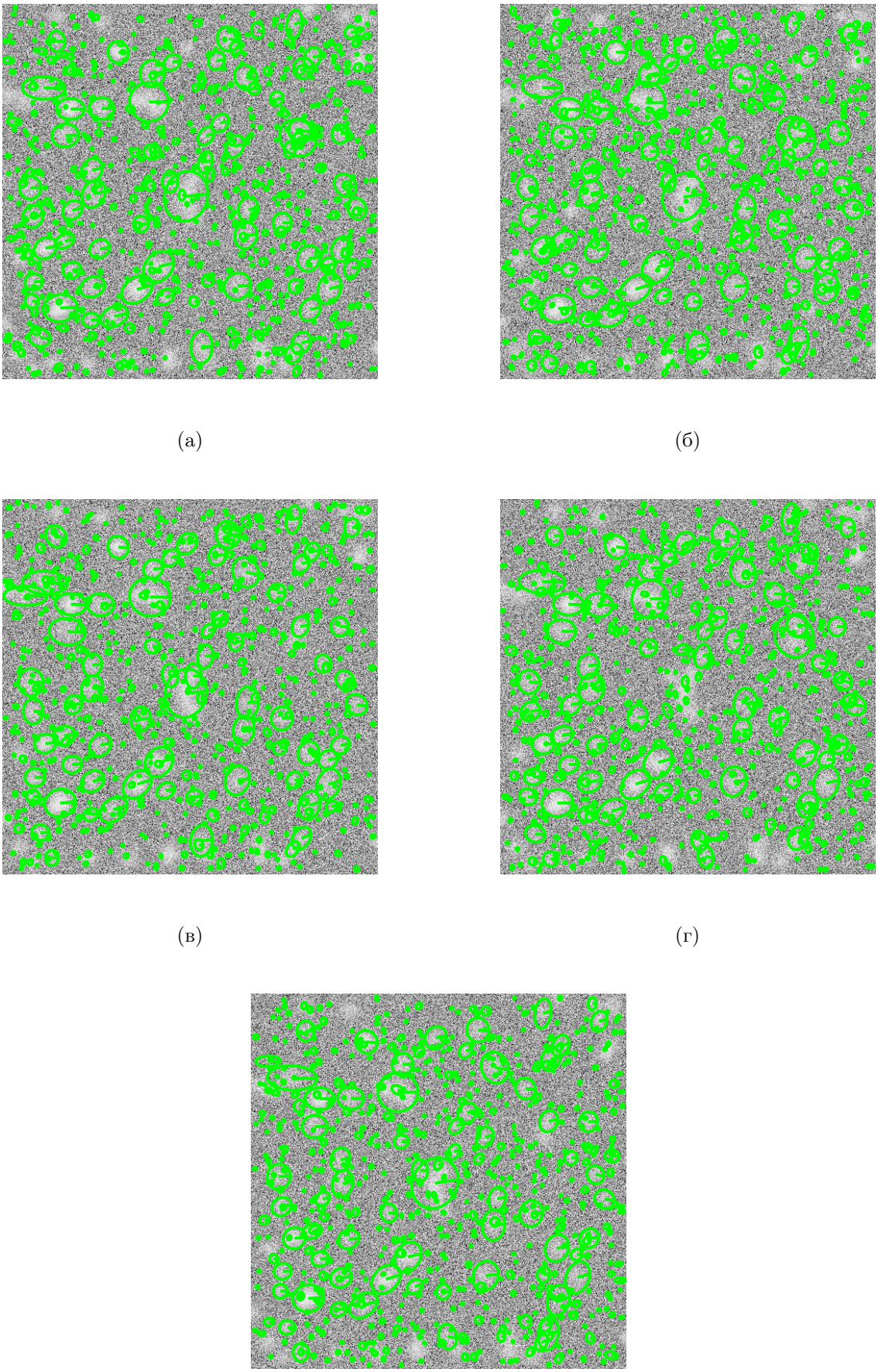
Моделирование высот. Дальнейшее моделирование данных будем проводить, учитывая следующие соображения. Функцию фазы $\varphi(x,y)$ будем считать суперпозицией

$$\varphi(x,y) = \varphi_0(x,y)\varphi_h(x,y),$$

где $\varphi_0(x,y)$ — фаза отраженной волны, смоделированная нами на предыдущем шаге. Эта составляющая зависит только от случайного шума, накладывающегося на действительную и мнимую части волны, и не зависит от высоты. Составляющая $\varphi_h(x,y)$ зависит от высоты соответствующего участка поверхности с учетом его смещений во времени. Пример этой составляющей фазы $\varphi_h(x,y)$ для построенной матрицы высот $\mathbf{H}$ продемонстрирован на рис. 3б.

Для генерации фазовых компонент будем изменять матрицу высот $\mathbf{H}$ согласно выбранному закону. Например, в простейшем случае будем уменьшать каждый элемент $h(i,j)$ на постоянную величину смещения поверхности за время между получением снимков. Такой закон соответствует равномерному уменьшению высот земной поверхности. При этом будет меняться составляющая $\varphi_{hi}(x,y)$ фазовой компоненты изображений, а соответственно, и вся суперпозиция $\varphi_i(x,y)$.

На рис. 4 показаны фазовые компоненты модельных изображений с учетом изменения высоты и наложения шума. На рис. 4а приведена фазовая компонента, соответствующая высотам на рис. 3а и амплитудной составляющей на рис. 2а. На рис. 4б приведена фазовая компонента, соответствующая изменившимся вследствие смещений поверхности высотам и амплитудной компоненте на рис. 2в. Отметим, что на реальных снимках фазовая состав-



(а) (б)

(в) (г)

(д)

Рис. 6: Последовательность амплитудных изображений с выделенными отражателями

ляющая зашумлена гораздо сильнее, и визуально практические невозможно распознать изменение фазы, соответствующее рельефу поверхности.

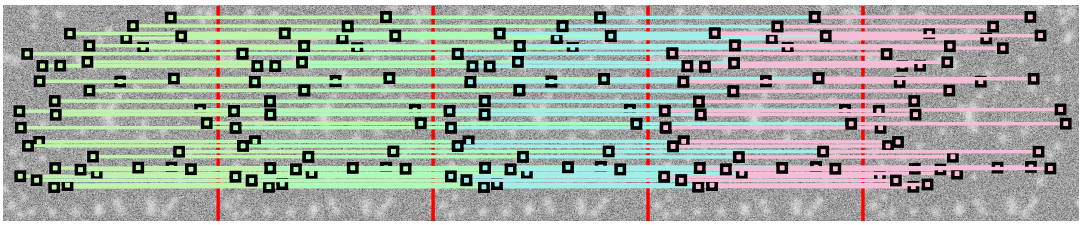


Рис. 7: Поиск системы отражателей

Итак, при моделировании данных мы получаем последовательность матриц высот $\mathbf{H}_i$, плавно изменяющихся во времени в соответствии с заданным законом, соответствующим сдвигу земной поверхности. Для каждой матрицы $\mathbf{H}_i$ мы строим матрицы $\mathbf{Z}_i$ амплитуды и $\mathbf{\Phi}_i$ фазы. Причем матрицы яркости зависят только от заранее выбранной системы устойчивых отражателей, положение которых описывается матрицей $\mathbf{\Sigma}$, а матрицы фазы зависят и от положения устойчивых отражателей, и от соответствующих значений высоты.

Вычислительный эксперимент

Для демонстрации работоспособности разработанного алгоритма проведен вычислительный эксперимент по обработке смоделированной последовательности радиолокационных снимков. В качестве исходных данных, была сгенерирована поверхность дисперсий, показанная на рис. 5а, характеризующая расположение устойчивых отражателей, а также поверхность высот, показанная на рис. 5б. Помимо этого, был сгенерирован набор из пяти SAR-изображений, состоящих из зашумленных амплитудных и фазовых компонент. При генерации фазовых компонент использовалась последовательность высотных изображений, причем каждое следующее высотное изображение формировалось из предыдущего путем преобразования линейного сдвига высот:

$$H_{k+1} = \eta(H_k) = \|h_{ij} + \big(\lfloor\frac{m}{2}\rfloor - i\big)\mu\|,$$

где μ — параметр, определяющий величину сдвига.

Для каждого из изображений алгоритм выделил свою систему устойчивых отражателей на рис. 6. На рис. 7 проиллюстрирована работа алгоритма поиска единой системы отражателей. Красные вертикальные линии показывают границы изображений. Горизонтальные разноцветные линии демонстрируют связи между соответствующими устойчивыми отражателями. Поскольку смещения отражателей в плоскости изображения от снимка к снимку не происходило, горизонтальность соединяющих линий на рисунке свидетельствует о правильности обнаружения пар соответствующих отражателей.

На рис. 8 показано обнаруженное алгоритмом движение отражателей. Отражатели, для которых зафиксировано положительное смещение относительно всей сцены, соответствующее увеличению высоты, выделены красным цветом. Устойчивые отражатели, для которых зафиксировано отрицательное относительное движение, выделены синим цветом. Черным цветом выделены устойчивые отражатели, оставшиеся неподвижными относительно всей сцены. Полученный результат свидетельствует о работоспособности алгоритма

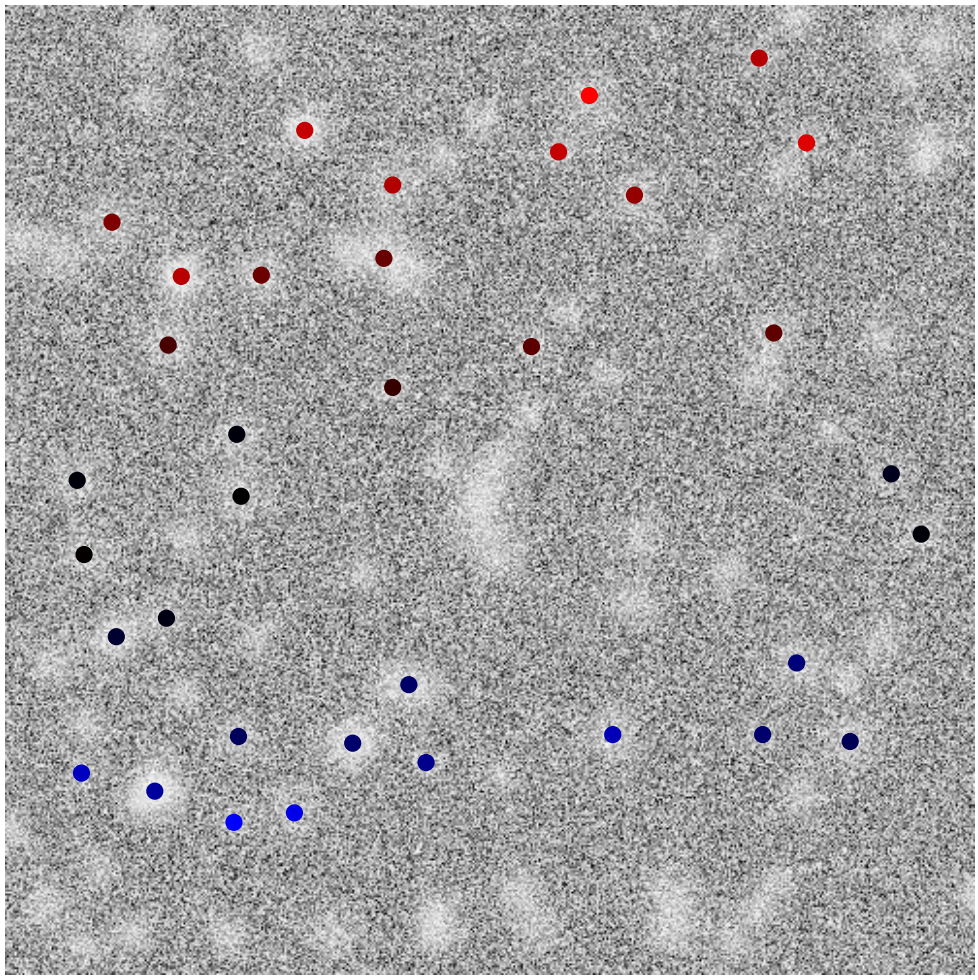


Рис. 8: Движение отражателей

и возможности с некоторой точностью восстановить первоначально заданные относительные линейные сдвиги высот.

Заключение

Исследована задача обнаружения движения системы устойчивых отражателей на последовательности SAR-изображений земной поверхности. Предложен алгоритм описания системы отражателей, основанный на построении метрической конфигурации отражателей. Исследуется изменение метрической конфигурации с течением времени. Предложен способ измерения относительного движения отражателей, основанный на изменении метрической конфигурации. Работоспособность алгоритма проиллюстрирована на примере обработки синтетического набора данных, метод генерации которых основывается на вероятностной модели отраженной волны.

Литература

[1] *Harger R. O.* Synthetic aperture radar fundamental and image processing // *EARSeL Advances in Remote Sensing*. 1993. Vol. 2. Pp. 268–286.

[2] *Hartl P., Bamler R.* Synthetic aperture radar interferometry // *Inverse Problems*. 1998. Vol. 14. Pp. R1–R54.

[3] *Quagliarini A., Schiavon G., Costantini M., Minati F.* A new method for baseline calibration in SAR interferometry // *Proceedings of the FRINGE 2003 Workshop (ESA SP-550)*. 2003.

[4] *Mallorqui J. J., Berardion P., Sansosti E., Lanari R., Mora O.* A small-baseline approach for investigating deformations on full-resolution differential SAR interferograms // *IEEE Trans. Geosci. Remote Sensing*. 2004. Vol. 42. Pp. 1377–1386.

[5] *Costantini M.* A new method for identification and analysis of persistent scatterers in series of SAR images // *IEEE Geoscience and Remote Sensing Symposium (International) (IGARSS 2008)*. 2008.

[6] Italian Space Agency. *COSMO-SkyMed SAR Products Handbook* // http://www.e-geos.it/products/pdf/csk-product

[7] *Rocca F., Ferretti A., Prati C.* Nonlinear subsidence rate estimation using permanent scatterers in differential SAR interferometry // *IEEE Transactions on Geoscience and Remote Sensing*. 2000. Vol. 38. Pp. 2202–2212.

[8] *Rocca F., Ferretti A., Prati C.* Permanent scatterers in SAR interferometry // *IEEE Transactions on Geoscience and Remote Sensing*. 2001. Vol. 39. Pp. 8–20.

[9] *Segall P., Kampes B., Hooper A., Zebker H.* A new method for measuring deformation on volcanoes and other natural terrains using inSAR persistent scatterers // *Geophysical Redearch Letters*. 2004. Vol. 31.

[10] *Agram P. S. Persistent Scatterer Interferometry In Natural Terrain* // PhD Thesis. Stanford University, 2012.

[11] *Lindeberg T.* Scale-space theory: A basic tool for analysing structures at different scales // *Journal of Applied Statistics*. 1994. Vol. 21, no. 2. Pp. 224–270.

[12] *Schmid C., Mikolajczyk K.* Scale & affine invariant interest point detectors // *International Journal of Computer Vision*. 2004. Vol. 60. Pp. 63–86.

[13] *Lowe D. G.* Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*. 2004.

[14] *Майсурадзе А. В. О Специальных представлениях метрических конфигураций* // Диссертация. Вычислительный центр им. А. А. Дородницына РАН. 2005.

[15] *Леонтьев В. К.* О мерах сходства и расстояниях между объектами // *Журнал вычислительной математики и математической физики*. 2009. Т. 49. С. 2041–2058.

[16] *Yihua Liu, Xiaofan Li, Liang Gao, Hua Zhang, Qiming Zeng.* The optimum selection of common master image for series of differential SAR processing to estimate long and slow ground deformation // *IEEE Geoscience and Remote Sensing Symposium (International) (IGARSS'05) Proceedings*. 2005. Vol. 7. Pp. 4586–4589.

[17] *Bethel J., Li Zh.* Image coregistration in SAR interferometry // *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. 2008. Vol. XXXVII. Pp. 433–438.

[18] *Василейский А. С., Карацуба Е. А., Карелов А. И., Кузнецов М. П., Рейер И. А.* Алгоритм выделения устойчивых отражателей на спутниковых снимках земной поверхности // *Машинное обучение и анализ данных*. 2012. Т. 1. С. 452–463.

[19] *Бондаренко А. В., Ососков М. В., Моржин А. В., Визильтер Ю. В., Желтов С. Ю.* // *Обработка и анализ изображений в задачах машинного зрения*. М.: Физматкнига, 2010.

Построение кросс-корреляционных зависимостей при прогнозе загруженности железнодорожного узла*

Вальков А. С.,[1] ***Кожанов Е. М.,***[2] ***Мотренко А. П.,***[3] ***Хусаинов Ф. И.***[4]
valkov@forecsys.ru, vinger4@gmail.com, anastasia.motrenko@gmail.com, f-husainov@yandex.ru
1 — Вычислительный центр РАН, 2 — Московский государственный технический университет имени Н. Э. Баумана, 3 — Московский физико-технический институт, 4 — Российская открытая академия транспорта Московского государственного университета путей сообщения (МИИТ)

Рассматривается проблема обнаружения причинно-следственных связей в разнородных временных рядах. Предлагается прогностическая модель, использующая выявленные связи. Модель предназначена для прогнозирования загруженности железнодорожного узла. Модель использует как исторические данные о загруженности, так и внешние данные: биржевые цены на основные инструменты и нормативные документы. При построении модели используются экспертные высказывания относительно вида связей. Предложен метод оценки достоверности экспертных высказываний. Метод проиллюстрирован данными грузовых перевозок РЖД.

Ключевые слова: *временные ряды, прогнозирование, загрузка железнодорожного узла, прогностическая модель, экспертное высказывание.*

Constructing a cross-correlation model to forecast the utilization of a railway junction station*

Valkov A. S.,[1] ***Kozhanov E. M.,***[2] ***Motrenko A. P.,***[3] ***Husainov F. I.***[4]
1 — Computing Center of the Russian Academy of Sciences; 2 — Bauman Moscow State Technical University; 3 — Moscow Institute of Physics and Technology; 4 — Moscow State University of Railway Engineering

The problem of detecting causal relationships between time series is studied. The authors propose a forecasting model that considers detected relationships. The model is aimed to forecast the utilization of a railway junction station. The model relies on the history of a junction station utilization as well as on the time series for the main financial instruments and regulations. Expert's assessments are used to construct the model. A method that evaluates plausibility of the expert's assessments is proposed. The method is illustrated with the Russian Railways data.

Keywords: *time series, forecasting, utilization of a railway junction station, forecasting model, expert assessment.*

Введение

При прогнозировании грузоперевозок необходимо учитывать как предысторию самих перевозок, так и различные внешние факторы. Так как множество внешних факторов, влияющих на объемы перевозимых грузов и, как следствие, на загруженность железнодорожных узлов, велико, то для выбора факторов предполагается использовать экспертные оценки. Эксперты высказывают гипотезы о влиянии внешних факторов на грузоперевоз-

Работа выполнена при финансовой поддержке РФФИ, проект № 12-07-31095.

ки. Требуется проверить эти гипотезы, обнаружить влияние внешних факторов и оценить достоверность экспертных суждений.

В данной работе экспертные высказывания о влиянии внешних событий на грузоперевозки представлены в таблице 1. В правой колонке перечислены факторы, которые, с точки зрения экспертов, оказывают влияние на объем грузоперевозок. В центральной колонке перечислены группы груза, соответствующие этим факторам. В левой колонке перечислены высказывания эксперта о характере влияния. В таблице2 перечислены факторы, влияние которых подтвердить или опровергнуть затруднительно из-за недостаточного количества данных.

Экспертные высказывания носят качественный характер и представлены в номинальных (оказывает данный фактор влияние на загруженность железнодорожного узла или нет) и ранговых шкалах (степень влияния — высокая, низкая, фактор не оказывает влияния). В связи с возможным несоответствием экспертного суждения измеряемым данным или с возможной несогласованностью [1] высказываний нескольких экспертов предлагается решить задачу определения достоверности экспертного высказывания и разработать алгоритм получения оценки достоверности. Методы получения экспертных оценок и высказываний и проверки их непротиворечивости описаны в [2].

Таблица 1: Факторы экономического и производственного характера, влияющие на динамику перевозок

№	Вид фактора, влияющего на объем грузоперевозок	Группы грузов и отрасли, на которые оказывается влияние	Степень и характер влияния
1	Мировые и внутренние цены на соответствующие активы	Нефть и нефтепродукты, черные металлы, цветные металлы, удобрения, уголь и др.	На экспортные перевозки влияние сильное. Связь бывает как прямой, так и обратной.
2	Курс рубля к доллару	Грузы, отправляемые на экспорт (нефть и нефтепродукты, металлы, уголь)	Степень влияния для экспортных перевозок зачастую высокая.
3	Сезонность производства природного-климатического характера	Зерно, овощи, бахчевые культуры	Степень влияния высокая. Динамика перевозки связана со сбором урожая. Повышенный объем перевозок приходится на июль, август и сентябрь, иногда на октябрь.
4	Сезонность спроса на продукцию	Строительные грузы (щебень, кирпич, цемент, промсырье)	Степень влияния высокая. Динамика перевозки связана с сезоном строительных работ. Пик приходится на летние месяцы.
5	Особенности технологического цикла в различных отраслях	Различные грузы. Например, в соледобыче в период «соленавигации» высокий уровень добычи, а в остальные периоды зависит от величины складских запасов	Во многих случаях являются ограничителем (как верхней, так и нижней границей) объема погрузки грузов.

Таблица 2: Нерассматриваемые факторы экономического и производственного характера, влияющие на динамику перевозок

№	Вид фактора, влияющего на объем грузоперевозок	Группы грузов и отрасли, на которые оказывается влияние	Степень и характер влияния
1	Сезонность, связанная с навигацией	Грузы, которые перевозятся водным транспортом	Степень влияния высокая для производств, расположенных в пределах досягаемости судоходных путей.
2	Производственные мощности заводов и других предприятий	Добывающие отрасли (угледобыча, добыча соли, камня гипсового, добыча нефти); нефтеперерабатывающие, нефтехимические, металлургические производства.	Во многих случаях являются верхней границей объема погрузки грузов.
3	Запасы готовой продукции на складах предприятий	—	Во многих случаях динамика этого показателя является мерой спроса на продукцию данного предприятия, а спрос в свою очередь оказывает влияние на планируемые объемы производства.
4	Доли погрузки предприятия, приходящиеся на другие виды транспорта	Все грузы, перевозка которых возможна альтернативными видами транспорта	На сверхдальних расстояниях перевозки влияние низкое, для массовых грузов низкая; для остальных грузов, особенно на короткие и средние расстояния.
5	Мощности станций погрузки	Все грузы	Высокое в случаях увеличения производственных мощностей предприятий-грузоотправителей или низких мощностей станции или низких перерабатывающих способностей грузовых фронтов.

Экспертное утверждение о влиянии некоторого фактора на загруженность железнодорожного узла интерпретируется как утверждение о причинно-следственной связи между временными рядами. Таким образом, для оценки достоверности экспертного высказывания необходимо исследовать ряды на наличие причинно-следственной связи.

Способы обнаружения причинно-следственной связи исследованы и широко распространены в [3, 4]. Подход Грейнджера, принятый в этих работах, основан на сравнении точности прогноза исследуемого временного ряда исключительно по его истории и при наличии информации о других временных рядах. В случае, если улучшение прогноза подтверждается статистически, говорят о G-зависимости [5] временных рядов. При этом никакой информации о структуре зависимости между рядами получить не удается. Например, если изменения двух исследуемых рядов вызваны третьим временным рядом, о котором исследователь не знает, ряду будут признаны G-зависимыми, хотя причинно-следственной связи между ними нет. В работе [6] описаны способы расширения подхода Грейнджера для обнаружения структуры связей между временными рядами.

Альтернативный подход, моделирование структурных уравнений (structural equation modelling), рассмотрен в работах [7, 8]. В этом случае исследователь строит модель, основываясь на своих предположениях о структуре причинно-следственных связей между измеренными временными рядами, а также рядами, значения которых по каким-либо причинам неизвестны, а затем настраивает ее в соответствии с данными. В работе [9] обсуждается возможность, сделав вывод о наличии связи между временными рядами, перенести этот результат на временные ряды того же характера, но измеренные в других условиях.

Метод сходящегося перекрестного отображения (англ. convergent cross mapping, CCM) [10, 11], предложенный в 2012 г., позволяет определить, что временные ряды принадлежат одной динамической системе. Этот метод был разработан для выявления причинно-следственных связей в случаях, когда тест Грейнджера не применим или не может обнаружить связи. Метод основан на преобразовании пространства состояний системы и сравнении ближайших соседей одной и той же точки в преобразованных системах и заключается в проверке сходимости коэффициента корреляции между спрогнозированными и исходными значениями исследуемого ряда при увеличении объема выборки.

Постановка задачи

Задано множество временных рядов $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$, в котором временной ряд $\mathbf{s}_i \in \mathbb{Z}^n,\ i \in \mathcal{I}$ — временной ряд, состоящий из элементов, соответствующих числу вагонов, проходящих через станцию. *Положительное значение элемента вектора s_{ij} означает приходящий вагон, отрицательное — отправляющийся.* Каждому $s_{ij},\ j \in \mathcal{J} = \{1, \ldots, n$ поставлены в соответствие две метки $g(s_{ij}) \in G$ и $h(s_{ij}) \in H$ — тип вагона и вид перевозимого груза.

Множество индексов $\mathcal{I}$ соответствует множеству железнодорожных узлов (станций), а множество $\mathcal{J}$ соответствует множеству отсчетов времени.

Задано множество внешних факторов $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, в котором временной ряд $\mathbf{x}_i \in \mathbb{R}^n,\ i \in \mathcal{I}$. Задан набор экспертных высказываний μ о влиянии внешних факторов $\mathbf{x}$ на временные ряды $\mathbf{s}$,

$$\mu = \mu(\mathbf{x}, \mathbf{s}) \in \{\text{«+»}, \text{«-»}\}.$$

Требуется проверить экспертные высказывания и поставить в соответствие каждому высказыванию значение достоверности.

Выявление G-зависимости между временными рядами

Тест Грейнджера применяется при прогнозировании с помощью линейных регрессионных моделей. Пусть $\mathbf{s}$ и $\mathbf{x}$ — исследуемые временные ряды, тогда линейная регрессионная модель имеет вид:

$$\hat{s}_j = \sum_{t=1}^{\tau} a_t s_{j-t} + \sum_{t=1}^{\tau} b_t x_{j-t} + \varepsilon_j, \quad j \in \mathcal{J}, \tag{1}$$

$$\hat{x}_j = \sum_{t=1}^{\tau} c_t s_{j-t} + \sum_{t=1}^{\tau} d_t x_{j-t} + \xi_j \quad j \in \mathcal{J}. \tag{2}$$

Здесь τ – количество предыдущих значений, принимаемых во внимание, ряды $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ содержат веса учитываемых объектов, а ε_j и ξ_j – ошибки прогнозирования. Будем считать, что временной ряд $\mathbf{x}$ влечет за собой временной ряд $\mathbf{s}$, если модуль ошибки $\varepsilon_j = s_j - \hat{s}_j$ прогнозирования ряда $\mathbf{s}$ уменьшается при включении в модель значений ряда $\mathbf{x}$. Для проверки значимости уменьшения ошибки при добавлении в модель ряда $\mathbf{x}$ воспользуемся

статистикой Фишера. Пусть для ряда $\mathbf{s}$ построена модель (1), учитывающая историю ряда $\mathbf{x}$. Построим также линейную регрессионную модель, основанную только на истории $\mathbf{s}$:

$$\tilde{s}_j = \sum_{t=1}^{\tau} a_t s_{j-t} + \tilde{\varepsilon}_j, \quad j \in \mathcal{J}. \tag{3}$$

Тогда применение теста Грейнджера сводится к вычислению статистики

$$F = \frac{(\mathrm{RSS}_{\tilde{\mathbf{s}}} - \mathrm{RSS}_{\hat{\mathbf{s}}})/\tau}{\mathrm{RSS}_{\hat{\mathbf{s}}}/(|J| - 2\tau)} \tag{4}$$

и сравнению ее с критическим значением при заданном уровне значимости. Здесь

$$\mathrm{RSS}_{\hat{\mathbf{s}}} = \sum_{j\in\mathcal{J}} \varepsilon_j^2 = \sum_{j\in\mathcal{J}} (s_j - \hat{s}_j)^2,$$

$$\mathrm{RSS}_{\tilde{\mathbf{s}}} = \sum_{j\in\mathcal{J}} \tilde{\varepsilon}_j^2 = \sum_{j\in\mathcal{J}} (s_j - \tilde{s}_j)^2.$$

Нулевая гипотеза заключается в предположении, что ряд $\mathbf{x}$ не оказывает влияния на значения ряда $\mathbf{s}$. При нулевой гипотезе F принадлежит распределению Фишера со степенями свободы τ и $|J| - 2\tau$. Полученную вероятность отклонить нулевую гипотезу $1 - p(\mathbf{s}, \mathbf{x})$, где $p(\mathbf{s}, \mathbf{x})$ — критическое значение F-статистики будем интерпретировать как достоверность экспертного высказывания $\mu(\mathbf{s}, \mathbf{x})$.

Исследование данных о влиянии цен на основные инструменты на перевозки соответствующих групп грузов приведено ниже, в разделе «Вычислительный эксперимент».

Тест Грейнджера применим к рядам, обладающим не зависящими от времени матожиданием и дисперсией. Если исследуемый ряд не обладает этими свойствами, необходимо привести его к соответствующему виду. Кроме того, в самом определении G-зависимости, данном Грейнджером, заключается предположение о разделимости исследуемых временных рядов. Под разделимостью рядов здесь имеется в виду, что для ряда $\mathbf{s}$ можно построить модель (3), не учитывающую никакой информации о временном ряде $\mathbf{x}$. В случае линейной зависимости это выполняется, однако в случае сложных динамических систем разделимость, как правило, отсутствует.

Использование метода сходящегося перекрестного отображения для выявления зависимости между временными рядами

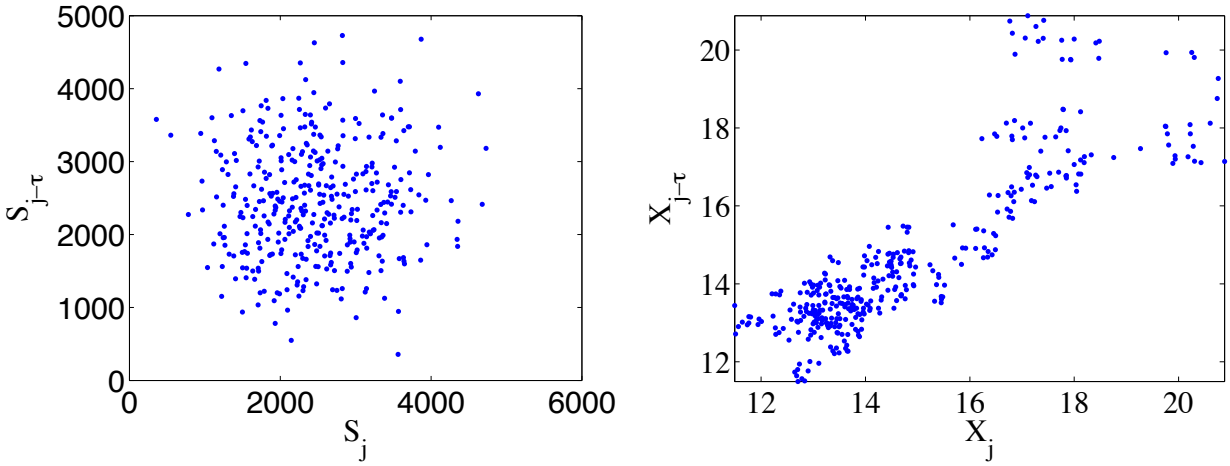
Для исследования более сложных динамических систем, находящихся вне области применения теста Грейнджера, существует метод сходящегося перекрестного отображения. Этот метод существенно опирается на теорему Такенса [12, 13]. Рассмотрим множество M состояний некоторой динамической системы

$$\mathbf{x}_j = \varphi(\mathbf{x}_{j-1}), \quad \mathbf{x} \in \mathbb{R}^d. \tag{5}$$

Согласно теореме Такенса, при соблюдении определенных условий динамическая система может быть реконструирована с помощью последовательности ее наблюдаемых состояний. А именно, утверждается, что множество $M_{\mathbf{x}} = f(M)$, где

$$f(x) = (x, \varphi(x), \ldots, \varphi^{d-1}(x)).$$

может быть использовано для описания системы (5).



(а) Множество $M_{\mathbf{s}} = \{(s_j, s_{j-\tau}),\ j \in \mathcal{J}\}$ (б) Множество $M_{\mathbf{x}} = \{(x_j, x_{j-\tau}),\ j \in \mathcal{J}\}$

Рис. 1: Множества $M_{\mathbf{s}}$ и $M_{\mathbf{x}}$

Опишем процедуру исследования зависимости между временными рядами $\mathbf{s}$ и $\mathbf{x}$ с помощью метода сходящегося перекрестного отображения. Пусть ряды $\mathbf{s}$ и $\mathbf{x}$ и имеют достаточно большую историю. Рассмотрим множество $M = \{(s_j, x_j), j \in |\mathcal{J}|\}$ пар значений временных рядов $\mathbf{s}$ и $\mathbf{x}$, измеренных в j-й момент времени. В силу предположений о линейной зависимости φ между рядами,

$$\varphi(x_\tau) = x_{j-\tau}, \quad \varphi(x_\tau) = x_{j-\tau},$$

где τ — выбранное в (1), (2) значение задержки. Тогда отображения $f_{\mathbf{s}} : M \to M_{\mathbf{s}}$ и $f_{\mathbf{x}} : M \to M_{\mathbf{x}}$, определенные как

$$f_{\mathbf{s}}(s_j, x_j) = (s_j, s_{j-\tau}) \equiv \boldsymbol{s}(j),$$

$$f_{\mathbf{x}}(s_j, x_j) = (x_j, x_{j-\tau}) \equiv \boldsymbol{x}(j),$$

должны сохранять свойства системы (5).

Рассмотрим точку $\boldsymbol{x}(j)$ множества $M_{\mathbf{x}}$. Найдем для нее $d+1$ ближайших соседей $\boldsymbol{x}(j_1), \boldsymbol{x}(j_2), \ldots, \boldsymbol{x}(j_{d+1})$. Предполагая связь между рядами $\mathbf{x}$ и $\mathbf{s}$, считаем, что индексы

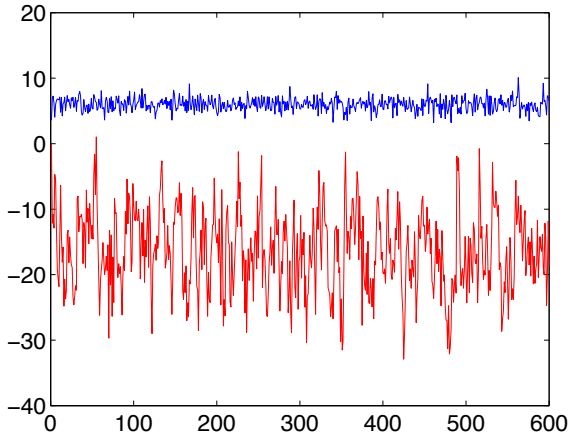
$$j_1, \ldots, j_{d+1} \tag{6}$$

ближайших к $\boldsymbol{x}(j)$ соседей в $M_{\mathbf{x}}$ являются также индексами ближайших к (j) соседей в $M_{\mathbf{s}}$. Используя значения $s_{j_1}, s_{j_2}, \ldots, s_{j_{d+1}}$ временного ряда $\mathbf{s}$, получим прогноз значения s_j:

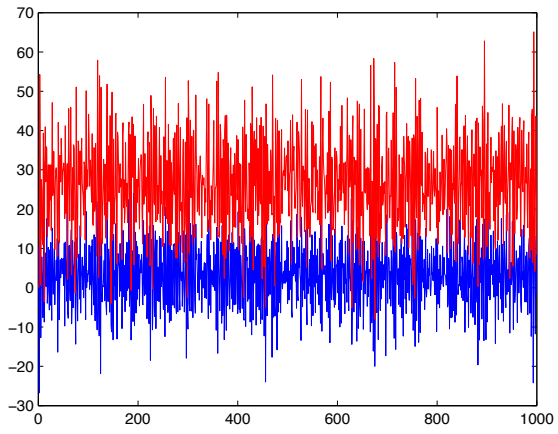
$$\hat{s}_j = \sum_{i=1}^{d+1} w_i s_{j_i}. \tag{7}$$

Веса w_i получаются экспоненциальным взвешиванием евклидовых расстояний $r(s_{j_i}, s_j)$ от s_j до элементов $s_{j_1}, s_{j_2}, \ldots, s_{j_{d+1}}$:

$$w_i = \frac{u_i}{\sum u_k}, \quad u_i = \exp\left(-\frac{r(s_j, s_{j_i})}{r(s_j, s_{j_1})}\right), \quad i = 1, \ldots, d+1,$$

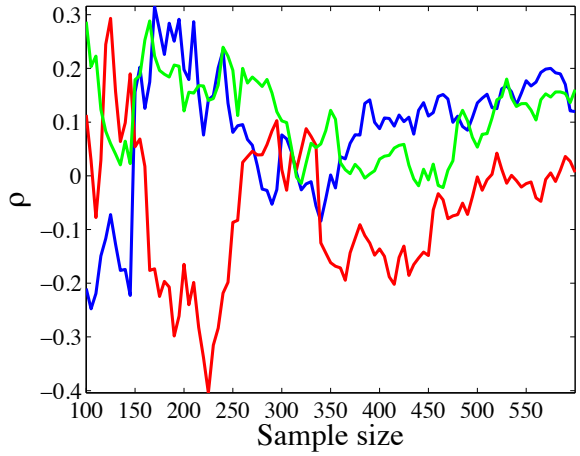


(а) Синий ряд независимый, красный зависит от своей истории и истории синего ряда

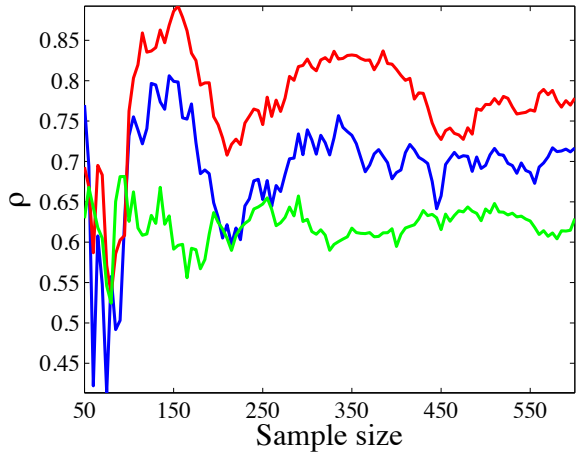


(б) Временные ряды с двусторонней связью

Рис. 2: Синтетические временные ряды



(а) Зависимость корреляция для рядов с односторонней причинно-следственной связью



(б) Зависимость корреляция для рядов с двусторонней связью

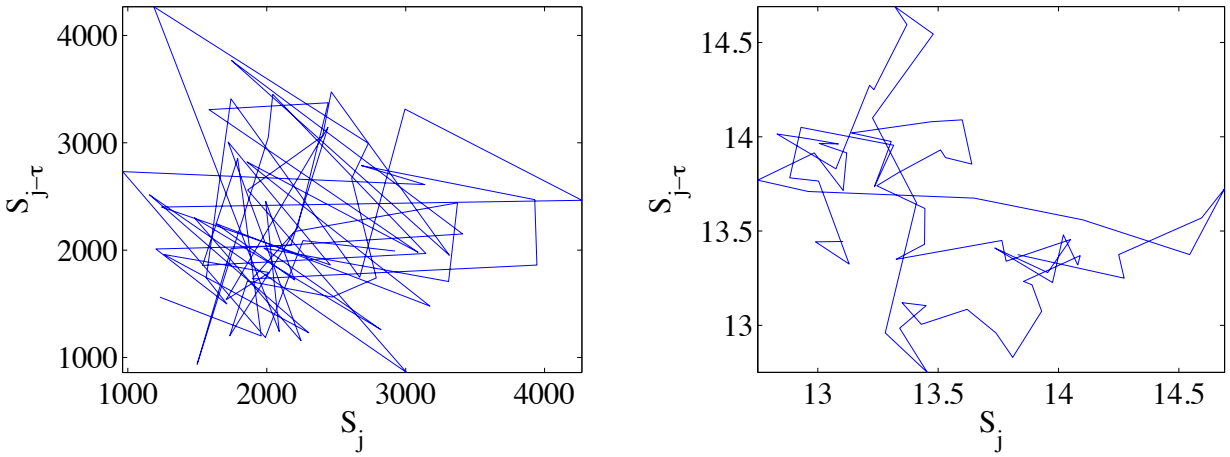
Рис. 3: Зависимость различных корреляций от объема выборки

$$r(s, s') = |s - s'|.$$

При увеличении истории $|\mathcal{J}|$ измеряемых временных рядов расстояния между соседними точками множеств $M_{\mathbf{s}}$ и $M_{\mathbf{x}}$ сокращаются. В случае если ряд $\mathbf{x}$ действительно влияет на $\mathbf{s}$, при $|\mathcal{J}| \to \infty$ прогноз (7) j-го значения ряда $\mathbf{s}$ становится точнее. Тогда коэффициент корреляции

$$\rho(\hat{s}_j, s_j) = \frac{1}{\sigma_s \sigma_{\hat{s}}} \mathsf{E}(\hat{s}_j - \mathsf{E}\hat{s}_j)(s_j - \mathsf{E}s_j) \tag{8}$$

должен стремиться к некоторому ρ_0, при этом значение ρ_0 не должно равняться нулю. Здесь E — математическое ожидание случайной величины, σ — ее дисперсия. Чем сильнее ρ_0 отклоняется от нуля, тем сильнее зависимость между рядами. Наличие этой сходимости будет проверяться в вычислительном эксперименте.



(а) (б)

Рис. 4: Первые сто точек множества $M_{\mathbf{s}}$ и $M_{\mathbf{x}}$, соединенные по возрастанию индекса j

В качестве примера рассмотрим ряд $\mathbf{s}$ загруженности железнодорожного узла цистернами с нефтью и ряд $\mathbf{x}$ цен на нефть. Опишем процедуру добавления элементов к выборке, предназначенную для выявления сходимости выражения (8) к ρ_0. Предположим, что ряд $\mathbf{s}$ зависит от ряда $\mathbf{x}$, тогда ближайшим соседям $\boldsymbol{x}(j)$ будут соответствовать ближайшие соседи $\boldsymbol{s}(j)$. Это позволит получать прогноз $\hat{s}_j$ значений ряда $\mathbf{s}$ по $s_{j_i},\ i = 1, \ldots, d+1$, соответствующим ближайшим соседям $\boldsymbol{x}(j)$. Представим множество индексов $|\mathcal{J}|$ значений временных рядов в виде

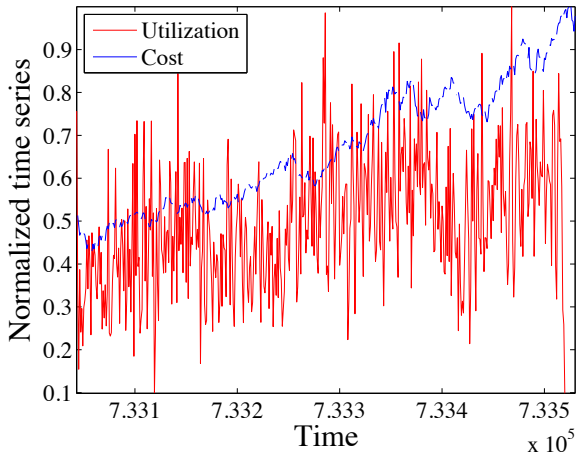
$$\mathcal{J} = \mathcal{L} \bigsqcup \mathcal{T},$$

причем в $\mathcal{T}$ содержатся последние исторические индексы временного ряда. Исключим элементы с индексами $j \in \mathcal{T}$ из множества M, и для каждого из них вычислим $\hat{s}_j$ по формуле (7). Спрогнозировав значения отрезка временного ряда, вычислим коэффициент корреляции (8) между спрогнозированными значениями и реальными значениями исследуемого ряда $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$. Здесь имеется в виду эмпирический коэффициент корреляции, для подсчета которого в формуле (8) матожидание E и дисперсия σ^2 заменяются на среднее значение ряда и среднеквадратичное отклонение:

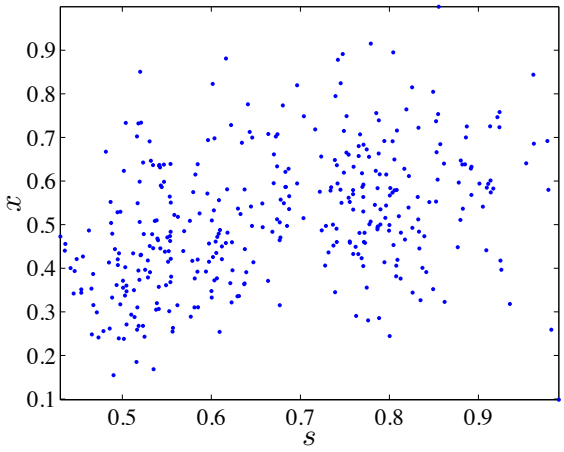
$$\mathsf{E}s \mapsto \bar{\mathbf{s}}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|} \sum_{j \in \mathcal{T}} s_j,$$

$$\sigma_s^2 \mapsto \frac{1}{|\mathcal{T}|} \sum_{j \in \mathcal{T}} (s_j - \bar{\mathbf{s}}_{\mathcal{T}})^2.$$

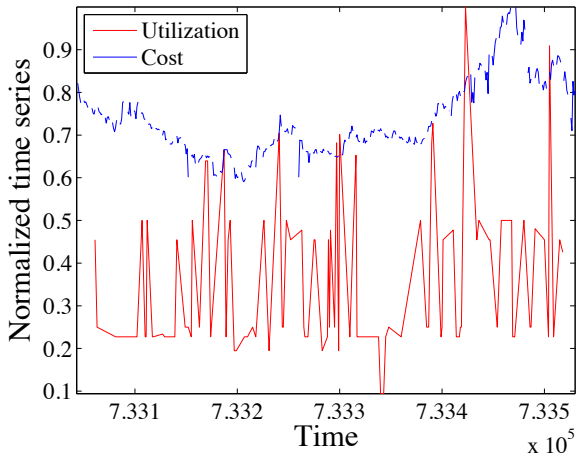
Рассмотрим зависимость коэффициентов корреляции $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$ исходного ряда $\mathbf{s}_{\mathcal{T}}$ и спрогнозированного ряда $\hat{\mathbf{s}}_{\mathcal{T}}$ и коэффициента корреляции $\rho(\hat{\mathbf{x}}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$ рядов $\hat{\mathbf{x}}_{\mathcal{T}}$ и $\mathbf{x}_{\mathcal{T}}$ от размера выборки. При увеличении объема выборки корреляция (8) между спрогнозированными значениями зависимого временного ряда и его измеренными значениями должна возрастать. Будем наращивать размер выборки $|\mathcal{J}|$, оставляя для прогноза последние $0.25|\mathcal{J}|$ значений временного ряда, $\mathcal{T}(|\mathcal{J}|) = \{\lfloor 0.75|\mathcal{J}| \rfloor, \ldots, |\mathcal{J}|\}$, где $\lfloor a \rfloor$ означает округление вниз. Увеличивая $|\mathcal{J}|$ и решая (7), будем вычислять $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$, $\rho(\hat{\mathbf{x}}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$ и корреляцию $\rho(\mathbf{s}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$ между оставленными для прогноза отрезками рядов.



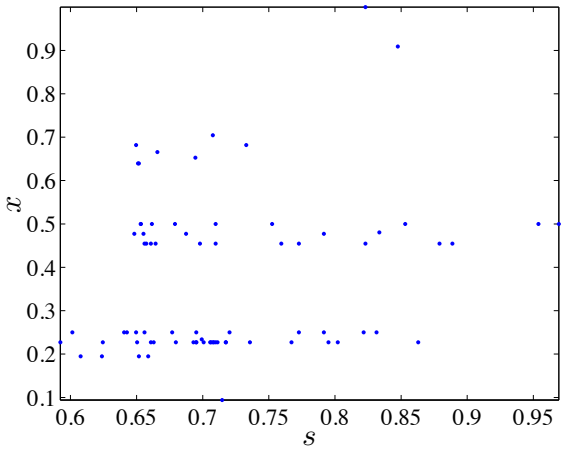
(а) «Нефть и нефтепродукты» —«Цены на нефть»



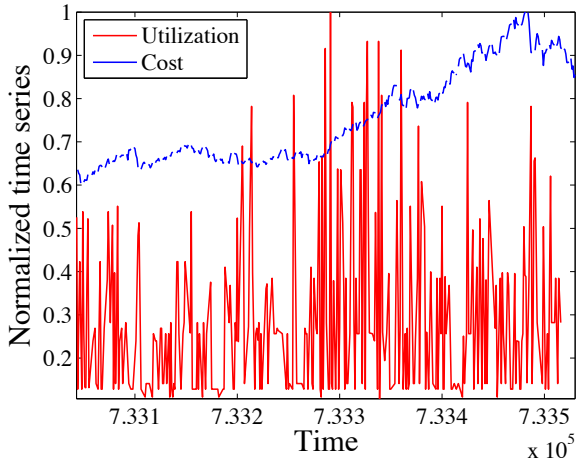
(б) «Нефть и нефтепродукты» —«Цены на нефть»



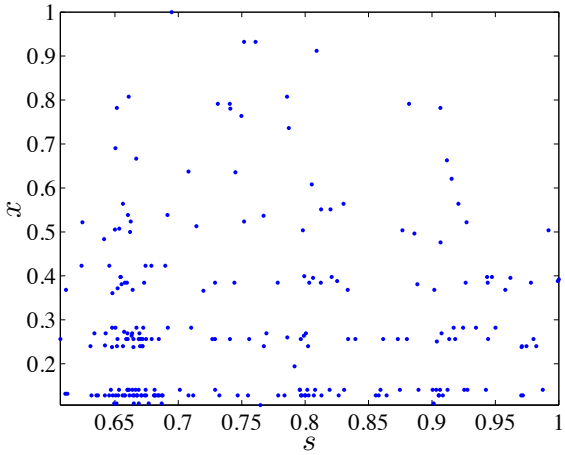
(в) «Сахар» —«Цены на сахар»



(г) «Сахар» —«Цены на сахар»



(д) «Цветные металлы, изделия из них и лом цв. металлов» —«Цены на золото»



(е) «Цветные металлы, изделия из них и лом цв. металлов» —«Цены на золото»

Рис. 5: Пары рядов инструменты-цены

Рассмотрим предполагаемую связь между биржевыми ценами на нефть и загруженностью железнодорожного узла вагонами с нефтью (ожидается, что цены на нефть влияют на загруженность). Тогда модуль коэффициента корреляции $\rho(\hat{\mathbf{s}}_\mathcal{T}, \mathbf{s}_\mathcal{T})$ должен возрастать с объемом выборки и выходить на ненулевую асимптоту. На рис. 8в изображены зависимости коэффициентов корреляции от объема выборки, по оси абсцисс отложена длина временного ряда, полученного добавлением в выборку $|\mathcal{J}|$-го элемента. На графике красного цвета по оси ординат отложены значения коэффициента корреляции спрогнозированых цен на нефть с их истинными значениями, синего — корреляция прогноза загруженности железнодорожного узла с измеренными значениями загруженности. На графике зеленого цвета отложены значения коэффициента корреляция истинных значений исследуемых рядов рядов. Хотя коэффициент $\rho(\hat{\mathbf{s}}_\mathcal{T}, \mathbf{s}_\mathcal{T})$ принимает значения в диапазоне от 0.6 до -0.8, сходимости нет. В вычислительном эксперименте в таких случаях делается вывод об отсутствии связи между временными рядами.

Рассмотрим пример с синтетическими данными, изображенными на рисунке 2а. Синий ряд, обозначим его $\mathbf{x}$, генерируется независимо от красного, $\mathbf{s}$. Ряд $\mathbf{s}$ зависит от $\mathbf{x}$. Зависимость коэффициентов корреляции между различными рядами от объема выборки для них изображена на рис. 3а. Здесь синим цветом изображена зависимость $\rho(\hat{\mathbf{s}}_\mathcal{T}, \mathbf{s}_\mathcal{T})$ от объема выборки, красным — зависимость $\rho(\hat{\mathbf{x}}_\mathcal{T}, \mathbf{x}_\mathcal{T})$ от объема выборки. Зеленый цвет соответствует зависимости $\rho(\mathbf{x}_\mathcal{T}, \mathbf{s}_\mathcal{T})$. Здесь $\rho_0(\hat{\mathbf{s}}, \mathbf{s}) \approx 0.1$, но наблюдается сходимость. Можно сделать вывод о наличии слабой связи между рядами.

Рассмотрим теперь синтетический пример с двусторонней связью 2б, 3б. В этом случае наблюдается сходимость к $\rho_0(\hat{\mathbf{x}}, \mathbf{x}) \approx 0.8$, $\rho_0(\hat{\mathbf{s}}, \mathbf{s}) \approx 0.7$, причем $\rho_0(\hat{\mathbf{x}}, \mathbf{x}) > \rho_0(\mathbf{x}, \mathbf{s})$, $\rho_0(\hat{\mathbf{s}}, \mathbf{s}) > \rho_0(\mathbf{x}, \mathbf{s})$, т. е. для рядов с двусторонней связью метод CCM не только выявляет наличие связи, но и справляется лучше чем кросс-корреляция. Метод CCM дает здесь лучшие результаты, так как он разработан специально для выявления взаимных связей, с которыми тест Грейнджера справляется не всегда. Кроме того, для получения адекватных результатов с помощью CCM необходимо, чтобы отображения $f_\mathbf{s}$, $f_\mathbf{x}$ были взаимно однозначными, т. е. не наблюдалось самопересечений траекторий $s(j)$ и $x(j)$. Из рис. 4 ясно, что для исследуемых данных это не так: наблюдается значительное число самопересечений.

Вычислительный эксперимент

Экспертами предоставлены временные ряды, содержащие данные о биржевых ценах на основные инструменты: сахар, бензин, медь, цинк, золото, никель, пшеницу, мазут, газ, олово, нефть, серебро, свинец за 2007–2008 гг. Эти временные ряды, нормированные на отрезок $[0, 1]$, изображены на рис. 6. Также есть данные о загруженности некоторого железнодорожного узла грузами различных групп каждый день с 01 января 2007 г. в течение 473 дней. В таблице 3 перечислены рассматриваемые группы грузов.

Опираясь на экспертные высказывания о наличии причинно-следственной связи между временным рядами, перечисленные в таблице 1, будем рассматривать различные пары $(\mathbf{s}, \mathbf{x})$ временных рядов. На рис. 5 слева изображены пары временных рядов «Группа груза» —«Цена на некоторый инструмент». Справа на этих рисунках изображено взаимное расположение значений этих рядов. Здесь по оси абсцисс отложены значения ряда загруженности узла грузами соответствующей группы, а по оси ординат — цены на некоторый основной инструмент. Каждая точка соответствует одному дню. Полный список исследуемых зависимостей приведен в таблице 4. На первом месте стоит прогнозируемый ряд $\mathbf{s}$, на втором — ряд $\mathbf{x}$, предположительно оказывающий влияние на $\mathbf{s}$. Для каждой пары

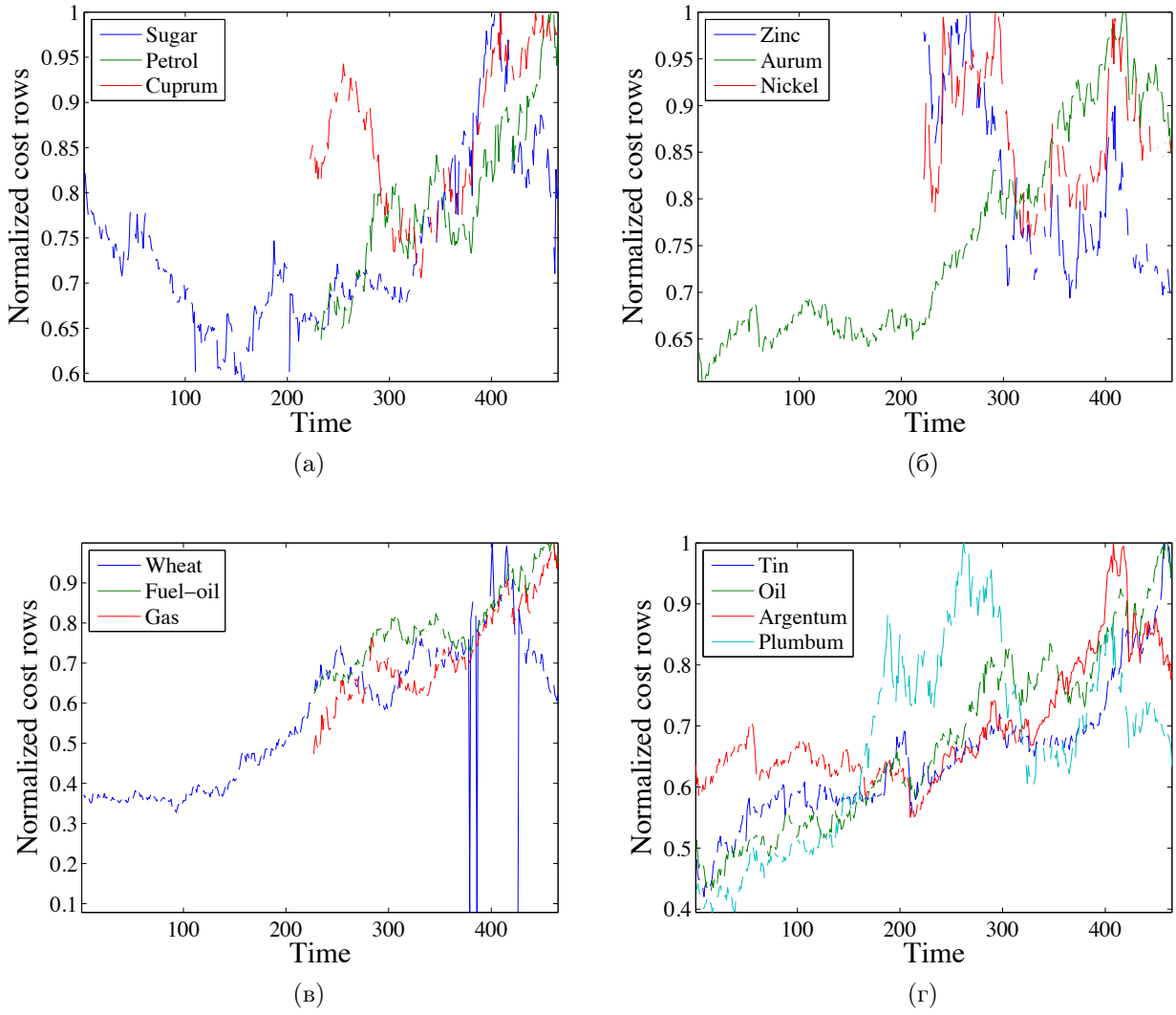


Рис. 6: Временные ряды для цен на основные инструменты

рядов в таблице 4 приведен номер соответствующего этой паре фактора из таблицы 1, а также результат теста на G-зависимость и метода сходящегося перекрестного отображения. В качестве результата теста Грейнджера в таблице приводится значение p-value для F-статистики (4). Решение о наличии G-зависимости принималось при $p < 0.1$ Также для каждой пары рядов приводится значение параметра задержки τ из модели (1). Пе-

Таблица 3: Группы грузов

3 — Нефть и нефтепродукты	13 — Лом черных металлов
7 — Руда железная и марганцевая	16 — Цветные металлы, изделия из них и лом цв. металлов
8 — Руда цветная и серное сырье	9 — Черные металлы
11 — Металлические конструкции	25 — Сахар
12 — Метизы	33 — Сахарная свекла и семена
34 — Зерно	35 — Продукты перемола

ред применением теста Грейнждера все ряды были сглажены, после чего тестировались на стационарность с помощью теста Дики-Фуллера. Нестационарные ряды были продифференцированы. В данной работе дифференцирование проводилось не более одного раза, большее количество преобразований может затруднить интерпретацию полученных результатов. В качестве результатов метода сходящегося перекрестного отображения приведены средние значения $\bar{\rho}_{\mathbf{s}}$ и $\bar{\rho}$ коэффициентов корреляции $\rho(\hat{\mathbf{s}}_{\mathcal{T}}, \mathbf{s}_{\mathcal{T}})$ и $\rho(\mathbf{s}_{\mathcal{T}}, \mathbf{x}_{\mathcal{T}})$. Решение о наличии или отсутствии сходимости коэффициента (8) принималось по виду графика зависимости соответствующего коэффициента от длины выборки $|\mathcal{J}|$. Графики для некоторых пар временных рядов изображены на рис. 8. На графике красного цвета по оси ординат отложены значения коэффициента корреляции $\rho(\hat{\mathbf{x}}, \mathbf{x})$ спрогнозированых цен на нефть с их истинными значениями, синего — корреляция $\rho(\hat{\mathbf{s}}, \mathbf{s})$ прогноза загруженности железнодорожного узла с измеренными значениями загруженности. На графике зеленого цвета отложены значения коэффициента корреляции истинных значений исследуемых рядов. В графе «CCM» встречается формулировка «Недостаточно данных». Так как метод сходящегося перекрестного отображения основан на проверке сходимости коэффициента корреляции (8) при неограниченном увеличении объема выборки, он требует большого объема выборки. Решения принимались при $|\mathcal{J}| > 50$. Для исследования влияния сезон-

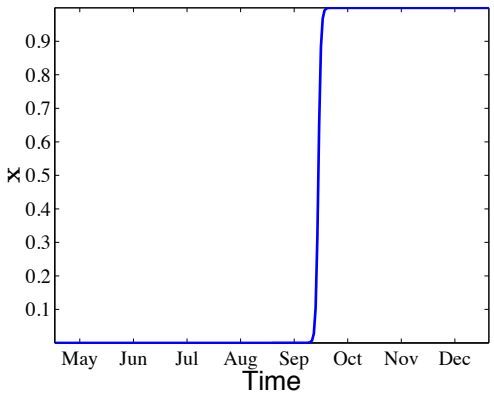


Рис. 7: Зависимость различных корреляций от объема выборки

ности строится вспомогательный временной ряд $\mathbf{x}$, $x_j \in [0, 1]$. Например, свекловичный сахар производится в сентябре–октябре. Для исследования влияния сезонности производства сахара на перевозки сахара будет составлен временной ряд, каждая точка которого соответствует одному дню. Значения ряда, соответствующие месяцам с января по август включительно, приравниваются нулю, а значения ряда, соответствующие месяцам после октября — единице. Значения в сентябре–октябре получим сглаживанием имеющихся значений. Описанный ряд изображен на рис. 7.

Результаты, приведенные в таблице, следует понимать следующим образом. Каждый из рассматриваемых методов исследует временные ряды на наличие некоторой связи, причем каждый в своем смысле. Таким образом, ни для одной из пар рядов, рассмотренных в вычислительном эксперименте, не принято решение о наличии связи в смысле метода сходящегося перекрестного отображения, хотя для некоторых из них сделан вывод о наличии G-зависимости. При этом ни наличие G-зависимости, ни зависимость в смысле метода сходящегося перекрестного отображения не доказывает наличие причинно-следственной связи между рядами. Однако определение G-зависимости согласуется с целью уточнить

прогноз временного ряда $\mathbf{s}$, используя значения временного ряда $\mathbf{x}$. Таким образом, если принимается решение о G-зависимости рядов, будем считать, что экспертное высказывание $\mu(\mathbf{s}, \mathbf{x})$ подтвердилось, и припишем ему оценку достоверности $1 - p(\mathbf{s}, \mathbf{x})$.

Таблица 4: Пары временных рядов, исследуемые на зависимость в вычислительном эксперименте

№	Пара временных рядов $\mathbf{s}$ —$\mathbf{x}$	№ факт.	Тест Грейнджера		CCM	
1	«Нефть и нефтепродукты» — «Цены на бензин»	1	$p = 0.1131,\ \tau = 6$	−	$\bar{\rho} = -0.1296,\ \bar{\rho}_{\mathbf{s}} = 0.1630$	−
2	«Нефть и нефтепродукты» — «Цены на мазут»	1	$p = 0.0581,\ \tau = 7$	+	$\bar{\rho} = 0.1731,\ \bar{\rho}_{\mathbf{s}} = 0.1773$	−
3	«Нефть и нефтепродукты» — «Цены на нефть»	1	$p = 0.0010,\ \tau = 7$	+	$\bar{\rho} = 0.1434,\ \bar{\rho}_{\mathbf{s}} = -0.3157$	−
4	«Сахар» —«Цены на сахар»	1	$p = 0.0038,\ \tau = 7$	+	$\bar{\rho} = 0.1071,\ \bar{\rho}_{\mathbf{s}} = 0.0175$	−
5	«Сахарная свекла и семена» — «Цены на сахар»	1	$p = 0.0111,\ \tau = 8$	+	N/A	N/A
6	«Зерно» —«Цены на пшеницу»	1	$p = 0.1667,\ \tau = 1$	−	N/A	N/A
7	«Продукты перемола» — «Цены на пшеницу»	1	$p = 0.5369,\ \tau = 1$	−	N/A	N/A
8	«Сахарная свекла и семена» — «Сезонность производства сахарной свеклы»	3	$p = 0.6601,\ \tau = 8$	−	N/A	N/A

Обозначения: «+» — сделан вывод о наличии связи между временными рядами, «−» — сделан вывод об отсутствии связи, «N/A» — не достаточно данных для проведения эксперимента.

Заключение

Рассматривается задача прогнозирования загруженности железнодорожного узла. При построении модели используются экспертные высказывания о влиянии внешних факторов на прогнозируемые ряды. В связи с этим ставится задача оценки достоверности экспертного высказывания, и возникает необходимость исследования временных рядов на наличие причинно-следственной связи. В работе рассмотрены два способа выявления связи между рядами, тест Грейнджера и метод перекрестного сходящегося отображения. Описаны области применимости методов и показано, что для достижения целей, описанных в работе, следует применять тест Грейнджера. Достоверность экспертного высказывания оценивается как вероятность на основе исследуемых временных рядов принять решение о наличии G-зависимости между ними.

Литература

[1] *Орлов А. И.* Экспертные оценки // *Заводская Лаборатория.* 1996. Т. 62, №. 1. С. 54–60.

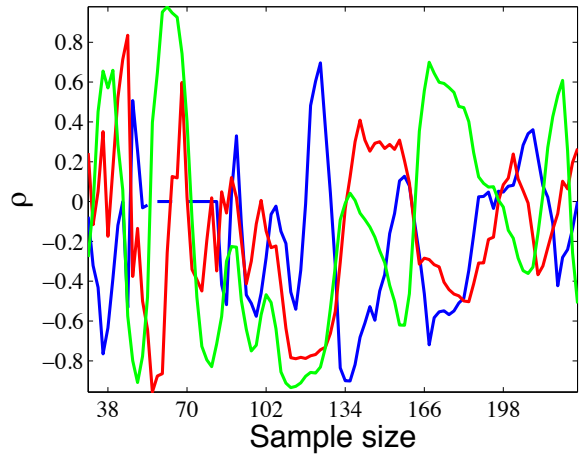
[2] *Орлов А. И.* Организационно-экономическое моделирование: учебник: в 3 ч. / Ч. 2: Экспертные оценки. М.: Изд-во МГТУ им. Н. Э. Баумана, 2011.

[3] *Barrett A. B., Barnett L., Seth A. K.* Multivariate Granger causality and generalized variance // *Phys. Rev. E.* 2010. Vol. 81, no. 4.

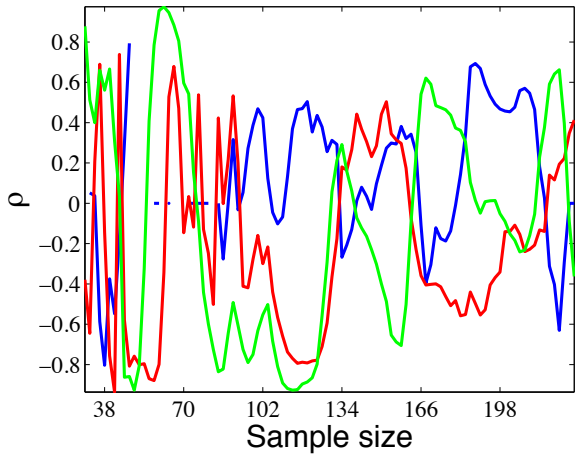
[4] *Мотренко А. П.* Использование теста Гренджера при прогнозировании временных рядов // *Машинное обучение и анализ данных.* 2011. Т. 1. С. 51–60.

[5] *Granger C. W. J.* Investigating causal relations by econometric models and cross-spectral methods // *Econometrica.* 1969. Vol. 37. Pp. 424–432.

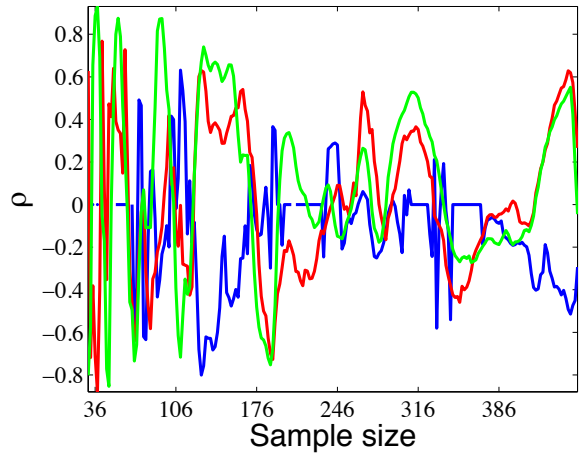
[6] *White H., Lu X.* Granger causality and dynamic structural systems // *Journal of Financial Econometrics.* 2012. Vol. 8, no. 2. Pp. 193–243.



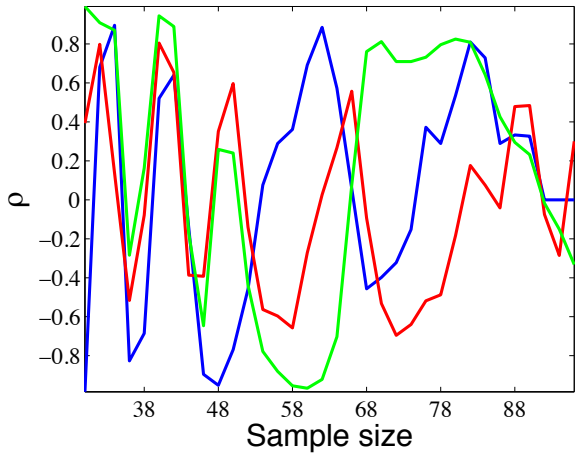
(а) «Нефть и нефтепродукты» —«Цены на бензин»



(б) «Нефть и нефтепродукты» —«Цены на мазут»



(в) «Нефть и нефтепродукты» —«Цены на нефть»



(г) «Сахар» —«Цены на сахар»

Рис. 8: Зависимости различных коэффициентов корреляции от объема выборки для пар инструменты-цены

[7] *Pearl J.* Graphs, causality and structural equation models // *Sociological Methods and Research.* 1998. Vol. 27, no. 2. Pp. 226–284.

[8] *Kline R. B.* Principles and practice of structural equation modeling. New York: Guilford, 2005.

[9] *Pearl J., Bareinboim E.* Transportability of causal and statistical relations: A formal approach // *25th AAAI Conference on Artificial Intelligence Proceedings.* 2011. Pp. 247–254.

[10] *Sugihara G., May R. M.* Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series // *Nature.* 1990. Vol. 344, no. 6268. Pp. 734–741.

[11] *Sugihara G. et al.* Detecting causality in complex ecosystems // *Science.* 2012. Vol. 338, no. 6106. Pp. 496–500.

[12] *Takens F.* Detecting strange attractors in Turbulence // Dynamical systems and turbulence. Eds. D. A. Rand and L.-S. Young. Lecture notes in mathematics ser. 1981. Vol. 898. Pp. 366–381.

[13] *Deyle E. R., Sugihara G.* Generalized theorems for nonlinear state space reconstruction // *PLoS ONE*, 2011. Vol. 6, no. 3. e18295.

Об эвристическом методе разрешения неоднозначности при морфологическом анализе незнакомых фамилий*

Сулейманова Е. А.[1]***, Константинов К. А.***[1]
yes@helen.botik.ru
[1]Институт программных систем имени А. К. Айламазяна РАН

Статья посвящена развитию подхода к морфологическому анализу незнакомых фамилий в русскоязычном тексте, реализованного в специальном модуле системы интеллектуального анализа текста ИСИДА-Т. Идея подхода состоит в первоначальном построении заведомо избыточного множества вариантов-гипотез и последующем сокращении числа вариантов с помощью различных эвристических методов: исключение невозможных вариантов на основании дополнительных проверок правилами-фильтрами; кластеризация словоформ и фильтрация результатов внутри кластера; ранжирование вариантов по предпочтительности. Анализируются ограничения на возможности метода, вытекающие, в частности, из его детерминированной природы.

Ключевые слова: *морфологический анализ собственных имен, неоднозначность, сокращение множества гипотез, эвристические методы.*

On a heuristic approach towards ambiguity resolution in unknown surname morphological analysis*

Suleymanova E. A.[1]***, Konstantinov K. A.***[1]
[1]Program Systems Institute RAS

The paper extends the approach towards morphological analysis of unknown Russian surnames, which has been implemented as part of the ISIDA-T information extraction software. The idea is first to construct a redundant set of hypotheses and then to reduce the number of hypotheses using various heuristic techniques like ruling out impossible options with filtering rules; clustering textual forms and filtering hypotheses within clusters; preference-based ranking of options. Some limitations of the method are analysed, including those due to its deterministic nature.

Keywords: *morphological analysis of proper names, ambiguity, hypotheses reduction, heuristic methods.*

Введение

Статья посвящена развитию подхода к морфологическому анализу незнакомых фамилий в русскоязычном тексте [1], реализованному в специальном модуле системы интеллектуального анализа текста «Исида-Т» [2]. Любая система, работающая с текстом и использующая морфологический анализатор со словарем основ, сталкивается с проблемой обработки незнакомых слов — т.е. лексем, по тем или иным причинам отсутствующих в словаре. Эта проблема не решается наращиванием объема словаря, поскольку, во-первых, словообразовательные возможности языка неисчерпаемы, а во-вторых, существуют принципиально открытые категории лексики.

Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-00307а.

В контексте задачи извлечения информации из текстов последние представляют наибольшую важность, поскольку ФИО лиц, названия всевозможных организаций, названия географических и административных объектов служат для идентификации участников фактов.

Для обработки незнакомой (отсутствующей в словаре) лексики морфологические анализаторы обычно используют вероятностные модели предсказания, основанные на сопоставлении правых концов словоформ. Применительно к фамилиям такое морфологическое предсказание «на общих основаниях» нередко приводит к ошибкам (точные замеры нами не проводились, но некоторые показательные примеры из практики системы «Исида-Т», использующей морфологический анализатор АОТ [3], приведены в упомянутой статье [1]).

В связи с этим для системы извлечения информации был разработан специальный модуль морфологического анализа фамилий (МАФ), который работает после общего морфологического анализа.

Общее описание МАФ в системе «Исида-Т»

МАФ опирается на специально построенную модель словоизменения фамилий, которая в текущей реализации включает 59 словоизменительных классов. При построении системы классов учитывались как общие закономерности субстантивного (по типу существительного) и адъективного (по типу прилагательного) словоизменения в русском языке, так и особенности склонения фамилий. Грамматический род фамилии считается классифицирующей категорией (как род у существительного). Таким образом, фамилиям *Иванов* и *Иванова* соответствуют разные классы, так же как и мужской и женской фамилиям *Петренко*, *Обама* и т.п.

В задаче МАФ можно выделить две составляющих.

1. Лексический анализ — определить, формой какой фамилии является данная словоформа. Считаем, что фамилия однозначно определяется леммой (канонической формой) и именем словоизменительного класса. В дальнейшем фамилию как результат распознавания словоформы будем называть *вариантом фамилии*.
2. Собственно морфологический анализ — определить грамматические характеристики словоформы как формы некоторого варианта фамилии.

Неоднозначность результатов характерна как для выбора варианта фамилии, так и для определения морфологических характеристик словоформы. Основной упор при МАФ делается на разрешение лексической неоднозначности, т.е. сокращение, в идеале — до одного, числа разных вариантов фамилии для словоформы. Задача однозначного определения грамматических характеристик словоформы при этом не ставится.

Каждый результат, полученный МАФ, называется *гипотезой*. Гипотеза включает в себя вариант фамилии, вариант морфологического разбора и некоторую служебную информацию.

МАФ в системе использует фильтровый подход: для словоформ строится заведомо избыточное множество гипотез, которое затем последовательно сокращается с помощью различных методов. При таком подходе минимизируется риск упустить правильный вариант.

При построении начального множества гипотез для словоформы фамилии по возможности учитывается ее левый контекст — «этикетная» лексема (*господин*, *госпожа*, *мсье*, *леди* и т.п.), имя (известное словарю системы), возможно с отчеством. Предполагается, что значения рода и падежа всех элементов такой именной группы, включая словоформу фамилии, совпадают (число рассматривается только единственное). Учет контекста,

особенно если он в целом неомонимичный, предотвращает построение заведомо неверных гипотез уже на начальном этапе. Например, для словоформы *Проди* после омонимичного по падежу мужского имени *Романо* будут построены как гипотезы с несклоняемым вариантом мужской фамилии *Проди* для всех падежей, так и гипотеза со склоняемой мужской фамилией *Продя* в родительном падеже. Но если перед именем будет стоять неомонимичная словоформа в дательном падеже *господину*, то гипотеза для склоняемого варианта фамилии не построится (в таблице словоизменительных классов для такого варианта не найдется соответствия).

Более подробно модель словоизменения фамилий и ее использование при построении начального множества гипотез описаны в уже упомянутой статье [1].

Для дальнейшего сокращения множества гипотез применяются следующие методы.

Методы сокращения множества гипотез

Исключение некорректных вариантов с помощью правил

Проверка дополнительных условий, таких как длина леммы, число гласных букв в лемме и т.п., позволяет исключить небольшое число «запланированных» ошибочных вариантов, которые могли быть получены в результате сопоставления словоформы с таблицей словоизменительных классов (например, вариант фамилии с классом «ов/ин муж» у фамилии *Скин* — ср. *Якин*).

Фильтрация результатов путем сравнения данных из одного текста

Проанализированные к данному моменту словоформы объединяются в кластеры при помощи простого алгоритма частичного сопоставления.

Предполагается, что кластер соответствует либо одной фамилии, либо двум парным по роду фамилиям, принадлежащим в нашей модели разным классам, но «в миру» считающимся одной фамилией (как *Иванов* и *Иванова*).

Сопоставление гипотез внутри кластера с учетом рода позволяет существенно сократить множество разных вариантов фамилии для каждой словоформы. Но очевидно, что при единичном или бесконтекстном употреблении фамилии в тексте для уменьшения неоднозначности результатов требуются другие методы.

Проверка по словарю известных фамилий

Чтобы не допускать казусов в результатах работы системы извлечения информации в тех случаях, когда речь идет о фамилии какого-либо известного персонажа, а текстовых данных недостаточно для того, чтобы выбрать правильный вариант, перед применением эвристических правил предпочтения имеющиеся в кластере варианты фамилии проверяются по словарю известных фамилий. Словарь совсем небольшой, содержит порядка двух десятков фамилий (в основном иностранных, но не только — *Обама*, *Буш*, *Шойгу*).

Выбор и маркировка предпочтительных вариантов

Правила-предпочтения делятся на частные и общие. Сначала применяются частные. Если применилось хотя бы одно частное правило (условия правил сформулированы таким образом, что больше одного не может примениться), конец работы алгоритма. Иначе — переход к общим правилам предпочтения.

Частные правила-предпочтения маркируют предпочтительные варианты фамилий на основе типичных для них концовок (которые, как предполагается, маловероятны для других фамилий), например:

Если в кластере все разделы имеют совпадающую текстовую форму фамилии и она оканчивается на: `ади`, `иди`, `изи`, `оди`, `ози`, `они`, `ори`, `рти`, `жоли`, `иани`, `швили` и т.п. (реальный список значительно длиннее) и в каждом разделе есть строка с именем класса «неизм

муж» или «неизм жен», то во всём кластере строки с «неизм» пометить специальным атрибутом, остальные строки пометить как имеющие низкий приоритет.

Общие правила предпочтения основаны на довольно формальном критерии: предпочтительной считается гипотеза с максимальным *индексом совпадения*. Индекс совпадения приписывается гипотезе при построении — это число букв в ячейке таблицы словоизменительных классов, с которой успешно сопоставилась словоформа (т.е. абсолютная длина буквенного совпадения, «ответственного» за эту гипотезу). Общие правила предпочтения выполняются отдельно для вариантов фамилий мужского и женского рода.

Вывод результатов работы МАФ

К концу работы алгоритма в кластере сохраняются все гипотезы, оставшиеся после фильтрации (до проверки по словарю и применения правил предпочтения). При этом возможны следующие случаи:

— среди гипотез нет помеченных как низкоприоритетные. Это возможно в двух случаях: (1) если все неправильные гипотезы отсеялись еще на этапах фильтрации (идеальный результат) и (2) наоборот, все неправильные гипотезы, оставшиеся после фильтрации, дошли до конца работы алгоритма;

— среди гипотез есть как «предпочтительные» — помеченные специальными атрибутами (найденные по словарю или выбранные правилами предпочтения), так и низкоприоритетные.

Аннотации класса Morpho (куда в системе записываются результаты морфологического анализа) строятся по всем гипотезам, независимо от помет. Однако при построении специальных аннотаций для фамилий те аннотации Morpho, которые построены по низкоприоритетным гипотезам, из рассмотрения исключаются.

Ограничения подхода

Предел полноты

Контекст. Для того чтобы попасть в число рассматриваемых, фамилия должна хотя бы один раз встретиться в заданном контексте — либо в сопровождении «этикетной» лексемы, либо с личным именем, известным словарю личных имен системы, либо, на худой конец, с инициалом. Однократно встретившаяся одиночная фамилия, так же как и фамилия, однократно встретившаяся с неизвестным системе именем, сейчас просто игнорируется алгоритмом. Очевидно, что включить в словарь все мыслимые личные имена (особенно иностранные) невозможно. Кроме того, пополнение словаря относительно редкими именами, особенно при наличии у них омоформ среди более распространенных имен или нарицательной лексики, неизбежно приводит к увеличению «шума» (*Марин*, *Светлан*, *Танка*).

Формат. Алгоритм рассчитан на распознавание и морфологический анализ фамилий в составе относительно простой модели:

имя + (отчество)? + (среднее имя)? + («элемент фамилии»[1])? + фамилия.

При этом среднее имя должно опознаваться словарем личных имен.

Предел точности

Омоформия личных имен разного грамматического рода. Нейтрализация по роду в личном имени, с которым употреблена в тексте фамилия (*Роберт — Роберта*, *Валентин — Валентина*, *Александр — Александра*), усложняет выбор правильного варианта.

[1]Артикль или частица *фон*, *ван*, *фог*, *ле*, *де*, *ди*, *делъ* и т.п.

Чем больше вхождений фамилии в тексте, тем больше шансов на правильный результат. При единичном упоминании фамилии с таким омонимичным именем алгоритм в его теперешнем виде часто оказывается не способен сделать правильный выбор.

Омонимия личных имен и прочих собственных названий. Пример регулярной ошибки: в тексте *президента Израиля Шимона Переса* алгоритм принимает за фамилию словоформу *Шимона* (*Израиль* — название государства и личное имя из словаря имен). Поскольку алгоритм фамилий работает до синтаксического анализа, предотвратить такую и подобные ей ошибки (*Рада* с фамилией *Украины*) удается только вмешательством в результаты общего морфологического анализа.

Омонимия (омоформия) личных имен и нарицательных имен. Алгоритм может «шуметь» из-за случайных совпадений слов, написанных с заглавной буквы в силу позиции, с формами личных имен из словаря. Например, любое слово с заглавной буквы после словоформы *Такая* в начале предложения может быть принято за фамилию (поскольку имя *Такай* есть в словаре имен).

Морфология составных фамилий. Фамилия из двух частей, разделенных дефисом, анализируется как одна словоформа (т.е. классифицирующая тройка для первой части не определяется), поэтому на безошибочный результат можно рассчитывать только для словоформы в канонической форме.

Зависимость от разнообразия текстовых форм. Чем меньше текстовых данных (чем более омонимичен лексический контекст и/или меньше вхождений фамилии), тем больше возрастает роль правил предпочтения. Для ограниченного числа типов фамилий с характерными «фамильными» окончаниями оказывается достаточно несложных эвристик, построенных вручную и опирающихся исключительно на буквенный состав словоформы. В огромном большинстве случаев правильный выбор может быть сделан лишь с учетом комплекса факторов разной природы, в том числе и трудно формализуемых:

— этноязыковая окрашенность имени (как минимум, русское-нерусское),
— этногеографическая привязка фамилии (может быть понятна из контекста),
— древесно-синтаксический контекст, снимающий падежную и родовую омонимию (этот уровень анализа недоступен алгоритму),
— наконец, фамилия может просто быть уже знакома из других текстов.

Примеры непростых задач для алгоритмического распознавания фамилий:

— по форме дательного падежа (*Вильгельму*) *Шойбле* предпочесть несклоняемый вариант *Шойбле* склоняемым вариантам *Шойбла* и *Шойбля* (ср. склоняемые в мужском роде фамилии *Воля*, *Сиривля*);
— по форме родительного падежа (*Марио*) *Драги* предпочесть несклоняемый вариант *Драги* склоняемому *Драга* (ср. *Брага*);
— по формам творительного падежа *Луневым*, *Кочевым* предпочесть варианты *Лунев*, *Кочев* вариантам *Луневой* и *Кочевой* (ср. *Броневой*, *Кошевой*).

Очевидно, выбор в таких случаях имеет вероятностную природу. Поэтому естественно было бы попытаться решать эту задачу статистическими методами с применением машинного обучения.

Формат. Уже упомянутое ограничение на формат приводит к потерям не только полноты, но и точности, поскольку к части «неформатных» случаев алгоритм всё-таки применяется, но с неверным результатом. Например, если в конструкции со средним име-

нем последнее не распознано словарем имен (например, *Энрике Пенья Ньето*), то оно будет разобрано как фамилия (*Энрике Пенья*).

Количественная оценка подхода

Как уже говорилось, алгоритм анализа фамилий встроен в систему извлечения информации. В настоящее время мы не располагаем отдельным инструментом для оценки количественных показателей работы алгоритма МАФ, но о качестве его работы тем не менее можно судить по результатам работы всей системы при решении задач выявления и извлечения собственных имен лиц (инструмент для оценки работы системы в целом имеется).

В задаче *выявления* оценивается факт распознавания собственного имени и правильность его границ (поскольку распознаваться должны собственные имена разной структуры и полноты). При этом правильность заполнения атрибутов собственного имени не оценивается. Можно считать, что применительно к фамилиям оценивается задача семантической классификации (является незнакомое слово фамилией или нет).

В задаче *извлечения* оценивается правильность заполнения атрибутов собственного имени. В частности, для фамилии оценивается только правильность определения леммы. Правильность определения словоизменительного класса фамилии не оценивается, т.к. для общей задачи извлечения информации это неважно, а эталонная разметка коллекции для оценки определения словоизменительного класса фамилии потребовала бы дополнительных усилий и соответствующей квалификации аннотатора.

Приведем оценку, полученную при решении задачи обнаружения и извлечения собственных имен лиц; оценка получена при очередном прогоне тестовой коллекции из 600 новостных текстов (6132 эталона) [4]:

— задача выявления собственных имен лиц: F-мера 95, 50 (при точности 96, 64 и полноте 94, 37);
— задача извлечения собственных имен лиц: F-мера 93, 58 (при точности 94, 71 и полноте 92, 48).

Разумеется, интересно было бы сравнить возможности предлагаемого алгоритма МАФ с другими аналогичными решениями. Однако в отсутствие общего стандарта для оценки такой задачи количественное сравнение результатов затруднительно. Нам известно описание только одного подхода к автоматическому выявлению и определению «морфологических и синтаксических характеристик» незнакомых фамилий [5]. Подход используется семантико-синтаксическим анализатором SemSin и опробован на коллекции объемом в 500 предложений. «Показано, что на новостных текстах удается опознать как фамилии, имена или инициалы до трети неизвестных слов с точностью свыше 95%». Из такой формулировки трудно понять, точность какой именно задачи оценивалась — только семантической классификации незнакомых слов или их распознавания с точностью до леммы. Кроме того, невозможно судить о показателе полноты.

Заключение

Анализ ограничений описанного подхода к морфологическому анализу незнакомых фамилий позволяет предположить, что для получения результатов приемлемого уровня на произвольной коллекции текстов требуется дальнейшее усовершенствование метода в следующих направлениях:

— использование в правилах для разрешения неоднозначности дополнительных параметров, характеризующих контекст;

— использование, наряду с «ручными» правилами, вероятностных алгоритмов, обучаемых на представительном размеченном корпусе.

Вероятностные модели вполне успешно используются для снятия частеречной омонимии при общем морфологическом анализе [6, 7, 8].

Кроме того, исследовательский интерес представляет использование результатов анализа более высоких уровней для снятия лексико-морфологической неоднозначности фамилий.

Литература

[1] *Сулейманова Е. А., Константинов К. А.* Морфологический анализ незнакомых фамилий в русскоязычном тексте // Программные продукты и системы: Международное научно-практическое приложение к международному журналу «Проблемы теории и практики управления». — 2009. — № 2. — С. 66–71.

[2] *Куршев Е. П., Кормалев Д. А., Сулейманова Е. А., Трофимов И. В.* Извлечение информации из текста в системе ИСИДА-Т // Труды XI Всероссийской научной конференции RCDL'2009, Петрозаводск: КарНЦ РАН, 2009. — С. 247–253.

[3] *Сокирко А. В.* Морфологические модули на сайте www.aot.ru // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. Диалог'2004, М.: Наука, 2004. — С. 559–564.

[4] Коллекция «Persons-600». — `http://ai-center.botik.ru/Airec/index.php/ru/collections/27-persons-600`.

[5] *Боярский К. К., Каневский Е. А.* Автоматическое выявление фамилий в тексте // Информационные системы для научных исследований: Сборник научных статей. Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (Санкт-Петербург, 10–12 октября 2012 г.), СПб, 2012. — С. 195–198.

[6] *Зеленков Ю. Г., Сегалович И. В., Титов В. А.* Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Тр. междунар. семинара Диалог'2005, 2005. — С. 188–197.

[7] *Сокирко А. В., Толдова С. Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп). — 2005. — `http://download.yandex.ru/company/grant/2005/01_Sokirko_92802.pdf`.

[8] *Sharoff S., Nivre J.* The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25-29 мая 2011 г.), М.: Изд-во РГГУ, 2011. — С. 591–604.

Исследование погрешности оценок скользящего экзамена*

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

В работе на модельных задачах проводится сравнение различных вариантов оценки скользящего экзамена, таких как leave-one-out и K-fold cross-validaton, а также оценки, основанной на эмпирическом риске с поправкой на смещение. Приводятся зависимости точности оценок от байесовского уровня ошибок. В качестве методов классификации рассмотрены дискриминант Фишера и гистограммный классификатор. В рамках исследования рассмотренные оценки риска показали достаточно близкие результаты по точности.

Ключевые слова: *распознавание образов, машинное обучение, решающая функция, вероятность ошибочной классификации, эмпирический риск, скользящий экзамен.*

Investigation of accuracy of crossvalidation*

Nedelko V. M.
Institute of Mathematics SB RAS

Several kinds of cross-validation, such as leave-one-out and K-fold cross-validaton, and also an empirical risk based bias adjusted estimate are investigated. The accuracies of the estimates on synthetic data are compared using Fisher's discriminant and histogram classifier. All estimates under consideration have shown similar accuracy.

Keywords: *pattern recognition, machine learning, decision function, misclassification probability, empirical risk, crossvalidation.*

Введение

Оценка скользящего экзамена является, пожалуй, наиболее часто используемой на практике оценкой вероятности ошибочной классификации [1, 2, 3, 4]. При этом существуют различные варианты этой оценки: leave-one-out, K-fold cross-validaton, случайные разбиения выборки, разбиения с соблюдением пропорции классов и другие.

Несмотря на активные исследования в этой области, до сих пор нет полного понимания, какая из оценок предпочтительнее и в каких условиях [5].

Данная работа посвящена исследованию вопроса, какую долю объектов следует выделять для контроля. Для наглядности будут рассмотрены два варианта: leave-one-out, когда для контроля выделяется один объект, и K-fold cross-validaton, когда выборка разбивается на K частей, каждая из которых поочередно выступает в качестве контрольной выборки. На самом деле, первый вариант является частным случаем второго при $K = N$, поэтому речь идет о выборе K.

В литературе можно встретить утверждение, что наилучшим выбором является использование K от 5 до 10 и что такой вариант предпочтительнее, чем leave-one-out. Действительно, есть ряд доводов в пользу такого утверждения. Первый довод состоит в том, что leave-one-out менее устойчива. Легко привести примеры выборок для случая двух

Работа выполнена при финансовой поддержке РФФИ, проект №11-07-00346-а.

классов, где эта оценка равна 1. Однако при малых K оценка cross-validaton становится существенно смещенной (поскольку строится при меньшем объеме обучающей подвыборки). Другой довод состоит в том, что при $K << N$ мы имеем оценку доверительного интервала для риска. Это классический доверительный интервал для параметра биномиального распределения, который применим для оценивания риска по одной контрольной выборке.

Очевидно, что оценка K-fold cross-validaton не хуже, чем оценка по одной контрольной выборке объема $\frac{N}{K}$. Фактически же оценка cross-validaton существенно лучше последней, но нет приемлемых аналитических оценок, насколько она лучше. Получается, что на практике мы рассчитываем на то, что оценка cross-validaton точнее, чем это удается доказать. Для leave-one-out вообще неизвестно приемлемых оценок точности. Но то, что мы не можем оценить ее точность, не означает, что она менее точна, чем cross-validaton.

В данной работе мы будем сравнивать точность оценок риска методом статистического моделирования.

При этом возникает еще одна нетривиальная проблема: выбор функции, характеризующей погрешность. Очевидно, что оценить вероятность 0,5 с погрешностью 0,1 совсем не то же самое, что оценить с той же погрешностью вероятность, к примеру, 0,15. Однако это соображение не позволяет однозначно определить выбор этой функции. В работе будет использован достаточно очевидный вариант функции погрешности, учитывающий приведенное соображение.

Основные понятия

Пусть X — пространство значений переменных, используемых для прогноза, а $Y = \{-1, 1\}$ — 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 $\langle D, B, \mathsf{P}_c \rangle$, где B — σ-алгебра, P_c — вероятностная мера.

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

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

$$R(c, \lambda) = \mathsf{E}\mathcal{L}(y, \lambda(x)) = \int_D \mathcal{L}(y, \lambda(x))\ \mathsf{P}_c(dx, dy), \quad x \in X, y \in Y.$$

При $\mathcal{L}(y, y') = I(y = y')$, где $I(\cdot)$ – индикаторная функция (равна 1, если условие истинно, и 0 — иначе), риск есть вероятность ошибочной классификации.

Заметим, что значение риска зависит от c — распределения, которое неизвестно. Поэтому возникает задача оценивания риска по выборке.

Пусть $V = \big((x^i, y^i) \in D \,|\, i = 1, \ldots, N\big)$, $V \in D^N$ — случайная независимая выборка из распределения P_c.

Алгоритм (метод) построения решающих функций есть отображение $Q\colon D^N \to \Lambda$, где Λ — заданный класс решающих функций, а $\lambda_{Q,V}$ — функция, построенная по выборке V методом Q.

Наиболее просто можно оценить риск по контрольной выборке:

$$R^*(V^*, \lambda) = \frac{1}{N^*} \sum_{i=1}^{N^*} \mathcal{L}(y_*^i, \lambda(x_*^i)),$$

где $V^* = \big((x^i_*, y^i_*) \in D \,|\, i = 1, \ldots, N^*\big)$, $V^* \in D^{N^*}$ — отличная от V выборка из распределения P_c. В этом случае доверительный интервал для риска есть классический односторонний доверительный интервал для параметра биномиального распределения. Такой интервал не зависит от метода классификации Q.

Для оценки риска естественно использовать эмпирические функционалы качества, т.е. точечные оценки риска, такие как эмпирический риск, оценка скользящего экзамена, оценка bootstrap и т.п.

Эмпирический риск определяется как средние потери на выборке:

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

Оценка скользящего экзамена определяется как

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

где $V'_i = V \backslash (x^i, y^i)$ — выборка, получаемая из V удалением i-го наблюдения,

Для вычисления оценки K-fold cross-validaton исходная выборка разбивается на K равных частей (для простоты полагаем, что N кратно K).

$$\breve{R}^K(V, Q) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y^i, \lambda_{Q,V^K_i}(x^i)),$$

где V^K_i — выборка, получаемая из V удалением всей подвыборки, которой принадлежит i-е наблюдение.

Функция погрешности оценки

Мерой качества точечных оценок обычно выступает средний квадрат отклонения от оцениваемой величины. Однако для оценивания риска такая функция погрешности мало подходит. Действительно, ситуация, когда мы оценили риск как нулевой, а он равен 0,1, совсем не равноценна ситуации, когда мы дали оценку 0,4 при истинном значении 0,5. В первом случае погрешность значительная, а во втором — несущественная.

Очевидным решением является использование не квадрата отклонения, а некоторой более сложной функции погрешности $\Delta(\cdot, \cdot)$.

Из содержательного смысла величины данная функция должна «усиливать» погрешность при приближении оцениваемой величины к крайним значениям 0 и 1. Однако это соображение не устраняет произвол в выборе конкретного вида функции.

Попытаемся уменьшить неопределенность, вводя дополнительные ограничения и требования.

Предположим, что мы (в схеме Бернулли) оцениваем вероятность p события по его частоте ν. Достаточно естественным представляется требование, чтобы мера качества такой частотной оценки не зависела от оцениваемой вероятности.

Это достигается выбором функции погрешности

$$\Delta_1(\bar{R}, R) = \frac{(\bar{R} - R)^2}{R(1 - R)}.$$

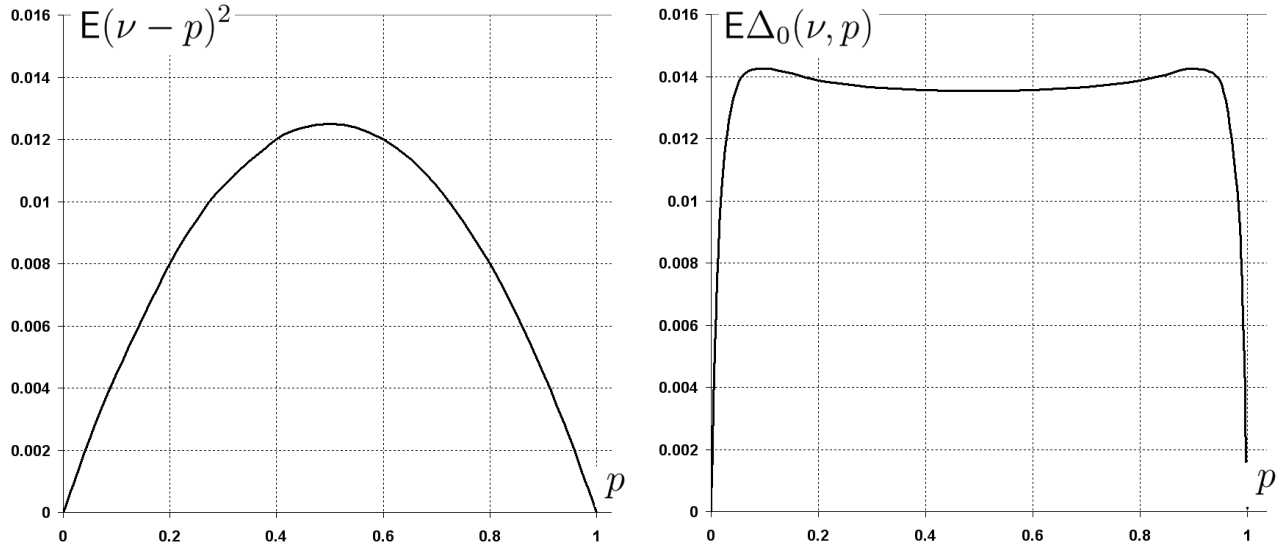


Рис. 1: Зависимости ожидаемой погрешности частотной оценки от вероятности события в схеме Бернулли при $N = 20$

Действительно, учитывая, что $\mathsf{E}(\nu - p)^2 = \mathsf{E}(\nu - \mathsf{E}\nu)^2 = \mathsf{D}\nu$, получаем $\mathsf{E}\Delta_1(\nu, p) \equiv \frac{1}{N}$. Таким образом, функция Δ_1 просто нормирует погрешность на дисперсию оценки. Схожее поведение имеет функция

$$\Delta_2(\bar{R}, R) = -\bar{R} \ln \frac{\bar{R}}{R} - (1 - \bar{R}) \ln \frac{1 - \bar{R}}{1 - R}.$$

Это есть дивергенция между эмпирическим распределением и фактическим распределением, что совпадает (с точностью до знака) с функцией правдоподобия.

Несмотря на то, что приведенные функции выглядят разумными способами выражения погрешности и имеют содержательное обоснование, они обладают неподходящим свойством: когда риск принимает крайние значения (0 и 1), погрешность равна бесконечности при любом значении оценки, отличающемся от истинного. В то же время, если оценка принимает крайние значения, а сама величина — не крайние, то функция погрешности конечна.

С точки зрения свойств, более подходящим вариантом функции погрешности были бы $\Delta_1'(\bar{R}, R) = \Delta_1(R, \bar{R})$ и $\Delta_2'(\bar{R}, R) = \Delta_2(R, \bar{R})$, т.е. функции с переставленными местами аргументами. Последние варианты, правда, фактически «запрещают» оценкам принимать крайние значения, но это приемлемо. Например, если скорректировать частотную оценку, используя $\frac{N\nu+1}{N+2}$, то средняя погрешность имеет приемлемый вид.

В данной работе для сравнения оценок будет использована функция погрешности

$$\Delta_0(\bar{R}, R) = \frac{(\bar{R} - R)^2}{2\bar{R}(1 - \bar{R}) + 2R(1 - R)}.$$

Этот вариант функции погрешности имеет подходящее качественное поведение, что иллюстрирует рис. 4. На правом графике функция гораздо ближе к константной, чем на левом.

Для сравнения вариантов оценки скользящего экзамена будет решаться модельная задача методом дискриминанта Фишера.

Дискриминант Фишера

Дискриминант Фишера использует исключительно метрические свойства конфигурации выборочных точек и не требует не только никаких предположений о распределениях, но и вообще статистической постановки задачи классификации.

Идея дискриминанта Фишера заключается в выборе такого направления в пространстве переменных, при проецировании выборки на которое образы классов оказываются в некотором смысле наиболее удаленными друг от друга. Формально это выражается в максимизации следующего критерия

$$\Phi(w) = \frac{(\tilde{\mu}_{w,1} - \tilde{\mu}_{w,-1})^2}{\tilde{S}_{w,1} + \tilde{S}_{w,-1}},$$

где $\tilde{\mu}_{w,y} = \frac{1}{N_y}\sum_{i=1}^{N} wx^i \cdot I(y^i = y)$ — среднее, а $\tilde{S}_{w,y} = \frac{1}{N_y}\sum_{i=1}^{N}(wx^i - \tilde{\mu}_{w,y})^2 \cdot I(y^i = y)$ — средний квадрат отклонений проекций точек выборки y-го класса на направление w, N_y — число объектов y-го класса в выборке.

Критерий приводится к виду

$$\Phi(w) = \frac{(w\tilde{\mu}_1 - w\tilde{\mu}_{-1})^2}{w^T(\tilde{S}_1 + \tilde{S}_{-1})w} = \frac{w^T(\tilde{\mu}_1 - \tilde{\mu}_{-1})(\tilde{\mu}_1 - \tilde{\mu}_{-1})^T w}{w^T \tilde{S} w},$$

где $\tilde{\mu}_y = \frac{1}{N_y}\sum_{i=1}^{N} x^i \cdot I(y^i = y)$ — среднее точек выборки у-го класса, $\tilde{S}_y$ — выборочная ковариационная матрица y-го класса, $\tilde{S} = \tilde{S}_1 + \tilde{S}_{-1}$.

Последняя форма критерия имеет вид отношения Релея. Известно, что максимум $\Phi(w)$ достигается при $w_\Phi = \tilde{S}^{-1}(\tilde{\mu}_1 - \tilde{\mu}_{-1})$.

Заметим, что выражение для w_Φ очень похоже на выражение для нормали к разделяющей гиперплоскости для случая нормальных распределений с равными матрицами ковариаций. Отличие лишь в том, что во втором случае вместо $\tilde{S}$ используется $\frac{1}{N}(N_1\tilde{S}_1 + + N_{-1}\tilde{S}_{-1})$.

Если количество объектов обоих классов в выборке одинаково, то направления в обоих методах совпадают. Такое сходство приводит к тому, что в литературе эти методы иногда смешиваются, несмотря на их принципиальное различие по подходу и предположениям.

После выбора направления w_Φ задача классификации становится одномерной и может быть достаточно легко решена, например, выбором границы по критерию минимизации эмпирического риска. Более того, в некоторых случаях полученное (проецированием) упорядочивание объектов само по себе может считаться решением, в частности, по нему можно вычислить AUC.

Численный эксперимент

Для сравнения точности различных вариантов скользящего экзамена проведем моделирование на следующем классе распределений. Пусть X — двумерное евклидово пространство. Безусловное распределение в X подчиняется нормальному закону с нулевым средним и единичной ковариационной матрицей. Условная вероятность $\mathsf{P}(y = 1 \,|\, x)$ имеет вид логистической функции $\frac{1}{1+e^{-\alpha w_1 x}}$, где $w_1 = (1, 1)$, а α — параметр, связанный с байесовским уровнем ошибки R_0.

Для иллюстрации на рис. 2 приведена выборка из распределения данного семейства.

Дискриминат Фишера обладает высокой статистической устойчивостью, поскольку требует оценивания по выборке всего лишь порядка n параметров (n — размерность пространства), поэтому моделирование следует проводить при относительно малых объемах выборки.

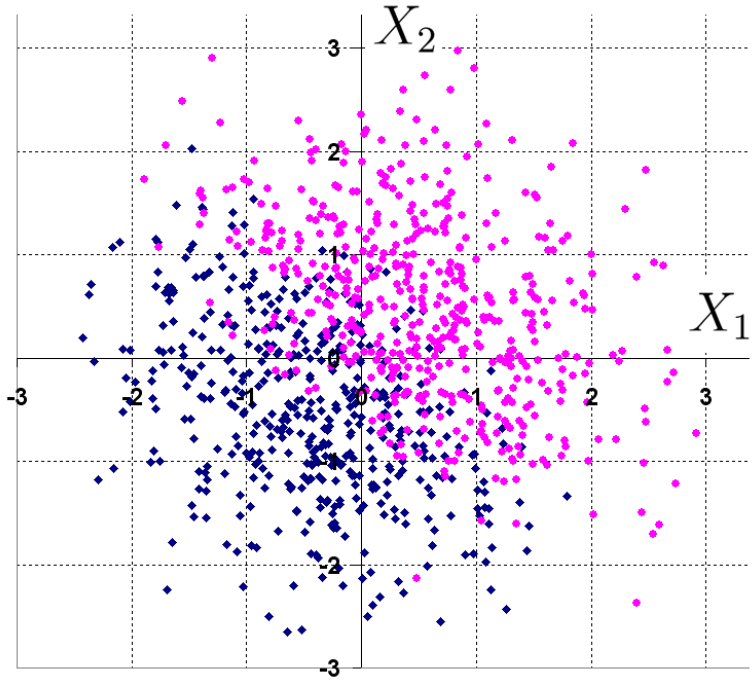


Рис. 2: Выборка из моделируемого распределения

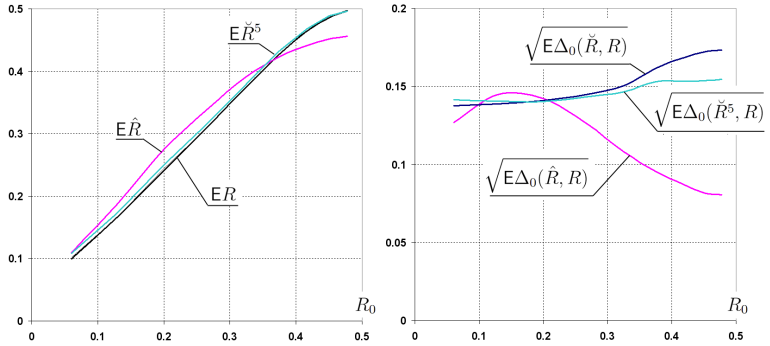


Рис. 3: Средние значения и средние погрешности различных оценок риска

На рис. 3 приведены результаты моделирования при $N = 20$. На левой диаграмме приведены средние (усреднение по 10000 выборкам) значения риска и его оценок в зависимости от байесовского уровня ошибки. Среднее значение оценки leave-one-out практически не отличается от среднего риска, поэтому эта кривая на данной диаграмме не изображена. На правой диаграмме приведены средние погрешности оценок риска.

В проводимое сравнение включена также $\hat{R}$ — оценка риска, получаемая добавлением к эмпирическому риску оценки максимального смещения [6]. Данная оценка подробно описана в [7].

Видим, что оценки leave-one-out и 5-fold cross-validation получились практически равноценными по качеству. Оценка $\hat{R}$ выглядит более предпочтительной при больших значениях R_0. Результаты моделирования при $N = 30$ и $N = 50$ качественно не отличались от приведенных.

Аналогичное исследование проводилось для гистограммного классификатора. В этом случае вычисления производились точно (не на основе статистического моделирования),

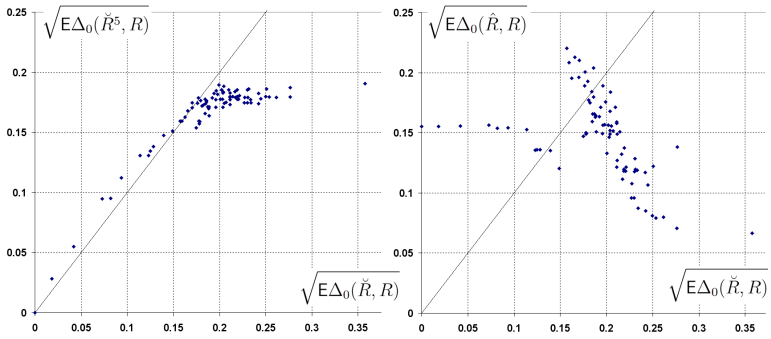


Рис. 4: Средние погрешности некоторых оценок риска для гистограммного классификатора при различных распределениях

однако ввиду экспоненциальной сложности численного суммирования результаты получены при малых размерностях: число «ячеек» равно 3, $N = 12$, $K = 4$.

На рис. 4 каждая точка соответствует некоторому распределению в дискретном пространстве. Распределения перебирались «по сетке» с достаточно большим шагом (ввиду экспоненциальной сложности). Полученный результат качественно согласуется с полученным для дискриминанта Фишера.

Заключение

В отсутствие теоретических результатов, которые бы давали точные оценки качества, важным инструментом исследования и сравнения оценок является статистическое моделирование. Хотя получаемые таким образом результаты носят частный характер, они дают возможность эмпирически делать определенные выводы.

Исследование, проведенное в данной работе, свидетельствует о практической равноценности оценок leave-one-out и 5-fold cross-validation. Хотя первая оценка в большинстве испытаний оказывается слегка точнее, в ряде случаев она дает существенно большую погрешность. На практике этот эффект, однако, вряд ли является существенным.

Также можно сделать вывод о перспективности оценки на основе эмпирического риска с поправкой на смещение. Данная оценка имеет сравнимую точность, но гораздо меньшую трудоемкость в вычислительном плане.

Литература

[1] *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск: Институт математики СО РАН, 1999. 211 с.

[2] *Langford J.* Quantitatively tight sample complexity bounds. Carnegie Mellon Thesis. 2002. `http://citeseer.ist.psu.edu/langford02quantitatively.html`. 130 p.

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

[4] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. 2008. Vol. 18, no. 2. Pp. 243–259.

[5] *Efron B., Tibshirani R.* Improvements on cross-validation: The .632+ Bootstrap method // J. Amer. Stat. Association. 1997. Vol. 92, no. 438. Pp. 548–560.

[6] *Неделько В. М.* Об интервальном оценивании риска для решающей функции // Таврический вестник информатики и математики. Изд-во НАН Украины. 2008. № 2. С. 97–103.

[7] *Недельko В. М.* Точные и эмпирические оценки вероятности ошибочной классификации // Научный вестник НГТУ. 2011. № 1 (42). С. 3–16.

Использование спектрального преобразования для распознавания напечатанного изображения*

Пушняков А. С.
aleksey.pushnyakov@phystech.edu
Московский физико-технический институт

Аннотация. В данной работе решается задача классификации двух типов изображений глаз: реально сфотографированного и впоследствии распечатанного. Используется метод спектрального преобразования изображения. В напечатанном изображении предполагается обнаружить периодическую структуру, которая порождает дополнительную гармонику в спектре. Рассматривается радиальная составляющая фурье-спектра, и по ней строится пространство признаков. Задача классификации решается с помощью метрического классификатора.

Ключевые слова: *iris recognition, дискретное преобразование Фурье, метрический классификатор.*

Recognition of printed image using spectral transform*

Pushnyakov A. S.
Moscow Institute of Physics and Technology

This paper is about a problem of classification real and printed eye-image. A spectral transform of image is used. We proposed to find a periodical structure that generates high frequency spectral magnitude. A feature space is based on radial component of the Fourier image. The problem of classification is solved by metric classifier.

Keywords: *iris recognition, discrete Fourier transform, metric classifier.*

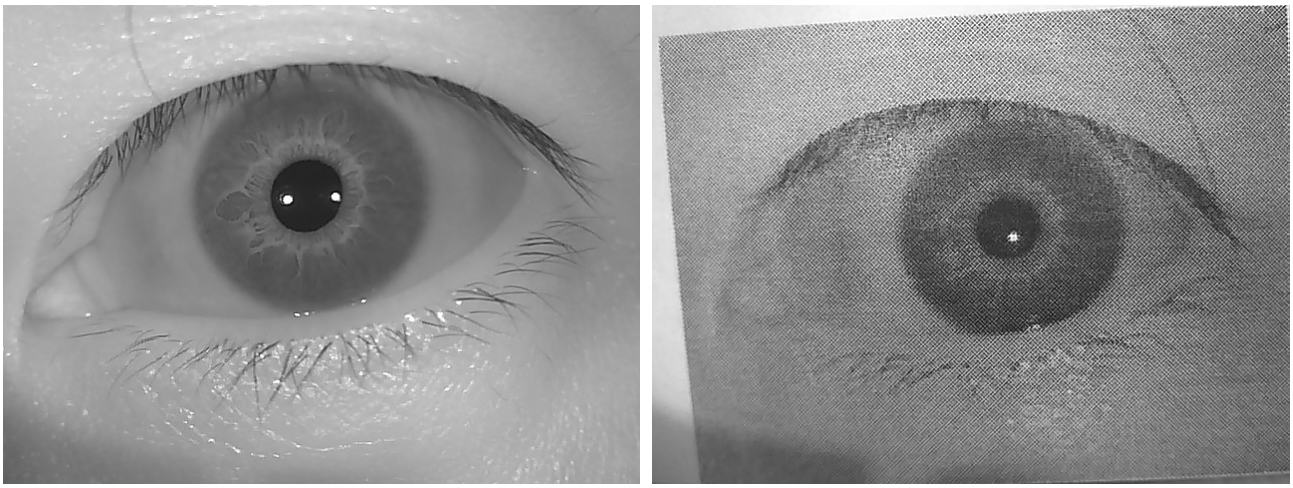
Введение

Задача классификации напечатанных и реальных изображении возникает в более общей задаче распознавания человека по радужной оболочке глаза [1, 2]. Изображение человеческого глаза может быть подделано. Одним из возможных способов такой фальсификации является фотографирование глаза и последующая печать изображения. Возникает вопрос, по каким критериям можно отличать два типа изображений при непосредственном сканировании. Предполагается, что при печати изображения появляется периодический шум высокой частоты (рис. 1). Данную частоту можно выделить с помощью спектрального преобразования изображения.

Спектральные преобразования используются в различных типах фильтрации изображений, при сжатии изображений с минимальными потерями качества и в других аспектах обработки изображений [3, 4]. В нашем случае предлагается использовать дискретное преобразование Фурье [3, 5]. Одним из примеров использования преобразования Фурье является задача распознавания изображений с учетом инвариантности относительно поворотов и масштабирования [6, 7].

Проблема фальсификации изображений при распознавании человека по радужке рассматривается в [8]. В числе прочего в это работе поднимается проблема классификации

Научные руководители И.А. Матвеев, В.В. Стрижов



(а) (б)

Рис. 1: (а) сфотографированное изображение глаза; (б) напечатанное изображение глаза

напечатанного и реального изображения. Некоторые решения, основанные на спектральном анализе, предложены в работах [9, 10].

В нашем случае коэффициенты Фурье составляют пространство первичных признаков. Используя эти признаки, предлагается классифицировать два выше описанных типа изображений. Так как размерность пространства первичных признаков велика, предлагается построить новые признаки которые уже будут непосредственно использоваться для классификации. Для построения новых (вторичных) признаков используется радиальная составляющая фурье-образа. Во-первых, такой подход может на порядок сократить размерность задачи классификации. Во-вторых, в силу того, что предполагаемая выделяющаяся гармоника имеет приблизительно равные пространственные частоты, то при переходе к радиальной компоненте, эти гармоники должны наложиться друг на друга. В итоге задача сводится к задаче классификации, которую, например, можно решить с помощью метрического классификатора [11], в частности, используя метод Парзеновского окна [12].

В работе также проводится оценка сложности предлагаемого алгоритма и проводится вычислительный эксперимент, проверяющий быстродействие и качество классификатора.

Постановка задачи

Рассмотрим наше изображение как сеточную функцию $f(x, y)$, где $x = 0, 1, \ldots M - 1$, $y = 0, 1, \ldots N - 1$. Дискретное преобразование Фурье можно записать в следующем виде

$$\varphi(u, v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-2\pi i \left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{1}$$

Элементы матрицы $\Phi = \|\varphi(u, v)\|_{u=0,v=0}^{M-1,N-1}$ — суть пространство первичных признаков. Под вторичным признаком понимается, некоторая функция $g(\Phi)$. Задача состоит в том, чтобы подобрать достаточно небольшой набор вторичных признаков $\boldsymbol{g} = \{g_k\}_{k=1}^{I}$, которые бы отражали распределение плотности спектра Φ (при этом предполагается, что I значительно меньше, чем MN). В частности, нас интересует наличие максимумов в области высоких частот. Наличие таких максимумов говорит о наличие периодического шума высокой частоты, который наблюдается в поддельных напечатанных изображениях.

Для классификации изображений используется метрический классификатор следующего вида

$$a(u, X_+, X_-) = \operatorname{sign}\left(W(u, X_+) - W(u, X_-)\right), \tag{2}$$

где u — классифицируемые изображение, X_+ и X_- — обучающие выборки настоящих и поддельных изображения соответственно, функция $W(u, X)$ определяет принадлежность объекта u к классу X.

Описание алгоритма

Пусть наше изображение $f(x, y)$ преобразовано согласно (1) в $\varphi(u, v)$ с помощью быстрого преобразования Фурье [13].

Первым шагом введем понятие радиальной компоненты фурье-образа Ф. Отметим, что нулевая гармоника спектра в нашем представлении есть $\varphi(0, 0)$ (малые частоты обычно самые интенсивные). Поэтому точку $(0, 0)$ примем за полюс. Далее сгруппируем все точки (u, v) по евклидовому расстояния до $(0, 0)$ с учетом тороидального замыкания сетки. Обозначим $\hat{u} = \min\{u, M - 1 - u\}$, $\hat{v} = \min\{v, N - 1 - v\}$ и рассмотрим множество $S(r) = \{(u, v) : \ r^2 \leqslant \hat{u}^2 + \hat{v}^2 < (r + 1)^2\}$. Здесь также учтено, что мы хотим иметь целые значения r. Тогда можно ввести радиальную компоненту фурье-образа как

$$R(r) = \frac{1}{|S(r)|} \sum_{(u,v)\in S(r)} |\varphi(u, v)| \tag{3}$$

При этом $r \leqslant r_{max} = \frac{1}{2}\sqrt{M^2 + N^2}$.

Теперь опишем процедуру получения вторичных признаков $\{g_k\}_{k=1}^{I}$. Для этого введем понятие интегрального признака для радиальной компоненты фурье-образа. Интегральным признаком $\theta(\alpha)$, $0 < \alpha < 1$ для $R(r)$ назовем следующую величину

$$\theta(\alpha) = \max\left\{\theta : \sum_{r=0}^{\theta} R(r) \leqslant \alpha \sum_{r=0}^{r_{max}} R(r)\right\} \tag{4}$$

Фактически, интегральный признак показывает то значение аргумента r, при котором площадь подграфика спектра есть доля α от общей площади подграфика. Основная идея состоит в том, что данные интегральные признаки отображают распределение плотности спектра. Поэтому для спектров настоящих изображений, которые почти монотонно убывают, и спектров поддельных изображений, которые имеют четко выделенные пики высокой частоты, эти характеристики должны значительно отличаться. Однако априори не понятно, какие значения α нужно выбирать. Также нужно учесть, что значительная часть плотности сосредоточена вблизи нулевых частот.

Предлагается построить последовательность α_k такую, чтобы последовательность $\delta_k = \alpha_k - \alpha_{k-1}$ достаточно быстро убывала. Положим $\delta_k = 2^{-k}$, $\alpha_0 = 0$ и соответственно $\alpha_k = \sum_{j=1}^{k} 2^{-j} = 1 - 2^{-k}$ (вообще говоря, можно брать δ_k как любую бесконечно убывающую геометрическую прогрессию). В силу дискретности задачи существует такой номер I, что $\theta(\alpha_I) = r_{max}$ и на этом последовательность обрывается. Тогда определим $g_k = \theta_k - \theta_{k-1}$, $k = 1, \ldots I$. Таким образом, вторичные признаки g_k есть скачки интегральных признаков. А именно: при изменении аргумента с θ_k на величину g_k значение площади подграфика увеличивается на долю δ_k от общей площади. Большое значение g_k указывает на наличие пика в спектре. Отметим, что, вообще говоря, I зависит от исследуемого

изображения, но последовательность θ_k можно продолжить стационарным образом, поэтому достаточно для всех изображений установить единый порог I_*. Например, его можно выбрать как максимальное значение I на обучающей выборке.

Также отметим, что можно каждый раз не пересчитывать сумму $\sum_{r=0}^{\theta} R(r)$, а определять $\theta_k = \theta(\alpha_k)$ следующим образом

$$\theta_k = \max\left\{\theta : \sum_{r=\theta_{k-1}}^{\theta} R(r) \leqslant \frac{1}{2}\sum_{r=\theta_{k-1}}^{r_{max}} R(r)\right\} \tag{5}$$

В качестве метрики на наборах $\boldsymbol{g}$ выбирается евклидова метрика в $\mathbb{R}^I$

$$\rho(\boldsymbol{g}_1, \boldsymbol{g}_2) = \left(\sum_{i=0}^{I} |g_{1i} - g_{2i}|^2\right)^{\frac{1}{2}}, \tag{6}$$

Для классификации используется метод Парзеновского окна с экспоненциальным ядром $K(x) = e^{-x^2}$. Ширина окна h выбирается равной максимальному расстоянию между классифицируемым объектом u и элементом обучающей выборки $x \in X_+ \cup X_-$

$$h(u) = \max_{x \in X_+ \cup X_-} \hat{\rho}(u, x) \tag{7}$$

где $\hat{\rho}(u, x) = \rho(\boldsymbol{g}(u), \boldsymbol{g}(x))$ определяется из (6). Отсюда получаем следующую функцию принадлежности объекта u классу X

$$W(u, X) = \frac{1}{|X|}\sum_{x \in X} K\left(\frac{\hat{\rho}(u, x)}{h(u)}\right), \tag{8}$$

и классификатор принимает следующий вид

$$a(u, X_+, X_-) = \operatorname{sign}\left(\frac{1}{|X_+|}\sum_{x \in X_+} K\left(\frac{\hat{\rho}(u, x)}{h(u)}\right) - \frac{1}{|X_-|}\sum_{x \in X_-} K\left(\frac{\hat{\rho}(u, x)}{h(u)}\right)\right) \tag{9}$$

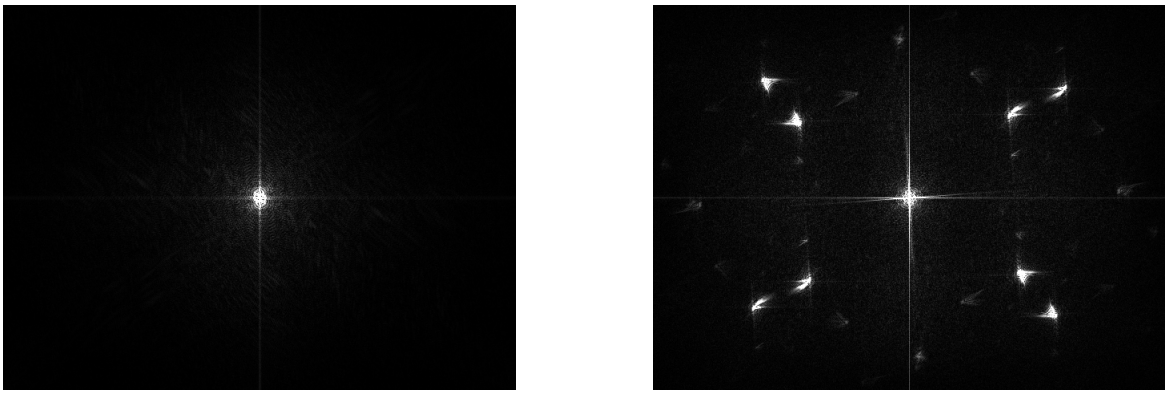
Оценка сложности алгоритма

Покажем что, время работы работы как классификатора, так и обучающего алгоритма определяется временем работы спектрального преобразования. Сложность быстрого преобразования Фурье есть $O(L \log L)$, где $L = NM$ — количество пикселов в изображении. В это время сложность остальной части алгоритма является линейной по L. Действительно, вычисление радиальной компоненты согласно формуле (3) выполняется за не более чем $3MN$ арифметических операций (с учетом разбиения всех точек на классы $S(r)$). Построение каждого интегрального признака θ_k и вместе с ним g_k не более чем за $2r_{max} = \sqrt{M^2 + N^2}$ арифметических операций (радиальная компонента $R(r) = 0$, при $r > r_{max}$). При этом количество признаков I фиксировано и мало по сравнению с L. Сложность самой классификации определяется только размером обучающей выборки l и количеством признаков I (также l значительно меньше L). Таким образом, общую сложность алгоритма можно оценить сверху как $O(L \log L)$.

Вычислительный эксперимент

В вычислительном эксперименте использовались изображения в формате jpeg с разрешением 640×480, то есть в формуле (1) нужно положить $N = 640$, $M = 480$. Изображения были разделены на 4 группы изображений: две группы использовались как обучающая выборка, две другие — контрольная выборка. В обучающей выборке было использовано 1000 реальных и 1000 поддельных изображений, в двух контрольных выборках было соответственно по 3000 изображений. Также классификация была проведена на меньших размерах выборок (порядка 100 изображений)

Как и ожидалось, наличие периодического шума приводи к сильному различию спектров полученных после дискретного преобразования Фурье согласно формуле (1). На рис. 2 изображен результат преобразования Фурье для изображения глаз на рис. 1.



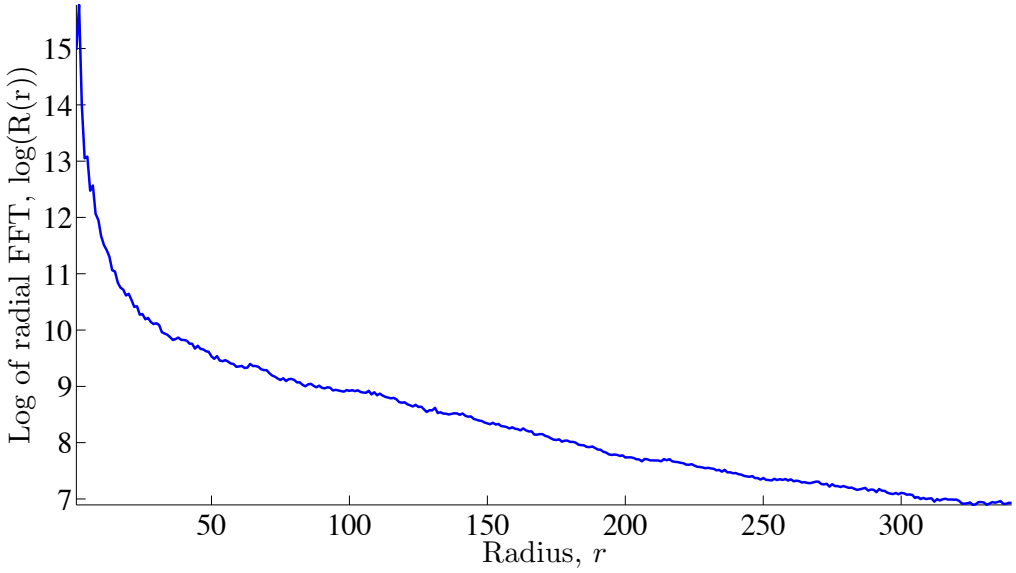
(а) реальное изображение (б) напечатанное изображение

Рис. 2: Двумерный фурье-образ изображения

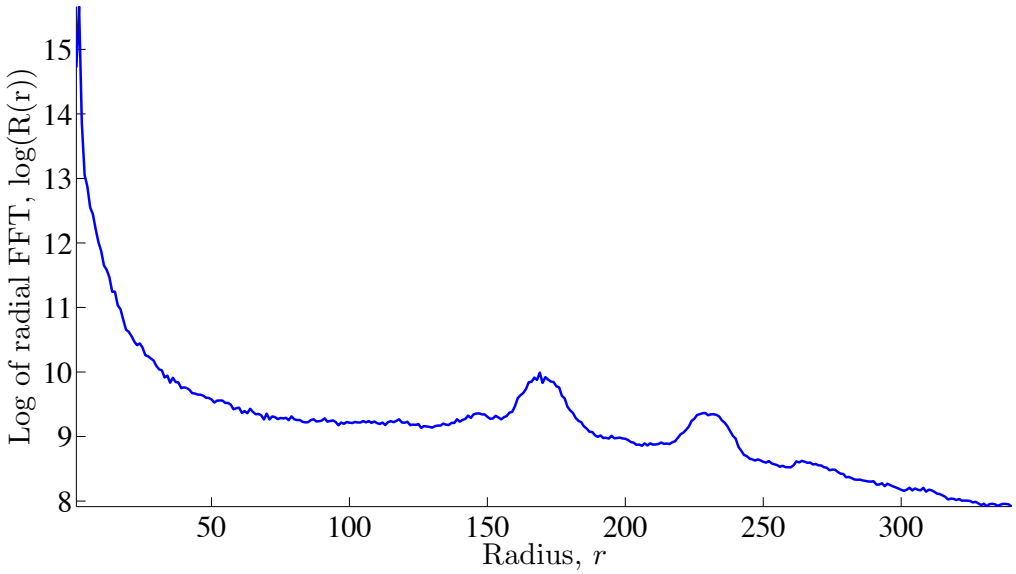
Как видно из последнего рисунка, в спектре напечатанного изображения есть 8 побочных максимумов; причем есть две группы по 4 максимума, находящиеся на приблизительно одинаковом расстоянии от нулевой частоты. При построении радиальной компоненты фурье-образа по формуле (3) мы получаем два пика у спектра напечатанного изображения. Для приведенных выше спектров радиальные компоненты приведены в логарифмическом масштабе на рис. 3.

Как показал эксперимент, для выявлений различий достаточно выбирать очень небольшое количество признаков. Во-первых, почти для всех изображений значение I (номер интегрального признака, после которого последовательность θ_k стабилизируется) не превышал 10. Во-вторых, значительные различия во вторичных признаках g_k возникают уже при $k < 5$. Поэтому в эксперименте полагалось $I = 10$. График значений g_k для тестовой обучающей выборки из 20 изображений показан на рис. 4. Каждая ломанная соответствует набору признаков одного изображения. Резкий пик для напечатанных изображений (красный цвет) при $k = 3 \div 4$ указывет на наличие побочного максимума в спектре.

Экспериментально была проверена скорость работы алгоритма. При этом, оказалось, что она почти не зависит от размера обучающей выборки. Так в основном эксперименте (обучающая выборка 2000 изображений) время в секундах работы классификатора в основном эксперименте (6000 изображений) составило $T_1 = 1704$, т. е. время обработки одного изображения $t_1 = 0{,}284$, а для обучающей выборки из 200 изображений $t_2 = 0{,}235$.



(а) реальное изображение



(б) напечатанное изображение

Рис. 3: Радиальная компонента фурье-образа

Время работы обучающего алгоритма на количество элементов выборки составило $\tau_1 = 0{,}241$, $\tau_2 = 0{,}225$ для выборок размера 4000 и 200 соответственно. Зависимость времени работы классификатора (на одном изображении) от размера длины обучающей выборки l показана на рис. 5. Из рисунка время работы вначале растет, а потом флуктуирует около значения $\hat{t} = 0{,}265$. Эти результаты согласуются с тем, что время работы алгоритма определяется сложностью спектрального преобразования и почти не зависит от размеров обучающей выборки.

Основной причиной возникновения ошибок при работе классификатора являлось отсутствие фокусировки и значительного числа изображений. Это связано с тем, что изображения получались сериями (съемка камерой), среди которых лишь одно или два изоб-

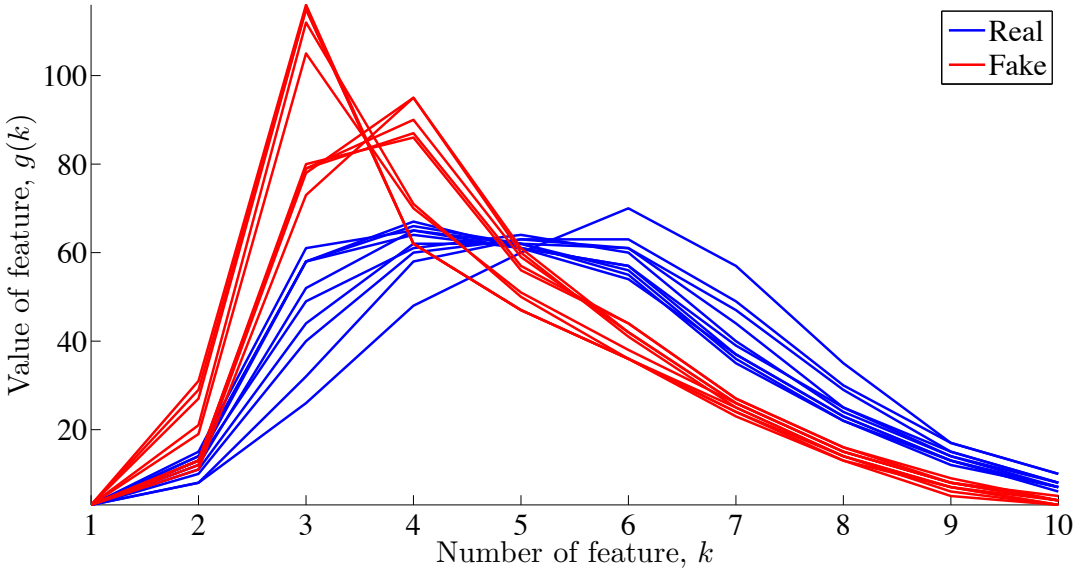


Рис. 4: Значения признаков g_k на тестовой обучающей выборке

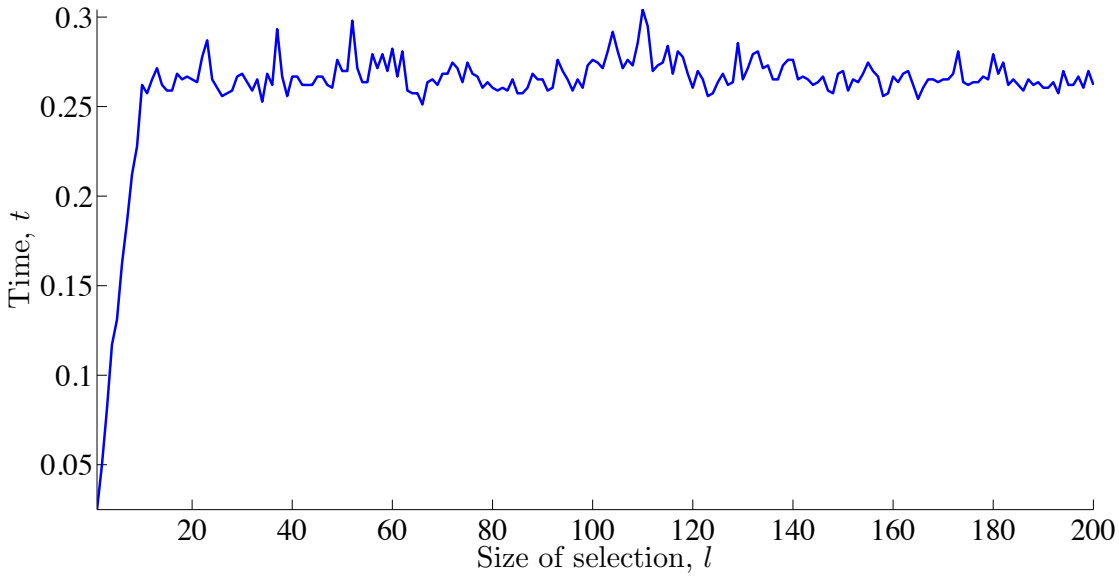


Рис. 5: Зависимость времени работы классификатора от размера обучающей выборки

ражения попадают в фокус. На хорошо сфокусированных фотографиях и при большой обучающей выборке (более 100 изображений) доля ошибок не превосходила $4-6\%$ (это относится к ошибочным классификациям как поддельных так и настоящих изображений). В общем случае при длине обучающей выборки более 100 изображений и наличии в ней плохо сфокусированных изображений доля ошибок не превосходила $10-15\%$. Однако на малых выборках (менее 50 изображений) доля ошибок второго рода (поддельное изображение распознается как настоящее) доходила до $25-30\%$. Возможными решения данной проблемы является предварительная фильтрация изображений, а также обработка изображений одного глаза сериями. В последнем случае серия определяется как напечатанная, если хотя бы на одно из изображений в серии определено как напечатанное.

Заключение

Предложен алгоритм классификации напечатанных и реальных изображений глаз, использующий дискретное преобразование Фурье. В качестве признаков для квалификации используются интегральные характеристики радиальной компоненты фурье-образа. Проведен вычислительный эксперимент, показывающий быстродействие алгоритма, достаточное для реального применения. Для повышения эффективности и надежности работы алгоритма возможна оптимизация метрического классификатора, например, с помощью выделения эталонных объектов в обучающей выборке.

Литература

[1] *Daugman J.*. How Iris Recognition Works // *IEEE Transactions on Circuits and Systems for Video Technology*. 2002. Vol. 14. Pp. 21–30.

[2] *Wildes R.P.* Iris recognition: an emerging biometric technology // *Proceedings of the IEEE*. Vol. 85, №9. Pp. 1348-1363

[3] *Гонсалес, Р and Вудс, Р.* Мир цифровой обработки Цифровая обработка изображений. ТЕХНО-СФЕРА. 2005.

[4] *Сорока Е. З., Хлбородов В. А. and Хуанг Т.* Обработка изображений и цифровая фильтрация. Мир. 1979.

[5] *Форсайт Д., Понс Ж.* Компьютерное зрение. Современный подход. ИД Вильямс М. 2004.

[6] *Hsu et Al.* Rotation-invariant digital pattern recognition using circular harmonic expansion // *Applied Optics*. 1982. Vol. 21, №22. Pp. 4012–4015.

[7] *Reddy B. S., Chatterji B. N.* An FFT-based technique for translation, rotation, and scale-invariant image registration // *Image Processing, IEEE Transactions on*. 1996. Vol. 5, №8. Pp. 1266–1271.

[8] *Daugman J.*. Recognising persons by their iris patterns // *Advances in Biometric Person Authentication*. 2005. Pp. 783–814.

[9] *He et Al.* A fake iris detection method based on fft and quality assessment // *Pattern Recognition, CCPR'08. Chinese Conference on*. 2008. Pp. 1–4.

[10] *He et Al.*, A new fake iris detection method // *Advances in Biometrics*. 2009. Pp. 1132–1139.

[11] *Ту, Дж. и др.* Принципы распознавания образов: Пер. с англ. Мир. 1978.

[12] *Pascal V., Yoshua B.* Manifold parzen windows // *Advances in Neural Information Processing Systems*. 2002. Vol. 15. Pp. 825–832.

[13] *Johnson, Steven G., Frigo M.* A modified split-radix FFT with fewer arithmetic operations // *Signal Processing, IEEE Transactions on*. 2007. Vol. 55, №1. Pp. 111–119.

Цифровые изображения на комплексном дискретном торе*

Мнухин В. Б.
mnukhin.valeriy@mail.ru
Южный Федеральный Университет

В работе предлагается алгебраический метод обработки цифровых изображений. Предлагается рассматривать изображения размера $p \times p$, где $p \geqslant 3$ — простое число вида $p = 4k-1$, как функции на комплексном дискретном торе. Вводится понятие комплексного вращения и определяется новое обратимое преобразование, являющееся дискретным аналогом непрерывного преобразования Меллина. Строится модулярное преобразование Фурье-Меллина, инвариантное относительно циклических сдвигов, масштабирования и комплексных вращений цифровых изображений.

Ключевые слова: *Цифровое изображение, конечное поле, дискретный тор, дискретное преобразование, преобразование Фурье-Меллина, полярно-логарифмические координаты, модулярный логарифм.*

Digital images on a complex discrete torus*

Mnukhin V. B.
Southern Federal University

An algebraic method for digital images analysis and processing is considered. It is based on the presentation of digital images of size $p \times p$, where $p \geqslant 3$ is the prime number such that $p = 4k - -1$, as functions on a complex discrete torus. For such images, complex rotations are introduced and a new invertible linear transform, called the modular Mellin transform, is presented. The discrete modular Fourier–Mellin transform, invariant under circular shifts, scaling, and complex rotations, is also defined.

Keywords: *Digital image, finite field, discrete torus, discrete transform, Fourier-Mellin transform, log-polar coordinates, modular logarithm.*

Введение

В настоящее время разработка методов решения задач обработки, распознавания и классификации изображений проводится, как правило, в предположении непрерывности заданного изображения. Это позволяет использовать для решения поставленных задач мощный аппарат математического анализа и интегральных преобразований. Однако на практике применение полученных таким образом решений к цифровым изображениям неизбежно приводит к появлению систематических ошибок, связанных с дискретностью реальных цифровых изображений.

В качестве одного из примеров подобной ситуации укажем на хорошо известный метод преобразования Фурье–Меллина [1–3], активно используемый, в частности, при решении задач регистрации изображений [4, с. 674]. Метод основан на рассмотрении Фурье-спектра исходного изображения в полярно-логарифмической системе координат, после чего к преобразованному таким образом спектру вторично применяется преобразование Фу-

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-07-00591 и № 13-07-00327.

рье. Нетрудно показать, что в *непрерывном случае* на базе такого преобразования легко определяются характеристики изображений, инвариантные относительно сдвигов, вращений и масштабирования. Вместе с тем в *дискретном* случае формальное использование этого метода сопряжено с существенными погрешностями [2, 5], связанными, в частности, с невозможностью адекватного преобразования цифровых изображений в полярную и полярно-логарифмическую системы координат.

В связи с этим возникает задача разработки методов, изначально ориентированных на цифровые изображения и использующих аппарат алгебры и теории чисел. В данной работе цифровые изображения размера $p \times p$, где $p \geqslant 3$ — простое число вида $p = 4k - 1$, предлагается рассматривать как функции на комплексных дискретных торах, которые можно считать «конечными комплексными плоскостями». Для таких изображений вводится понятие комплексного вращения, и строятся дискретные аналоги преобразований Меллина и Фурье–Меллина, позволяющие получать инварианты цифровых изображений относительно циклических сдвигов, масштабирования и комплексных вращений. Сохранение ортогональности при комплексных вращениях позволяет эффективно использовать их, в частности, в задачах анализа симметрии цифровых изображений [6, 7].

Комплексный дискретный тор

Кольцо классов вычетов по модулю целого $n > 1$ условимся обозначать как $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$.

Определение 1. *Пусть $m > 1$ и $n > 1$ — два целых числа. Множество*

$$\mathbb{Z}_m \times \mathbb{Z}_n = \big\{(x, y) \ : \ x \in \mathbb{Z}_m, \ y \in \mathbb{Z}_n\big\}$$

будем называть дискретным $m \times n$-тором. Пары $(x, y) \in \mathbb{Z}_m \times \mathbb{Z}_n$ будем называть пикселами, а классы вычетов x и y — координатами данного пиксела. Рисунок 1 иллюстрирует введенное понятие.

Очевидно, что любой тор $\mathbb{Z}_m \times \mathbb{Z}_n$ является кольцом относительно естественных операций $+$ и $\odot$ покомпонентного сложения и умножения: если $(a, b) \in \mathbb{Z}_m \times \mathbb{Z}_n$ и $(c, d) \in \mathbb{Z}_m \times \mathbb{Z}_n$, то

$$(a, b) + (c, d) = \big(a \underset{m}{+} c, \ b \underset{n}{+} d\big), \qquad (a, b) \odot (c, d) = \big(a \underset{m}{\cdot} c, \ b \underset{n}{\cdot} d\big).$$

Вместе с тем нетрудно заметить, что некоторые торы обладают особыми свойствами. Например, для простого p тор $\mathbb{Z}_p^2 = \mathbb{Z}_p \times \mathbb{Z}_p$, является векторным пространством размерности 2 над конечным полем $\mathbb{Z}_p$.

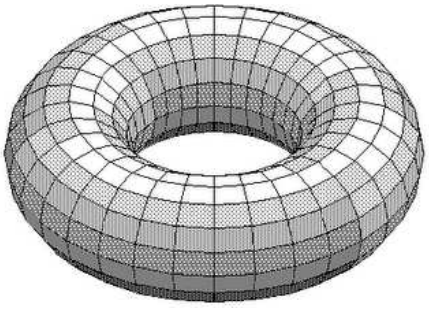


Рис. 1: Дискретный тор

Покажем, что в некоторых случаях на торе $\mathbb{Z}_p^2$ можно ввести структуру конечного поля со свойствами, аналогичными свойствам поля комплексных чисел $\mathbb{C}$. Вспоминая [8, с. 438]

построение кольца Гауссовых целых чисел $\mathbb{Z}\big[\sqrt{-1}\,\big]$, рассмотрим факторкольцо

$$\mathbb{C}_p = \mathbb{Z}_p[x]/(x^2+\overline{1})$$

кольца многочленов над конечным полем $\mathbb{Z}_p$ по главному идеалу, порожденному двучленом $x^2+\overline{1}$. Как очевидно, $\mathbb{C}_p$ содержит p^2 элементов вида $a+bi$, где $a, b \in \mathbb{Z}_p$, а i обозначает класс вычетов $\overline{x}$, так что $i^2 = \overline{-1} = \overline{p-1} \in \mathbb{Z}_p$.

Допустим, что простое число $p \geqslant 3$ имеет вид $p = 4k-1$, (где $k \in \mathbb{Z}$), или, другими словами, что $p \equiv 3 \bmod 4$. (Таковы, в частности, числа 71, 127, 251, 257 и 509.) Тогда, как известно ([9], с. 176), в поле $\mathbb{Z}_p$ элемент $\overline{-1} \in \mathbb{Z}_p$ не является квадратом, так что двучлен $x^2+\overline{1}$ неприводим в кольце $\mathbb{Z}_p[x]$. Следовательно, кольцо $\mathbb{C}_p$ является полем [10, с. 293], и, согласно фундаментальной теореме [8, с. 428] теории конечных полей, $\mathbb{C}_p \simeq \mathbb{GF}(p^2)$. Остается заметить, что естественное отображение

$$\mathbb{C}_p \ni a+ib \quad \longleftrightarrow \quad (a,b) \in \mathbb{Z}_p^2$$

переносит структуру поля на тор $\mathbb{Z}_p^2$, превращая его в «конечную комплексную плоскость».

Определение 2. *Если $p \geqslant 3$ — простое число такое, что $p \equiv 3 \bmod 4$, то дискретный тор $\mathbb{Z}_p^2$ со структурой введенного выше конечного поля, будем называть комплексным дискретным тором (КДТ) и обозначать $\mathbb{C}_p$.*

(Мы постараемся избегать словосочетания «дискретный комплексный тор», поскольку понятие «комплексный тор» хорошо известно в математике и используется в ином смысле.)

Учитывая аналогию $\mathbb{C}_p$ с комплексным полем $\mathbb{C}$ и кольцом $\mathbb{Z}\big[\sqrt{-1}\,\big]$, естественно называть элементы конечного поля $\mathbb{C}_p$ *дискретными комплексными числами* и использовать при работе с ними стандартную терминологию комплексного анализа. В частности, если $z = a+ib \in \mathbb{C}_p$, то будем писать $\mathrm{Re}(z) = a$ и $\mathrm{Im}(z) = b$. Число $z^* = a-ib$ будем называть *сопряженным* к $z = a+ib$, а величину $N(z) = zz^* = a^2+b^2 \in \mathbb{Z}_p$ называть *нормой* числа $z \in \mathbb{C}(p)$. (Отметим, что в отличие от комплексных чисел модуль $|z| = \sqrt{N(z)}$, вообще говоря, в $\mathbb{C}_p$ не определен.) Непосредственно проверяется, что $N(z_1z_2) = N(z_1)N(z_2)$; таким образом, норма является мультипликативной функцией. Кроме того, заметим, что

$$N(z) = 0 \;\Leftrightarrow\; z = 0.$$

(Действительно, если $N(z) = a^2+b^2 = 0$ и $b \neq 0$, то $(a/b)^2 = -1$, что невозможно по предположению.)

Условимся далее считать, что число p удовлетворяет условию определения 2. Кроме того, там, где это не может привести к недопониманию, условимся отождествлять классы вычетов поля $\mathbb{Z}_p$ с их представителями и опускать верхнюю черту над обозначениями классов.

Отметим, что дискретные комплексные числа допускают также матричное представление. Пусть $M_2(\mathbb{Z}_p)$ означает алгебру 2×2-матриц над полем $\mathbb{Z}_p$. Рассмотрим следующее инъективное отображение $\mu : \mathbb{C}_p \to M_2(\mathbb{Z}_p)$:

$$\mu(a+bi) = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}, \quad \text{и пусть} \quad P = \mu\big(\mathbb{C}_p\big) = \left\{ \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \;:\; a,b \in \mathbb{Z}_p \right\} \subset M_2(\mathbb{Z}_p).$$

Утверждение 1. *Биекция $\mu : \mathbb{C}_p \to P$ является изоморфизмом,*

$$\mu(z+w) = \mu(z)+\mu(w), \qquad \mu(zw) = \mu(z)\mu(w) \qquad \textit{для всех} \quad z,w \in \mathbb{C}_p.$$

При этом нормы дискретных комплексных чисел переходят в определители соответствующих матриц, $N(z) = \det \mu(z)$.

Доказательство полностью повторяет доказательство [8, с. 195] аналогичного утверждения для комплексных чисел.

Комплексные вращения цифровых изображений

Под *цифровым изображением* размера $m \times n$ условимся понимать комплекснозначную функцию $f(x, y) : \mathbb{Z}_m \times \mathbb{Z}_n \to \mathbb{C}$, определенную на всех пикселах тора $\mathbb{Z}_m \times \mathbb{Z}_n$. (Заметим, что такое определение позволяет считать изображениями и Фурье-образы. Поскольку тор является естественной областью определения дискретного преобразования Фурье [11, гл. 19], такой подход представляется полезным и будет использован ниже.)

На всяком торе $\mathbb{Z}_p^2$ определены следующие обратимые преобразования изображений:

— *циклический сдвиг* $\mathcal{T}_w$ на $w = (a, b) \in \mathbb{Z}_p^2$:

$$\mathcal{T}_w[f] = f(x - a, y - b);$$

— *масштабирование* $\mathcal{S}_a$ с коэффициентом $0 \neq a \in \mathbb{Z}_p$:

$$\mathcal{S}_a[f] = f(ax, ay);$$

— *отражения* $\mathcal{M}_x$ и $\mathcal{M}_y$:

$$\mathcal{M}_x[f] = f(-x, y), \qquad \mathcal{M}_y[f] = f(x, -y).$$

В то же время задача введения для цифровых изображений математически корректного понятия *вращения* представляется нетривиальной. Обычно в этом качестве используют ([4], с. 390) преобразования вида

$$\widetilde{\mathscr{R}}_\alpha[f] = f\big([x\cos\alpha - y\sin\alpha],\ \ [x\sin\alpha + y\cos\alpha]\big)\ ,$$

где $[a]$ означает округление действительного числа a до ближайшего целого. К сожалению, преобразование $\widetilde{\mathscr{R}}_\alpha$ не коммутирует с дискретным преобразованием Фурье, что приводит к систематическим ошибкам при его использовании. Аналогичные проблемы возникают [5] при попытках введения на цифровых изображениях полярной и полярно-логарифмической систем координат.

Рассмотрим теперь цифровые изображения на комплексном дискретном торе. Заметим, что вращение непрерывной функции комплексного переменного $f(z)$ на угол α в плоскости $\mathbb{C}$ естественно определяется как

$$\mathscr{R}_\alpha[f] = f\left(e^{i\alpha}z\right).$$

Поэтому, рассматривая изображения на КДТ как функции $f(z) : \mathbb{C}_p \to \mathbb{C}$ «дискретного комплексного переменного», для всякого ненулевого $w \in \mathbb{C}_p$ введем следующее обратимое преобразование изображения $f(z)$:

$$\mathscr{R}_w[f] = f(wz), \qquad \mathscr{R}_w^{-1}[f] = f\left(w^{-1}z\right). \tag{1}$$

Определение 3. *Преобразование $\mathscr{R}_w[f]$ будем называть комплексным вращением цифрового изображения f.*

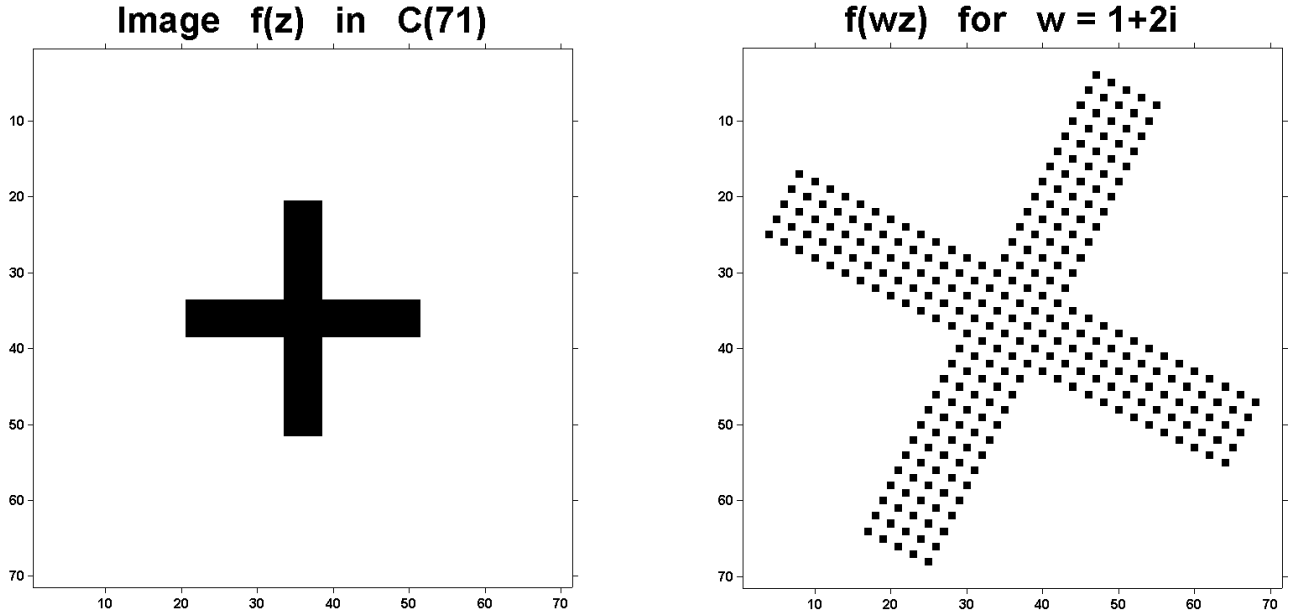


Рис. 2: Комплексное вращение на торе $\mathbb{C}_{71}$

Отметим, что масштабирование является частным случаем комплексного вращения, а сдвиги и отражения на КДТ принимают вид

$$\mathcal{T}_w[f] = f(z+w), \qquad \mathcal{M}_y[f] = f(z^*), \qquad \mathcal{M}_x[f] = f(-z^*).$$

Проиллюстрируем введенное понятие. Рисунок 2 показывает изображение креста на КДТ $\mathbb{C}_{71}$ и результат комплексного вращения этого изображения при $w = 1 + 2i$. Как нетрудно заметить, при этом «площадь увеличилась в 5 раз». В целом, преобразование можно рассматривать как непрерывное вращение на угол $\operatorname{arctg} 2$, совмещенное с масштабированием на $\sqrt{5}$.

Несмотря на то что на КДТ понятия угла, расстояния, площади и отношений порядка утрачивают привычный смысл, понятие ортогональности на нем сохраняется.

Определение 4. *Два вектора $z_1 = (a, b)$ и $z_2 = (c, d)$ на комплексном дискретном торе $\mathbb{Z}_p^2$ назовем ортогональными, если $ac + bd = 0$ или, что равносильно, $\operatorname{Re}(z_1 z_2^*) = 0$.*

Поскольку

$$\operatorname{Re}\big((wz_1)(wz_2)^*\big) = N(w)\operatorname{Re}(z_1 z_2^*),$$

немедленно получаем

Утверждение 2. *Комплексные вращения сохраняют ортогональность: если z_1 и z_2 ортогональны, то wz_1 и wz_2 тоже ортогональны.*

На рис. 3 показаны преобразования известного изображение из базы данных Brodatz, редуцированного из исходного размера 512×512 в формат 251×251 и рассматриваемого на комплексном торе $\mathbb{C}_{251}$. Оригинал показан в левой верхней части рисунка.

В правой верхней части показан результат вращения при $w = 1 + i$. Поскольку при этом «площадь увеличивается» в 2 раза, части преобразованного изображения накладываются друг на друга, что визуально выглядит как его искажение. Важно отметить, что в действительности преобразованное изображение полностью сохраняет всю информацию об оригинале и является просто перестановкой его пикселов.

Результат вращения при $w_1 = (1+2i)^{-1} = -50+100i$ показан в левой нижней части рисунка. В этом случае «площадь уменьшается» в 5 раз, что выглядит как появление четырех уменьшенных версий исходного изображения и ряда фрагментов, в совокупности составляющих пятую. Заметим, что эти уменьшенные версии, визуально представляющиеся идентичными, в действительности составлены из различных пикселов оригинала.

В правой нижней части показан результат вращения на $w_2 = 9+13i$, когда $N(w_2) = 9^2+13^2 = 250 = -1 \in \mathbb{Z}_{251}$. Изображение состоит из 250 фрагментов оригинала, что объясняет визуальные искажения. Вместе с тем хорошо заметно, что фрагменты повернуты на угол $\operatorname{arctg}(13/9) \approx 55^\circ$.

Подводя итоги, назовем угол α *допустимым*, если $\operatorname{tg}(\alpha)$ является рациональным числом, $\operatorname{tg}(\alpha) \in \mathbb{Q}$. Тогда комплексные вращения можно рассматривать как повороты на допустимые углы с соответствующим масштабированием и с учетом периодичности на торе.

Отметим, что сохранение ортогональности при комплексных вращениях позволяет эффективно использовать их, в частности, в задачах анализа симметрии цифровых изображений [6, 7].

Модулярные логарифмы и преобразование Меллина

Воспользуемся аналогией между $\mathbb{C}_p$ и $\mathbb{C}$ для переноса понятия комплексного логарифма в смысле главного значения (или, что равносильно, полярно-логарифмической системы координат) на комплексный дискретный тор. Понятно, что это невозможно сделать с помощью «наивного» подхода, основанного на понятиях полярного угла и полярного радиуса. Поэтому обратимся к алгебраическим методам введения полярных координат на комплексной плоскости.

Вспомним [10], что мультипликативная группа $\mathbb{C}^*$ поля комплексных чисел $\mathbb{C}$ допускает разложение в прямое произведение

$$\mathbb{C}^* \simeq \mathbb{R}^+ \times U(1) \tag{2}$$

где $U(1) = \left\{\, e^{i\theta} \;:\; 0 \leqslant \theta < 2\pi \,\right\}$ — одномерная унитарная группа, которую геометрически можно рассматривать как единичную окружность на комплексной плоскости, а $\mathbb{R}^+ = \left\{\, r > 0 \;:\; r \in \mathbb{R} \,\right\}$ — мультипликативная группа неотрицательных действительных чисел, геометрически представляющая собой на $\mathbb{C}$ луч с выколотым началом. Тем самым устанавливается взаимно-однозначное соответствие

$$z = (a,b) \;\leftrightarrow\; (r,\theta) \in \mathbb{R}^+ \times [0, 2\pi),$$

между ненулевыми комплексными числами $z \in \mathbb{C}^*$ и парами действительных чисел (r,θ), называемых *полярными координатами* точки z. Полагая $l = \ln r$, приходим к *полярно-логарифмическим координатам* (l,θ).

Пусть $\mathbb{C}_p^* = \mathbb{C}_p \setminus \{0\}$ — *мультипликативная группа* тора $\mathbb{C}_p$. Поскольку [8, стр. 429] мультипликативная группа всякого конечного поля является циклической, $\mathbb{C}_p^* = \langle g \rangle$, где g — некоторый примитивный элемент поля $\mathbb{C}_p$. Например, как нетрудно проверить, $\mathbb{C}_{71}^* = \langle 2+7i \rangle$, а $\mathbb{C}_{251}^* = \langle 1+5i \rangle$. Нам понадобится следующий элементарно проверяемый результат:

Утверждение 3. *Для всякого целого $p = 4k+3 \geqslant 3$, числа $m = (p-1)/2$ и $n = 2(p+1)$ являются взаимно-простыми:* $\gcd(m,n) = 1$.

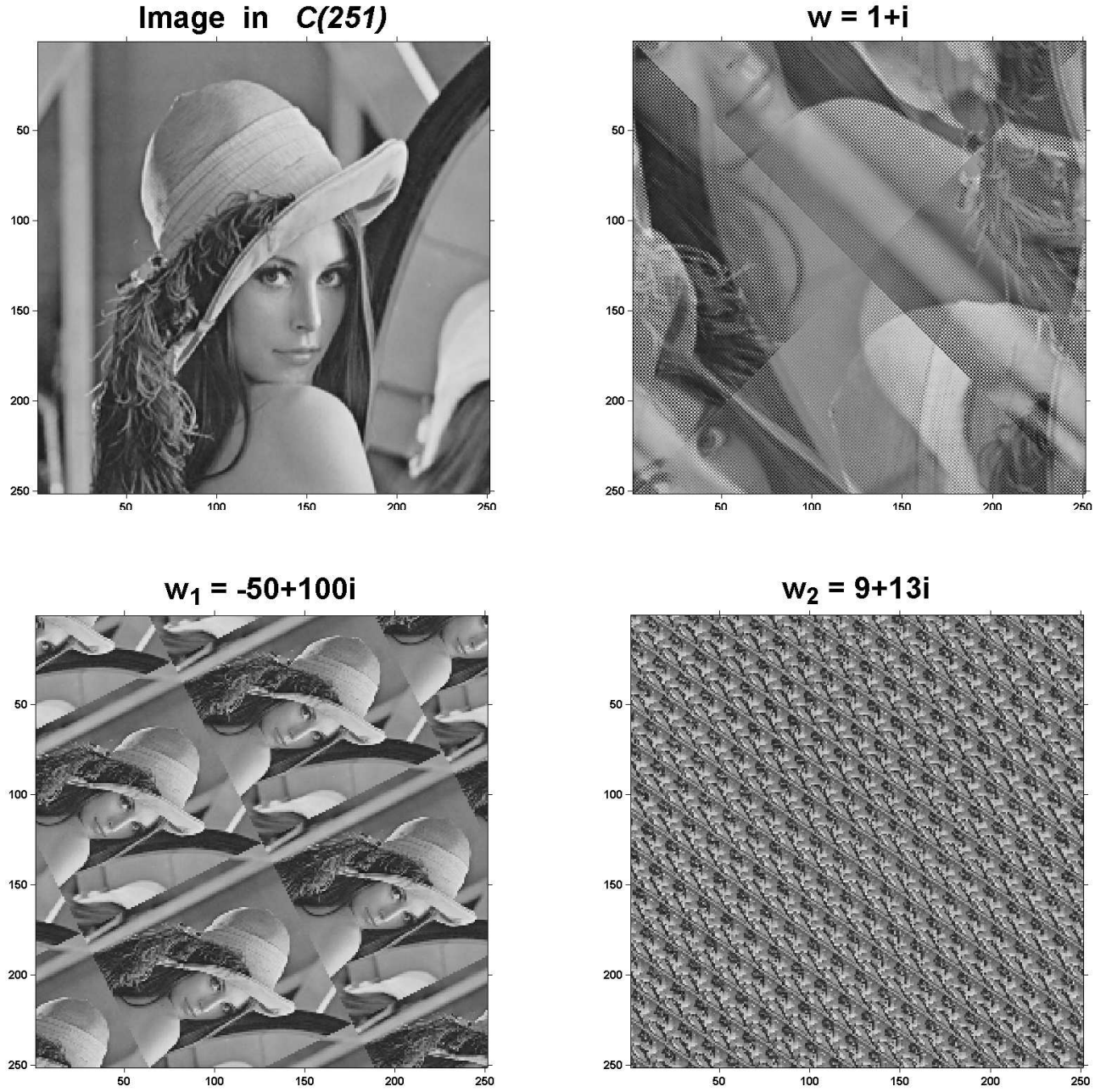


Рис. 3: Комплексные вращения на $\mathbb{C}_{251}$

Заметим, что $mn = p^2 - 1 = |\mathbb{C}_p^*|$. Поэтому для комплексного дискретного тора возникает следующий аналог разложения (2):

Теорема 1. *Мультипликативная группа $\mathbb{C}_p^* = \langle g \rangle$ комплексного дискретного тора разлагается в прямое произведение циклических групп порядков $m = (p-1)/2$ и $n = 2(p+1)$,*

$$\mathbb{C}_p^* \simeq \mathbb{Z}_m \times \mathbb{Z}_n \,, \tag{3}$$

причем

$$\mathbb{Z}_n = \langle g^m \rangle \simeq \{ z \in \mathbb{C}_p^* \ : \ N(z) = \pm 1 \} \qquad \text{и} \qquad \mathbb{Z}_m = \langle g^n \rangle \simeq \{ a^2 \ : \ 0 \neq a \in \mathbb{Z}_p \}.$$

Таким образом, если $(l,\theta) \in \mathbb{Z}_m \times \mathbb{Z}_n$ соответствует $z \in \mathbb{C}_p^*$, то $l \in \mathbb{Z}_m$ может рассматриваться как «логарифм модуля», а $\theta \in \mathbb{Z}_n$ — как «аргумент» дискретного комплексного числа z.

Определение 5. *Дискретный тор $\mathbb{Z}_m \times \mathbb{Z}_n$ будем называть полярной областью для комплексного дискретного тора $\mathbb{C}_p$.* (Заметим, что сама полярная область комплексным дискретным тором не является.)

Зафиксируем примитивный элемент g и определим отображение

$$\mathrm{Exp}_g : \mathbb{Z}_m \times \mathbb{Z}_n \to \mathbb{C}_p^*$$

следующим образом:

$$\text{если} \qquad (l,\theta) \in \mathbb{Z}_m \times \mathbb{Z}_n \ , \qquad \text{то} \qquad \mathrm{Exp}_g(l,\theta) = g^{nl+m\theta} \in \mathbb{C}_p^* \ .$$

Как легко видеть, Exp_g изоморфно отображает аддитивную группу кольца $\mathbb{Z}_m \times \mathbb{Z}_n$ на мультипликативную группу $\mathbb{C}^*(p)$. Назовем его *модулярной экспонентой*, а обратное к нему отображение $\mathrm{Ln}_g : \mathbb{C}^*(p) \to \mathbb{Z}_m \times \mathbb{Z}_n$ — *модулярным логарифмом*. Непосредственно из определения вытекает «основное логарифмическое тождество»

$$\mathrm{Ln}_g(z_1 z_2) = \mathrm{Ln}_g(z_1) + \mathrm{Ln}_g(z_2).$$

Понятие модулярного логарифма позволяет однозначно сопоставить каждому цифровому изображению $f(z)$ размера $p \times p$ цифровое изображение $\psi(l,\theta)$ размера $m \times n$ в полярной области $\mathbb{Z}_m \times \mathbb{Z}_n$ следующим образом:

$$f(z) = \psi\big(\mathrm{Ln}_g(z)\big).$$

Определение 6. *Преобразование $\mathscr{P}[f] = \psi$ назовем полярным преобразованием изображения f, а изображение ψ будем называть полярным образом для f.*

Как показывает следующий результат, преобразование $\mathscr{P}[f]$ может рассматриваться как дискретный аналог перехода в полярно-логарифмическую систему координат.

Утверждение 4. *Комплексным вращениям изображения на КДТ соответствуют циклические сдвиги его образа в полярной области: если $\mathscr{P}[f(z)] = \psi(l,\theta)$, то для всякого $w \in \mathbb{C}_p^*$ выполняется соотношение*

$$\mathscr{P}\big[f(wz)\big] = \psi(l - l_0, \theta - \theta_0) \ , \qquad \textit{где} \quad (l_0,\theta_0) = \mathrm{Ln}_g(w) \ . \tag{4}$$

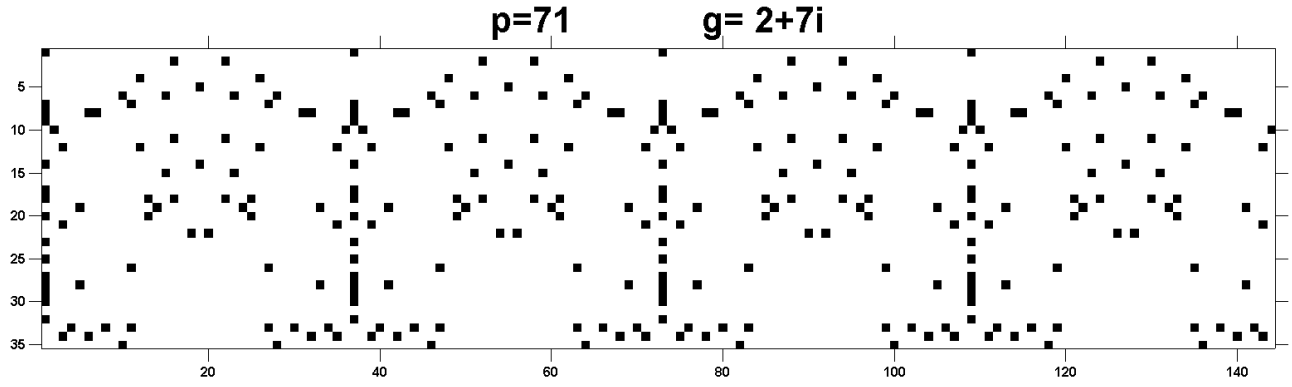


Рис. 4: Полярный образ изображения на рис. 2

В качестве примера рассмотрим изображение f на торе $\mathbb{C}_{71}$, показанное в левой части рис. 2. Его полярный образ имеет размер 35×144 и для $g = 2 + 7i$ показан на рис. 4. Поскольку $\mathrm{Ln}_g(1 + 2i) = (14, 22) \in \mathbb{Z}_{35} \times \mathbb{Z}_{144}$, полярный образ изображения $f(wz)$ из правой части рис. 2 получается из полярного образа оригинала f циклическим сдвигом на 22 единицы вдоль оси «углов» θ и на 14 единиц вдоль оси «логарифмов радиусов» l. Хорошо заметный периодический характер полярного образа обусловлен симметриями оригинала.

Понятно, что полярный образ изображения зависит от выбора примитивного элемента g, выступающего в качестве «основания» модулярного логарифма $\mathrm{Ln}_g(z)$. Следующий результат можно рассматривать как аналог «формулы замены основания».

Утверждение 5. *Если g и h — два различных примитивных элемента группы $\mathbb{C}_p^*$, а z — произвольный элемент из $\mathbb{C}_p^*$, то*

$$\mathrm{Ln}_g(h) \odot \mathrm{Ln}_h(z) = \mathbb{I}_p \odot \mathrm{Ln}_g(z), \tag{5}$$

где $\odot$ означает умножение в кольце $\mathbb{Z}_m \times \mathbb{Z}_n$, а $\mathbb{I}_p := \mathrm{Ln}_g(g)$ является константой, зависящей только от p, и равной $\mathbb{I}_p = (4, m)^{-1} \in \mathbb{Z}_m \times \mathbb{Z}_n$. В частности, $\mathbb{I}_{71} = (9, 107)$, $\mathbb{I}_{251} = (94, 125)$.

Отметим, что полярное преобразование $\mathscr{P}[f]$, являясь обратимым на $\mathbb{C}_p^*$, не будет обратимым в $\mathbb{C}_p$ только потому, что пиксел $(0, 0)$ не принадлежит полярной области. Формально добавляя к ней этот пиксел и считая его значение $f(0, 0)$ неизменным, можем говорить об обратимости $\mathscr{P}$ в *расширенной полярной области.*

Очевидно, что на всяком дискретном торе $\mathbb{Z}_m \times \mathbb{Z}_n$ естественно определяется *дискретное преобразование Фурье* $\mathscr{F}$: под Фурье-образом изображения $f : \mathbb{Z}_m \times \mathbb{Z}_n \to \mathbb{C}$ понимается функция $\mathscr{F}[f] = F : \mathbb{Z}_m \times \mathbb{Z}_n \to \mathbb{C}$, такая, что

$$F(\overline{u}, \overline{v}) = \frac{1}{\sqrt{mn}} \sum_{(\overline{x},\overline{y}) \in \mathbb{Z}_m \times \mathbb{Z}_n} f(\overline{x}, \overline{y})\, e^{-2\pi i \left(x\frac{u}{m} + y\frac{v}{n}\right)}, \tag{6}$$

где $x, y, u, v \in \mathbb{Z}$ являются произвольными представителями соответствующих классов вычетов $\overline{x}, \overline{u} \in \mathbb{Z}_m$, $\overline{y}, \overline{v} \in \mathbb{Z}_n$, а $i \in \mathbb{C}$ — обычная мнимая единица поля комплексных чисел.

Определение 7. *Зафиксировав примитивный элемент g, будем называть дискретное преобразование Фурье полярного образа $\mathscr{P}[f]$ модулярным преобразованием Меллина изображения f, обозначая его как*

$$\mathscr{M}\big[f\big] = \mathscr{F}\big[\mathscr{P}[f]\big].$$

Нетрудно заметить, что $\mathscr{M}$ является линейным и обратимым на расширенной полярной области. Из утверждения 4 и теоремы о сдвиге для преобразования Фурье [11] немедленно получаем

Следствие 1. *Спектр модулярного преобразования Меллина цифрового изображения инвариантен относительно комплексных вращений (а значит, и масштабирования): для любого ненулевого $w \in \mathbb{C}_p$,*

$$\Big|\mathscr{M}\big[f(wz)\big]\Big| = \Big|\mathscr{M}\big[f(z)\big]\Big|.$$

По аналогии с $\mathscr{M}[f]$ можно ввести такие преобразования, как $\mathscr{P}^{-1}\mathscr{F}\mathscr{P}$, $\mathscr{F}\mathscr{P}\mathscr{F}^{-1}$ и $\mathscr{P}^{-1}\mathscr{F}\mathscr{P}\mathscr{F}^{-1}$. В частности, пусть

$$\widetilde{\mathscr{H}}\big[f\big] = \mathscr{M}\Big[\big|\mathscr{F}^{-1}[f]\big|\Big] = \mathscr{F}\mathscr{P}\Big[\big|\mathscr{F}^{-1}[f]\big|\Big].$$

Как легко видеть, $\left|\widetilde{\mathscr{H}}[f]\right|$ инвариантно относительно сдвигов, масштабирования и комплексных вращений цифрового изображения f. Это позволяет считать $\widetilde{\mathscr{H}}$ дискретным аналогом преобразования Фурье-Меллина.

Заключение

Можно надеяться, что представленные выше результаты предоставят дополнительные возможности для решения различных проблем качественной обработки цифровых изображений. Вместе с тем вне рамок настоящей работы остались такие вопросы, как детальное исследование свойств введенных дискретных преобразований, изучение возможностей их 3D-обобщений, а также детали их практического применения, в частности, для анализа симметрии изображений. Автор рассчитывает вернуться к рассмотрению указанных вопросов в дальнейшем.

Литература

[1] *Reddy S., Chatterji B.* An FFT-based technique for translation, rotation, and scale invariant image registration // *IEEE Trans. on Image Processing*. 1996. Vol. 5, no. 8. Pp. 1266–1271.

[2] *Derrode S., Ghorbel F.* Robust and efficient Fourier–Mellin transform approximations for gray-level image reconstruction and complete invariant description // *Computer Vision and Image Understanding*. 2001. Vol. 83, no. 1. Pp. 57–78.

[3] *Derrode S., Ghorbel F.* Shape analysis and symmetry detection in gray-level objects using the analytical Fourier–Mellin representation // *Signal Processing*. 2004. Vol. 84, no. 1. Pp. 25–39.

[4] *Pratt W. K.* Digital image processing. — John Wiley and Sons, 2007. 782 p.

[5] *Averbuch A., Coifman R. R., Donoho D. L., Israeli M., Shkolnisky Y.* A framework for discrete integral transformations I: The pseudo-polar Fourier transform // *SIAM J. Sci. Comput.* 2008. Vol. 30. Pp. 764–784.

[6] *Каркищенко А. Н., Мнухин В. Б.* Преобразование симметрии периодических структур в частотной области // *ММРО-15*. М.: МАКС Пресс, 2011. С. 386–389.

[7] *Каркищенко А. Н., Мнухин В. Б.* Распознавание симметрии изображений в частотной области // *Интеллектуализация обработки информации (ИОИ-2012)*. М.: ТОРУС ПРЕСС, 2012. С. 426–429.

[8] *Кострикин А. И.* Введение в алгебру. М.: Наука, 1977. 496 с.

[9] *Burton D. M.* Elementary number theory. McGrow Hill, 2007. 434 p.

[10] *Dummit D. S., Foote R. M.* Abstract algebra. John Wiley and Sons, 2004. 932 p.

[11] *Poularikas A. D.* The transform and applications handbook. CRC Press, 2010. 1336 p.

Повышение эффективности алгоритма классификации на основе Анализа Формальных Понятий

Прокашева О. В.
prokasheva@gmail.com
МГУ имени М.В. Ломоносова

В настоящей работе исследуется метод классификации на основе построения минимальных гипотез с использованием решётки формальных понятий. Алгоритм тестируется на реальных данных с номинальными и вещественными признаками. Также сравниваются различные модификации метода для уменьшения количества отказов от классификации на основе введения метрик и процедуры голосования.

Ключевые слова: *анализ формальных понятий, классификация, машинное обучение.*

Efficiency improvement of the FCA-based classification algorythm

Prokasheva O. V
Lomonosov Moscow State University

In this paper we investigate the FCA classification algorithm on the basis of minimal hypothesis. The algorithm is tested on various benchmarks with nominal and real features. Also various modifications of the algorithm are presented and compared in terms of their efficiency. These modifications are aimed to reduce classification errors and refusals by using various metrics and voting procedures. **Keywords**: *formal concept analysis, classification, machine learning.*

Введение

Анализ формальных понятий — относительно новый метод анализа данных, относящейся к прикладной ветви алгебраической теории решёток. Данный метод нашел широкое применение в различных областях машинного обучения, таких как информационный поиск, обработка документов и текстов, построение таксономий и классификации.

Существует несколько подходов к классификации в рамках АФП [3]. В данной работе исследуется подход классификации на два класса на основе ДСМ-метода генерации гипотез [2]. В рамках данного подхода гипотезой называется некоторый набор признаков, который присутствует в описании объектов одного класса и не присутствует в описании объектов из других классов. Таким образом гипотезы способны классифицировать новые недоопределенные объекты. Согласно методу АФП гипотезы извлекаются из решёток понятий.

Ранее в работе [1] проводилась оценка эффективности метода классификации на основе генерации гипотез. Метод основан на поиске гипотез, которым удовлетворяет большинство объектов из класса, для каждого класса была построена одна гипотеза. Было показано, что алгоритм в большинстве случаев отказывается от классификации по недостатку информации.

В настоящей работе для каждого класса строится набор всевозможных минимальных гипотез. Алгоритм классификации с использованием извлеченных из обучающей выборки гипотез тестировался на реальных данных. Было установлено, что алгоритм в большом количестве случаев отказывается от классификации по противоречию, то есть гипотезы

относят объект сразу к двум классам. Поэтому было предложено применять различные модификации алгоритма для уменьшения отказов от классификации. В частности были использованы метрики и процедура голосования по большинству.

Анализ Формальных Понятий. Основные определения

Напомним некоторые основные определения Анализа Формальных понятий, согласно [2].

Определение 1. *Формальный контекст — тройка $K = (G, M, I)$, где G — множество объектов, M — множество признаков, I — соответствие между G и M: gIm означает, что объект $g \in G$ обладает признаками $m \in M$*

Определение 2. *Для произвольных $A \subseteq G$ и $B \subseteq M$ определены операторы Галуа:*

$$A' = \{m \in M | \forall g \in A\ (gIm)\};$$

$$B' = \{g \in G | \forall m \in B\ (gIm)\}$$

Оператор $''$ (композиция двух применений оператора $'$) — *оператор замыкания.*

Определение 3. *Пара множеств $H = (A, B) : A \subseteq G, B \subseteq M$, $A' = B$ и $B' = A$, называется формальным понятием контекста K с формальным объемом A $(Extent(H) = A)$ и формальным содержанием B $(Intent(H) = B)$.*

Определение 4. *Понятия (A_1, B_1) и (A_2, B_2) связаны отношением частичного порядка $(A_1, B_1) \leqslant (A_2, B_2)$, если $A_1 \subseteq A_2$ (что эквивалентно $B_2 \subseteq B_1$)*

Определение 5. *Частично упорядоченное по вложению объемов множество формальных понятий контекста K обозначается $L(K)$ и называется решеткой понятий контекста K.*

Далее рассматривается задача классификации по положительным и отрицательным примерам.

Пусть у нас имеется множество объектов G. Все объекты имеют описание на некотором формальном языке, указывающем степень обладания объектами некоторыми признаками, множество которых обозначим M. Множество G разбито на два непересекающихся класса G_+ (*положительный*) и G_- (*отрицательный*) относительно обладания их объектами некоторым *целевым признаком* $z \notin M$. Элементы данных классов, предъявленные в качестве исходных данных для решения задачи классификации, называются, соответственно, *положительными* или *отрицательными примерами.*

В терминах АФП задача классификации по положительным и отрицательным примерам формулируется следующим образом [2] :

Входные данные:

$K_+ = (G_+, M, I_+)$ – положительный контекст по отношению к целевому признаку z;
$K_- = (G_-, M, I_-)$ – отрицательный контекст по отношению к целевому признаку z;
$K_\tau = (G_\tau, M, I_\tau)$ – недоопределённый контекст по отношению к целевому признаку z;

Здесь M — множество признаков, G_+, G_- и G_τ — совокупности соответственно положительных, отрицательных и недоопределённых примеров, а $I_\varepsilon \subseteq G_\varepsilon \times M$, где $\varepsilon \in \{+ +, -, \tau\} \triangleq E$ — соответствия, определяющие их признаки.

Задача:

Для недоопределённых объектов G_τ определить неизвестное значение предиката обладания признаком z.

Определение 6. *Формальное понятие положительного контекста K_+ называется положительным. Если (A, B) — положительное понятие, то множество A называется его положительным формальным объёмом, а множество B — положительным формальным содержанием.*

Аналогично для отрицательных и недоопределённых понятий, формальных объёмов и содержаний контекстов K_- и K_τ.

Определение 7. *Положительное формальное содержание B_+ положительного понятия $(A_+, B_+) \in K_+$ называется:*

1) *положительной гипотезой, если $\underset{K_-}{\forall} (A_-, B_-)\ (B_+ \neq B_-)$, т. е. оно не является формальным содержанием ни одного отрицательного понятия;*
2) *фальсифицированной положительной гипотезой, если $\underset{K_-}{\exists} (g, g^-)\ (B_+ \subseteq g^-)$.*

Аналогично для отрицательных и фальсифицированных гипотез.

Модель классификации в терминах АФП основана на общем принципе: для заданных положительных и отрицательных примеров целевого понятия необходимо построить «обобщение» положительных понятий, которое не покрывало бы отрицательных.

Если формальное содержание g^τ содержит положительную (отрицательную) гипотезу, то говорят, что последняя является гипотезой в пользу положительной (отрицательной) классификации g соответственно.

Если формальное содержание g^τ содержит одновременно или не содержит положительную и отрицательную гипотезу, то алгоритм отказывается от классификации по противоречию (в 1 случае) и по недостатку гипотез (во 2 случае).

На практике удобно выделять не все положительные H_+(отрицательные H_-) гипотезы, а минимальные положительные (отрицательные) гипотезы, которые эвивалентны множествам всех гипотез по отношению к возможным классификациям.

Определение 8. *Положительная гипотеза h_+ есть минимальная положительная гипотеза, если ни одно подмножество $h \in h_+$ не является положительной гипотезой.*

Минимальные отрицательные гипотезы определяются аналогично.

Алгоритм классификации

В статье использован следующий алгоритм классификации по положительным и отрицательным примерам.

Алгоритм распознавания:

1) Бинаризация признаков;
2) Вычисление минимальных гипотез, соответствующих положительным и отрицательным примерам;
3) Классификация недоопределенных примеров на основе выбранных гипотез:
   (а) Если объект содержит гипотезы только из положительного (отрицательного) класса, то объект классифицируется положительно (отрицательно);
   (б) Если объект содержит гипотезы из обоих классов, алгоритм отказывается от классификации по противоречию;
   (в) Если объект не содержит никакие гипотезы из обоих классов, алгоритм отказывается от классификации по недостатку информации.

Поскольку АФП применим к данным с номинальными признаками, то в случае данных с вещественными признаками приходится применять процедуру обработки.

Задача бинаризации признаков сама по себе является отдельным полем для исследования. Применение более тонких методов обработки признакового пространства способно существенно улучшить результаты классификации. В данной работе ставилась задача сравнения качества работы нескольких алгоритмов и выявления возможностей дальнейшего совершенствования подхода на основе АФП, а не поиск наилучшего решения конкретной задачи классификации.

Поэтому было принято решение о применении простого метода бинаризации признаков:

1) Линейное нормирование признаков, то есть отображение в интервал $[0, 1]$;
2) Интервал $[0, 1]$ делился на 10 частей и каждому вещественному признаку соответствовал один интервал, в который он попал из 10 возможных.

Для вычисления минимальных гипотез был использован модифицированный алгоритм «Замыкай-по-одному» [2]. Согласно алгоритму процесс порождения формальных понятий начинается от понятий с наименьшими содержаниями к понятиям с наибольшими содержаниями (то есть «сверху-вниз» согласно порядку $\leqslant$ в решетке понятий).

Алгоритм порождает формальные понятия отдельно для каждого из классов до тех пор, пока не будут найдены все минимальные гипотезы.

Будем считать, что все признаки из M упорядочены и могут быть представлены своими номерами. Обозначим за $min(X)$ функцию, которая выдает объект из множества X с наименьшим номером, за $max(X)$ функцию, которая выдает объект с наибольшим номером. Переменная $prev(B)$ задает единственного предшественника формального понятия с содержанием B в решётке. Переменная $next\_i(B)$ обозначает номер следующего признака, который должен быть проверен как признак, порождающий потомка множества B. min_hyp обозначает список минимальных гипотез.

Алгоритм выделения минимальных гипотез:

$B := \varnothing,\ next\_i(B) := 1,\ prev(B) := \varnothing,\ min\_hyp := \varnothing$
пока $B \neq \varnothing$ или $next\_i(B) \leqslant |M|$
  пока $next\_i(B) \leqslant |M|$ и B'' не является гипотезой
    $i := next\_i(B)$
    если $min((B \cup \{i\})'' \setminus B) \leqslant i$ **то**
      $prev((B \cup \{i\})'') := B$
      $next\_i(B) := next\_i(B) + 1$
      $next\_i((B \cup \{i\})'') := min(\{j \mid i < j \,\&\, j \notin (B \cup \{i\})''\})$
      $B := (B \cup \{i\})''$
    иначе
      $next\_i(B) := next\_i(B) + 1$
  если B'' является гипотезой **то**
    $min\_hyp \leftarrow B''$
  $B := prev(B)$

Согласно алгоритму, процедура генерации «потомков» формального понятия прекращается, как только формальное содержание понятия является гипотезой. Таким образом вычисляются минимальные гипотезы, так как порождения других гипотез, подмножеством которых являются выделенные ранее гипотезы, не происходит.

Временная сложность вычисления всех минимальных положительных и отрицательных гипотез в худшем случае равна $O((|L(K_+) + |L(K_-)|) \cdot |M|^2 \cdot |G_+| \cdot |G_-|)$, $|L(K_+)|$ ($|L(K_-)|$) — размер решетки понятий контекста K_+ (K_- соответственно).

Однако стоит отметить, что результат алгоритма зависит от порядка предоставления признаков и выделенное множество гипотез может также содержать гипотезы, не являющиеся минимальными. Поэтому необходимо производить дополнительную фильтрацию списка min_hyp для удаления гипотез $h \in min\_hyp$, таких, что $\exists\ h_1 \in min\_hyp : h_1 \subset h$.

Эксперименты на реальных данных

Для тестирования алгоритма были использованы данные из UCI Machine Learning repository: [1]

1) *Liver Disorders (Заболевания печени).* Цель задачи — определить пациентов с заболеванием печени. Выборка содержит данные медицинских анализов крови 345 пациентов. Признаки представляют собой результаты анализов — 6 показателей: средний объем эритроцитов, щелочная фотосфаза, аланиновая трансаминаза (АЛТ), аспартатаминотрансфераза (АСТ), гаммаглютамилтранспептидаза, средний объём употребляемого алкоголя в день.
2) *Tic-tac-toe (Крестики-нолики).* Цель задачи — определить выигрышные ситуации. Выборка состоит из 958 игровых ситуаций, представляющих всевозможные конфигурации на поле в конце игры, где крестики ходят первыми. Признаки представляют собой каждый из квадратов на игровом поле (всего 9) и имеют значения «крестик» , «нолик» или «пусто».
3) *House-votes-*84 *(Голосование конгресса США в* 1984 *году).* Цель задачи — определить партию головавшего когрессмена (республиканец или демократ). Выборка состоит из результатов голосования 435 членов палаты представителей США. Признаками являются результаты голосования по 16 вопросам. Значениями является «высказался за», «высказался против», «воздержался».

Данные из задачи *Liver Disorders (Заболевания печени)* подверглись бинаризации согласно алгоритму, описанному в предыдущей главе. В итоге было сформировано 60 признаков.

Алгоритм распознавания реализован в среде MATLAB. Тестирование производилось с помощью скользящего контроля по 10 блокам. Результаты работы представлены в таблице 1. Также в таблице 1 указано время работы программы на системе с процессором Intel Core i5, 2.3 ГГц, 4 Гб ОЗУ.

Как видно из результатов тестирования, алгоритм в большинстве случаев отказывается от классификации по противоречию.

Модифицированный алгоритм

Для увеличения количества распознанных объектов была проведена модификация данного алгоритма. Первая модификация представляет собой отбор гипотез H, которым удовлетворяет некоторое количество объектов, большее чем параметр алгоритма $P : |Extent(H)| > P$. Это позволяет не учитывать шум и значительно уменьшает время работы алгоритма.

Далее были использованы следующие алгоритмы классификации на основе отобранных гипотез:

[1] http://archive.ics.uci.edu/ml/

Таблица 1: Результаты тестирования простого алгоритма

Задача	*% отказа по недостатку*	*% отказа по противоречию*	*% верно*	*% ошибок*	*Время работы (сек.)*	*Кол-во гипотез*
1. Liver Disorders	13.53%	22.8%	42.8%	20.88%	12, 8	258
2. Tic-tac-toe	0%	99.3%	0.7%	0%	51	1887
3. House votes 84	0%	37.96%	62%	0.04%	390	1777

Таблица 2: Результаты тестирования модифицированного алгоритма

Задача	*Алгоритм классификации*	*% отказ*	*% верно*	*% ошибок*	*Время работы (сек.)*	*Кол-во гипотез*
1. Liver Disorders	АФП, голосование, $P = 0,3\%$	$0,28\%$	$65,5\%$	$34,1\%$	$9,71$	110
	SVM	0%	$60,7\%$	$39,30\%$	$0,11$	-
	$C4.5$	0%	$60,6\%$	$39,40\ \%$	$0,44$	-
2. Tic-tac-toe	АФП, метрика (а), $P = 3\%$	0%	100%	0%	$3,16$	17
	SVM	0%	$98,3\%$	$1,7\%$	$0,4$	-
	$C4.5$	0%	$89,1\%$	$10,90\%$	$0,16$	-
3. House votes	АФП, голосование, $P = 2,8\%$	$0,6\%$	$96,3\%$	$3,1\%$	$94,7$	510
	SVM	0%	$95,7\%$	$4,3\%$	$0,09$	-
	$C4.5$	0%	$93,1\%$	$6,9\%$	$0,17$	-

1) Простое голосование (голосование по большинству);
2) Введение метрики $\rho(\boldsymbol{x}, \boldsymbol{y})$:
   (а) $\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{|\boldsymbol{x} \cap \boldsymbol{y}|}{|\boldsymbol{y}|}$;
   (б) $\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{|\boldsymbol{x} \cap \boldsymbol{y}|}{|\boldsymbol{x} \cup \boldsymbol{y}|}$;
   (в) $\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{2} \cdot |\boldsymbol{x} \cap \boldsymbol{y}|}{|\boldsymbol{x}| + |\boldsymbol{y}|}$;
   (г) $\rho(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{|\boldsymbol{x} \oslash \boldsymbol{y}|}{|\boldsymbol{x} \cup \boldsymbol{y}|}$, $\boldsymbol{x} \oslash \boldsymbol{y} = (\boldsymbol{x} \backslash \boldsymbol{y}) \cup (\boldsymbol{y} \backslash \boldsymbol{x})$.

Результаты тестирования модифицированного алгоритма сравнивались с результатами алгоритмов классификации SVM и $C4.5$. В таблице 2 представлены лучшие результаты на скользящем контроле по 10 блокам.

Как видно из результатов тестирования, качество работы алгоритма на основе Анализа Формальных Понятий не уступает известным алгоритмам SVM и $C4.5$, а в лучшем случае

превосходит их. Однако алгоритмы на основе Анализа Формальных Понятий проигрывают по времени работы. Таким образом применение таких алгоритмов целесообразно в задачах, требующих точного решения и не имеющих сильного ограничения по времени решения.

Также стоит отметить, что отбор гипотез позволяет существенно улучшить качество распознавания. Процент верно распознанных объектов на стадии тестирования у алгоритма на основе голосования по большинству в среднем больше, чем у других алгоритмов на основе метрик (а)-(г).

Модифицированный алгоритм классификации эффективно решает задачи с номинальными и бинарными признаками. Решение задачи «Liver Disorders» может быть улучшено с помощью другой процедуры бинаризации признаков, например применения интервалов не фиксированной длины.

Заключение

В статье рассматриваются модификации простого алгоритма на основе АФП и сравнивается их эффективность.

Установлено, что при введении ограничения на размер формального объема гипотез и применения метода голосования по большинству классификатор способен эффективно решать задачи распознавания с номинальными и бинарными признаками. Также благодаря порождению всевозможных минимальных гипотез значительно уменьшается количество отказов от классификации по сравнению с алгоритмами, описанными в [1].

Анализ Формальных Понятий представляет собой перспективное направление для исследований в области обработки данных и извлечения зависимостей.

Дальнейшие исследования могут быть направлены на решение задач классификации с вещественными признаками и применения методов АФП для предобработки признакового пространства, в частности для повышения компактности данных. Также важной задачей является реализация быстрых алгоритмов построения решетки формальных понятий.

Литература

[1] *Онищенко А. А., Гуров С. И.* Классификация на основе АФП и бикластеризации: возможности подхода // Прикладная математика и информатика: Труды факультета Вычислительной математики и кибернетики. — 2011. — Т. 38. — С. 77–87.

[2] *Kuznetsov S.* Mathematical aspects of concept analysis. // Journal of Mathematical Science. — 1996. — Vol. 80, No. 2. — Pp. 1654–1698.

[3] *Meddouri N., Meddouri M.* Classification Methods based on Formal Concept Analysis // CLA 2008 (Posters), Palackŕy University, Olomouc, 2008. —Pp. 9–16.

Анализ трехмерных текстур с позиции стохастической геометрии и функционального анализа*

Федотов Н. Г.[1]***, Голдуева Д. А.***[1]
nikolayfedotov@mail.ru, daria-a-m@yandex.ru
[1]Пензенский государственный университет

Предложен новый подход к анализу трехмерных текстур, основанный на аппарате стохастической геометрии и функционального анализа. Приведены результаты исследования новых триплетных признаков на устойчивость к масштабным преобразованиям на примере трехмерных текстур, полученных с помощью атомно-силовой микроскопии.

Ключевые слова: *трехмерные текстуры, стохастическая геометрия, триплетные признаки.*

Analysis of three-dimensional textures from a position of stochastic geometry and functional analysis*

Fedotov N. G.[1]***, Goldueva D. A.***[1]
[1]Penza State University

A new approach is suggested on three-dimensional (3D) texture analysis based on stochastic geometry and functional analysis. A sensivity of triple features to scale transformation was examined on 3D textures scanned by the atomic-force microscope.

Keywords: *three-dimensional textures, stochastic geometry, triple feature.*

Введение

Во многих отраслях знаний существенная часть информации заключается в сложноструктурированных изображениях, многие из которых содержат текстуры. Сложноструктурированные изображения содержат множество объектов, относящихся к различным видам, каждый из которых обладает своими собственными значимыми характеристиками. Наряду с общетеоретическим значением задача распознавания подобных изображений исключительно актуальна и с прикладной точки зрения. От ее успешного решения зависит эффективность обработки информации в области аэрокосмических исследований, анализа Земли из космоса, медицинской и технической диагностики.

Согласно существующим в настоящий момент методам анализа трехмерных поверхностей, формирование признаков осуществляется на основе контурного и текстурного описания наблюдаемой проекции изображения трехмерного объекта, целью которых является инвариантное представление объекта при изменении его масштаба и поворота в наблюдаемой плоскости.

Большинство методов распознавания трехмерных объектов, как правило, предполагают анализ плоских проекций наблюдаемых изображений. Такой подход объясняется относительной новизной задачи распознавания 3D-образов, в то время как по задачам распознавания плоских изображений накоплен богатый опыт. Поэтому методы 3D-распознавания, в основном, базируются на обработке двумерных изображений.

Работа выполнена при финансовой поддержке РФФИ, проект № 12-07-00501.

Общая идея подобных алгоритмов заключается в следующем. Предположим, имеется изображение проекции некоторого трехмерного объекта из известного класса эталонных объектов. Тогда самый простой метод его распознавания будет заключаться в переборе всех возможных комбинаций проекций эталонных объектов и выбора среди них наиболее подходящих к представленной проекции в соответствии с выбранной мерой сходства. В результате работы такого алгоритма сразу несколько объектов из эталонных могут подходить к представленной проекции. Например, по одной из проекции цилиндра можно сделать вывод, что это еще и шар. Поэтому ошибки распознавания такого рода в алгоритмах распознавания 3D-образов по одной проекции не исключаются.

Почти все алгоритмы распознавания трехмерных образов реализуют описанную выше идею и отличаются лишь способом подбора проекций трехмерных объектов. Таким образом, они предполагают предварительное упрощение анализируемого трехмерного объекта, что ведет к потере существенной доли полезной информации и, как следствие, к снижению точности анализа 3D изображений.

В настоящей статье предлагается новый универсальный подход к анализу трехмерных текстур, основанный на аппарате стохастической геометрии и функционального анализа [1, 2, 3], позволяющий анализировать трехмерные текстуры непосредственно, без предварительного их упрощения.

Кроме того, большинство методов анализа трехмерных объектов оперируют небольшим количеством признаков, имеющих конкретную интерпретацию в терминах решаемой задачи. Предлагаемый в настоящей статье метод анализа трехмерных текстур позволяет в режиме автоматической генерации формировать десятки тысяч признаков изображений, что повышает надежность их классификации.

Теория триплетных признаков

Ключевым элементом теории триплетных признаков является геометрическое трейс-преобразование, связанное со сканированием анализируемого объекта по сложным траекториям, введенное в [1] и исследованное в [1, 4, 5, 6, 7] и последующих работах. Трейс-преобразование сводимо в частных случаях к преобразованиям Радона, Фурье, Хо, Радона–Хо, но не эквивалентно им [2]. Трейс-преобразование является удобным инструментом для распознавания движущихся объектов. На его основе возможно получение нового класса признаков распознавания, позволяющих обеспечить инвариантное или сенситивное распознавание к группе движений и масштабным преобразованиям [7]. С помощью признаков, посредствам которых организовано сенситивное распознавание, можно дополнительно определить параметры преобразования, которое претерпел исходный объект анализа [7]. Исследование движения с помощью трейс-преобразования рассмотрено в [4, 6]. В [8] было введено в рассмотрение двойственное трейс-преобразование, которое позволяет совместно с трейс-преобразованием и триплетными признаками произвести нелинейную фильтрацию изображений с целью их предварительной обработки, предшествующей формированию признаков, а именно: полигональную аппроксимацию, выделение выпуклой оболочки, уменьшение зашумленности и т. д. [8, 9]. На основе трейс-преобразований и теории триплетных признаков возможно эффективное решение широкого круга практических задач распознавания и анализа изображений в различных сферах человеческой деятельности. Применение трейс-преобразования и теории триплетных признаков для распознавания дефектов сварных соединений рассмотрено в [10]. Применение указанного аппарата для решения задач распознавания человеческих лиц и биометрического поиска рассмотрено в [11, 12, 13]. Анализ и распознавание объектов нанотехнологий из области

биологии рассмотрены в [2]. В этой же книге дается анализ точности формирования триплетных признаков. В работе [14] описывается анализ изображений ультразвуковых исследований из области медицины. Таким образом, предлагаемый подход имеет обширное применение.

Признаки изображения в рассматриваемом подходе имеют структуру в виде композиции трех функционалов [1]:

$$\Pi(F) = \Theta \circ \mathrm{P} \circ \mathrm{T}(F \cap l(\theta, \rho)), \tag{1}$$

где θ, ρ полярные координаты сканирующей прямой $l(\theta, \rho)$ [2], с которыми связаны функционалы Θ и P соответственно; функционал T связан с параметром t, задающим точку на сканирующей прямой $l(\theta, \rho)$; $F(,)$ — функция изображения на плоскости $(,)$. Подобное представление признака оказалось продуктивным в силу своей геометричности: многие известные формулы стохастической геометрии и известные преобразования Радона, Хо, Фурье и др. укладываются в такую трехкомпонентную форму. В связи с характерной структурой такие признаки были названы триплетными. Функционал T называют трейс-функционалом, P — диаметральным функционалом, Θ — круговым функционалом. Области определения и области значения функционалов T, P и Θ являются подмножествами множества действительных чисел. Функционалы T, P и Θ выбираются из различных областей математики: теории вероятности, математической статистики, теории рядов и фракталов, стохастической геометрии и т. д. Таким образом, триплетные признаки сохраняют следы генезиса соответствующих областей математики, чем объясняется гибкость и интеллектуальность алгоритмов распознавания, базирующихся на триплетных признаках. Подробное описание формирования триплетных признаков бинарных изображений можно найти в [2], для полутоновых текстур в [15].

Формирование триплетных признаков трехмерных текстур

Под текстурой в настоящей работе будет пониматься изображение поверхности с повторяющимися примитивами, каким-либо образом распределенными на ней. Трехмерные текстуры, в отличие от двумерных, имеют еще одну группу характеристик, описывающую особенности высот. Поэтому для анализа трехмерных текстур целесообразно построить совокупность признаков, учитывающую как особенности проекции точек минимума текстуры на плоскость (или особенности яркости поверхности текстуры), так и особенности высот анализируемого изображения.

В настоящей работе для примера рассматривались трехмерные текстуры, полученные при помощи атомно-силового микроскопа (рис. 1).

Триплетные признаки трехмерных текстур имеют следующий вид:

$$\Pi(F) = \Theta \circ \mathrm{P} \circ \mathrm{T}(F \cap \alpha), \tag{2}$$

где α — сканирующая плоскость перпендикулярная плоскости , $F(,,z)$ — функция изображения в пространстве $(,,z)$, которая каждой точке поверхности текстуры $A(,,z)$ ставит в соответствие z. Результат пересечения функции изображения с плоскостью α есть кривая $q(\theta, \rho, z)$ (рис. 2), заданная в цилиндрической системе координат параметрами θ, ρ и z. Проекция кривой q на плоскость есть прямая $l(\theta, \rho)$, где θ, ρ — ее полярные координаты, с которыми связаны функционалы Θ и P соответственно; функционал T связан с параметром t, задающим точку на прямой l.

Для решения задачи анализа трехмерных текстур, были выделены две группы триплетных признаков:

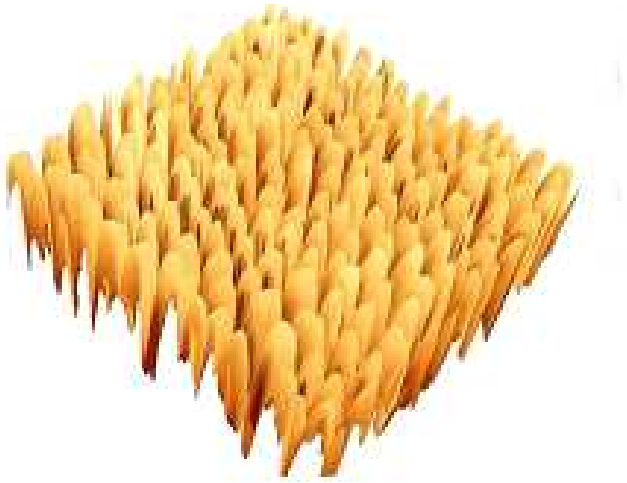


Рис. 1: Пример трехмерной текстуры, полученной при помощи атомно-силового микроскопа

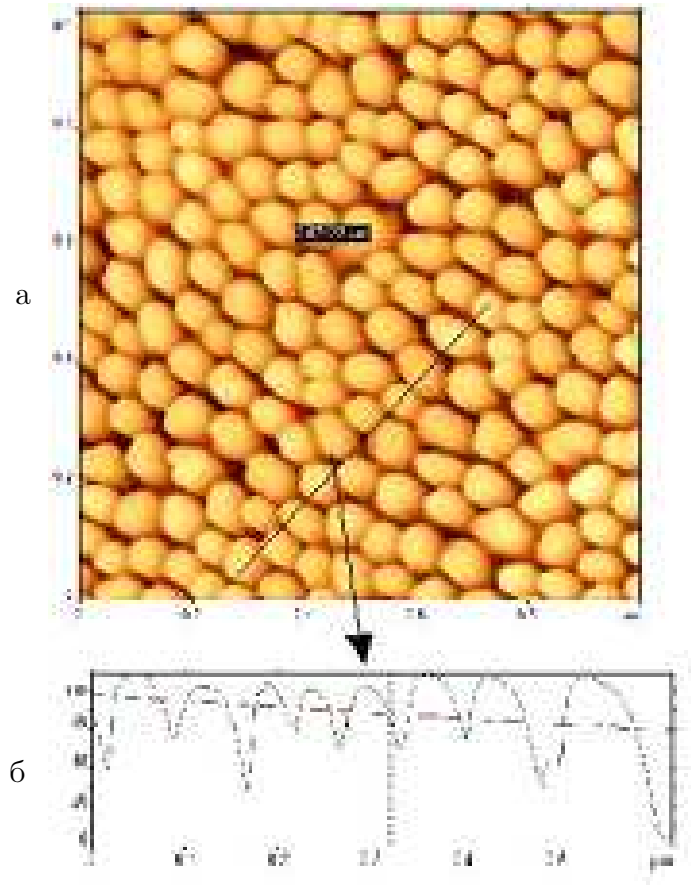


Рис. 2: Пример построения кривой q: (а) трехмерная текстура с выделенным направлением сканирования; (б) кривая q вдоль выделенного направления на координатной плоскости $tO'z$

1. признаки, характеризующие особенности проекции минимумов трехмерных текстур на плоскость ;
2. признаки, характеризующие особенности высот трехмерных текстур.

Причем вторую группу можно подразделить на три: 2.1, 2.2 и 2.3.

Признаки первой и второй группы имеют одинаковую трехфункциональную структуру. Отличие между ними заключается лишь в подходе к заданию характеристик отрезков прямой $l(\theta, \rho)$, соответствующих одному участку кривой q между двумя значимыми соседними минимумами. Для построения признаков, характеризующих особенности проекции минимумов трехмерных текстур на плоскость , каждому отрезку a_i прямой $l(\theta, \rho)$, соответствующему i-му участку кривой q между двумя значимыми соседними минимумами, ставится в соответствие его длина. Для построения признаков типа 2.1 каждому отрезку a_i прямой $l(\theta, \rho)$, соответствующему i-му участку кривой q между двумя значимыми соседними минимумами, ставится в соответствие максимальная высота на отрезке a_i. Для построения признаков типа 2.2 каждому отрезку a_i прямой $l(\theta, \rho)$, соответствующему i-му участку кривой q между двумя значимыми соседними минимумами, ставится в соответствие коэффициент асимметрии. Для построения признаков типа 2.3 каждому отрезку a_i прямой $l(\theta, \rho)$, соответствующему i-му участку кривой q между двумя значимыми соседними минимумами, ставится в соответствие радиус кривизны в точке с максимальной координатой z отрезка a_i.

Для формирования признаков первого и второго типа функция $f(\theta, \rho, t)$ принималась равной:

$$f(\theta, \rho, t) = z\,,$$

где z — высота в точке t прямой $l(\theta, \rho)$ такой, что $F \cap \alpha \neq \varnothing$, $l(\theta, \rho)$ — проекция пересечения $F \cap \alpha$ на плоскость xOy.

Далее посредством трейс-функционала $\mathrm{T}(F \cap \alpha) = \mathrm{T}f(\theta,\ \rho,\ t)$ составляется трейс-матрица, по которой путем последовательной свертки диаметральным и круговым функционалом определяется признак $\Pi(F) = \Theta \circ \mathrm{P} \circ \mathrm{T}(F \cap \alpha)$. Понятие трейс-матрицы подробно описано в [2].

Зачастую анализируемые трехмерные текстуры, относящиеся к одному классу, имеют высокий уровень вариабельности формы и положения повторяющихся включений. Поэтому для обеспечения гибкости, надежности и универсальности систем распознавания трехмерных текстур целесообразно применять признаки, инвариантные к группам движений и масштабным преобразованиям.

Триплетные признаки трехмерных текстур, инвариантные к группе движений и масштабным преобразованиям

Решение задачи формирования инвариантных признаков изображений целесообразно начать с исследования свойств функционалов, образующих композицию в структуре искомых признаков, поскольку именно от них зависит инвариантность признака по отношению к движению и масштабным преобразованиям изображений.

Рассмотрим более подробно каждый вид преобразований изображений.

1) Перенос. Для того, чтобы признак $\Pi(F)$ был инвариантен к переносу, необходимо и достаточно, чтобы P функционал был инвариантен к переносу [2].

Приведем примеры признаков, инвариантных к переносу:

$\Pi_1 = \mathrm{T}_1 \circ\ \mathrm{P}_2 \circ \Theta_7$,

$\Pi_2 = \mathrm{T}_2 \circ\ \mathrm{P}_2 \circ \Theta_5$,

$\Pi_3 = \mathrm{T}_3 \circ\ \mathrm{P}_2 \circ \Theta_7$.

Здесь $\mathrm{T}_1 = \max\limits_i x_i$, где $x_i = \max\limits_j f(\theta, \rho, t_j)$, t_j — точки отрезка $_i$ прямой $l(\theta, \rho)$, где j — порядковый номер точки прямой $l(\theta, \rho)$ в дискретном ее представлении; $\mathrm{T}_2 = n$, где n — число

отрезков $_i$, выделяемых на сканирующей прямой; $\mathrm{T}_3 = \frac{\sum_{i=1}^{n} x_i}{n}$, $x_i = \frac{\left(\sqrt{1+\left(\frac{df(\theta,\rho,t)}{dt}|_{t=t_S}\right)^2}\right)^3}{\left|\frac{d^2 f(\theta,\rho,t)}{dt^2}|_{t=t_S}\right|}$, где $f(\theta,\rho,t_s) = \max\limits_j f(\theta,\rho,t_j)$, t_j - точки отрезка $_i$ прямой $l(\theta,\ \rho)$, если $n \neq 0$. В противном случае, $\mathrm{T}_3 = 0$; $\mathrm{P}_2 = \sqrt{\sum_{i=1}^{m} g^2(\theta_j,\rho_i)}$, где $g(\theta_j,\ \rho_i) = \mathrm{T}(F \cap \alpha)$, m — число дискретных значений ρ; $\Theta_5 = \max\limits_j \mathrm{h}(\theta_j)$, где $\mathrm{h}(\theta_j) = \mathrm{P}(g(\theta_j,\ \rho))$; $\Theta_7 = \sum_{j=1}^{n} \ln|\mathrm{h}(\theta_j)+1| \cdot \Delta\theta$, где $\Delta\theta$ — шаг сканирования по θ.

2) Поворот, при котором точка O переходит сама в себя, ось Ox в Ox', ось Oy в Oy', ось Oz сама в себя. Для того чтобы признак $\Pi(\mathrm{F})$ был инвариантен к повороту, необходимо и достаточно, чтобы Θ функционал был инвариантен к повороту [2].

Приведем примеры признаков, инвариантных к повороту:

$\Pi_4 = \mathrm{T}_2 \circ \mathrm{P}_1 \circ \Theta_3$,

$\Pi_5 = \mathrm{T}_8 \circ \mathrm{P}_6 \circ \Theta_9$,

$\Pi_6 = \mathrm{T}_3 \circ \mathrm{P}_1 \circ \Theta_4$.

Здесь $\mathrm{T}_8 = \sum_{i=1}^{k} |x_{i+1} - x_i|$, если $k \geqslant 2$,

$$x_i = \frac{\sum_{j=s}^{d} t_j^3 f(\theta,\rho,t_j) - 3\left(\sum_{j=s}^{d} t_j f(\theta,\rho,t_j)\right)\sum_{j=s}^{d} t_j^2 f(\theta,\rho,t_j) + 2\left(\sum_{j=s}^{d} t_j f(\theta,\rho,t_j)\right)^3}{\sqrt{\sum_{j=s}^{d} t_j^2 f(\theta,\rho,t_j) - \left(\sum_{j=s}^{d} t_j f(\theta,\rho,t_j)\right)^2}},$$

t_j — точки отрезка a_i прямой $l(\theta,\ \rho)$, $t_{s-1} = \mathrm{t}_{d+1} = 0$ или t_{s-1}, t_{d+1} не принадлежат области сканирования; $\mathrm{P}_1 = \sum_{i=1}^{m} g(\theta_j,\rho_i) \cdot \Delta\rho$, где $\Delta\rho$ - шаг сканирования по ρ; $\mathrm{P}_6 = \sum_{i=1}^{m} g(\theta_j,\rho_i)$; $\Theta_3 = \sqrt{\sum_{j=1}^{n} (\mathrm{h}(\theta_j))^2}$; $\Theta_4 = \sum_{j=1}^{n} \mathrm{h}(\theta_j)$; $\Theta_9 = \sum_{j=1}^{n} \mathrm{h}(\theta_j) \cdot \Delta\theta$;

3) Масштабные преобразования. Для того, чтобы признак $\Pi(F)$ был инвариантен к масштабным преобразованиям, необходимо и достаточно, чтобы T или P функционал был инвариантен к масштабным преобразованиям и P функционал был инвариантен к переносу [2].

Пусть F — функция исходного изображения. Пусть F′ — функция изображения F, претерпевшего масштабное преобразование, т. е. $F' = kF$. Под действием того же преобразования плоскость α, пересекающая изображение F, перейдет в плоскость α'. Требуется, чтобы $\mathrm{T}(F \cap \alpha) = \mathrm{T}(F' \cap \alpha')$, или $\mathrm{P}(\mathrm{T}(F \cap \alpha)) = \mathrm{P}(\mathrm{T}(F' \cap \alpha'))$. Прямая l' будет при этом смещена относительно исходного положения прямой l, поэтому целесообразно требовать от P функционала инвариантности к переносу.

Вышесказанное является строгим только для непрерывного случая. Но мы имеем дело с дискретным сканированием, в силу которого появляется некоторое дополнительное условие. Поясним его суть.

При гомотетии диапазон ρ трейс-матрицы расширяется ($k > 1$) или сужается ($k < 1$) в k раз, т. е. $\rho' = k\rho$, так как изображение F′ пересечет в k раз больше (при $k > 1$) или в k раз меньше (при $k < 1$) прямых, чем изображение F. Таким образом, количество ненулевых значений в i-м столбце трейс-матрицы, построенной для изображения F', и количество ненулевых значений в i-м столбце трейс-матрицы, построенной для изображения F будут отличаться в k раз. Поэтому степень нестабильности признака к рассматриваемому преобразованию зависит от выбора P функционала. Если P функционал имеет накопительный характер, например, $\mathrm{P} = \sum_{i=1}^{m} g(\theta_j,\rho_i)$, то содержащий его признак будет иметь высокую степень нестабильности. Обратная же ситуация наблюдается с признаками, в структуру которых входят P функционалы, такие как $\mathrm{P} = \max\limits_i g(\theta_j,\rho_i)$, $\mathrm{P} = \frac{\sum_{i=1}^{m} g(\theta_j,\rho_i)}{m}$ и т. п.

Приведем примеры признаков, инвариантных к масштабным преобразованиям изображений, и экспериментально определим степень нестабильности каждого на трехмерных текстурах, полученных с помощью атомно-силового микроскопа. Степень нестабильности признака определим как относительную его погрешность, выраженную в процентах.

Пусть под действием семейства масштабных преобразований u_1, u_2, …, u_n изображение F переходит в новые изображения F_1, F_2, …, F_n соответственно. Пусть $\Pi(F_1)$, $\Pi(F_2)$, …, $\Pi(F_n)$ — значения признака П для изображений F_1, F_2, …, F_n соответственно. Тогда среднее значение признака:

$$\bar{\Pi} = \frac{\sum_{i=1}^{n} \Pi(F_i)}{n}.$$

Средняя абсолютная погрешность признака:

$$\bar{\varepsilon} = \frac{\sum_{i=1}^{n} \left|\Pi(F_i) - \bar{\Pi}\right|}{n}.$$

Тогда искомая относительная погрешность инвариантности признака есть:

$$\beta = \frac{\bar{\varepsilon}}{\bar{\Pi}} \cdot 100\%.$$

В таблице 1 приведены результаты экспериментов с использованием следующих функционалов.

Таблица 1: Результаты эксперимента

Функционалы, составляющие признак			Значение признака изображения						$\beta, \%$
T	P	Θ	$k = 1$	$k = 1{,}3$	$k = 1{,}5$	$k = 0{,}5$	$k = 0{,}6$	$k = 0{,}75$	
T_{12}	P_{10}	Θ_2	3,200	3,760	3,200	3,610	3,790	3,900	7,02
T_{11}	P_4	Θ_3	54,34	59,43	57,29	51,67	53,86	60,56	5,16
T_{11}	P_4	Θ_4	1032,43	958,41	1014,39	924,53	942,54	1010,99	3,94
T_{11}	P_4	Θ_5	5,69	4,38	4,25	5,71	5,22	4,88	10,32
T_{11}	P_4	Θ_6	2,44	2,62	2,89	2,51	2,77	3,1	7,28
T_{11}	P_4	Θ_7	6,98	6,32	6,81	6,64	7,03	7,07	3,21
T_{11}	P_4	Θ_8	4,21	4,85	4,57	4,29	4,53	4,25	4,49
T_{11}	P_4	Θ_9	8,87	8,54	8,91	8,46	9,23	9,35	3,03
T_{11}	P_4	Θ_{10}	394,74	376,62	388,23	359,34	421,46	408,52	4,27
T_{13}	P_4	Θ_5	0,4	0,35	0,39	0,36	0,41	0,37	5,26
T_{13}	P_4	Θ_7	1,82	1,79	1,92	2,43	1,96	2,03	7,97
T_{11}	P_{10}	Θ_2	2,62	2,45	3,04	2,41	3	3,02	9,55

$T_{11} = \frac{n \cdot \sqrt{\sum_{i=1}^{n} x_i^2}}{\sum_{i=1}^{n} x_i}$, если $\sum_{i=1}^{n} x_i \neq 0$, в противном случае, $T_{11} = 0$, $x_i = \max\limits_{j} f(\theta, \rho, t_j)$, t_j - точки отрезка a_i прямой $l(\theta,\ \rho)$;

$T_{12} = \frac{\sum_{i=1}^{n} x_i}{n \cdot \max\limits_{i} x_i}$, если $n \neq 0$. В противном случае, $T_{12} = 0$,

$$x_i = \frac{\sum_{j=s}^{d} t_j^3 f(\theta,\rho,t_j) - 3\left(\sum_{j=s}^{d} t_j f(\theta,\rho,t_j)\right)\sum_{j=s}^{d} t_j^2 f(\theta,\rho,t_j) + 2\left(\sum_{j=s}^{d} t_j f(\theta,\rho,t_j)\right)^3}{\sqrt{\sum_{j=s}^{d} t_j^2 f(\theta,\rho,t_j) - \left(\sum_{j=s}^{d} t_j f(\theta,\rho,t_j)\right)^2}},$$

t_j - точки отрезка a_i прямой $l(\theta,\ \rho)$, $\mathrm{t}_{s-1} = t_{d+1} = 0$ или t_{s-1}, t_{d+1} не принадлежат сетчатке;

$\mathrm{T}_{13} = \frac{\max\limits_i x_i - \min\limits_i x_i}{\max\limits_i x_i}$, где $x_i = \frac{\left(\sqrt{1+(f'(\theta,\rho,t_s))^2}\right)^3}{|f''(\theta,\rho,t_s)|}$, $f(\theta,\rho,t_s) = \max\limits_j f(\theta,\rho,t_j)$, t_j - точки отрезка a_i прямой $l(\theta,\ \rho)$;

$\mathrm{P}_4 = \max\limits_i g(\theta_j,\rho_i)$;

$\mathrm{P}_{10} = \frac{\sum_{i=1}^m g(\theta_j,\rho_i)}{m}$;

$\Theta_2 = \sum_{j=1}^n |\mathrm{h}(\theta_{j+1}) - \mathrm{h}(\theta_j)|$;

$\Theta_5 = \max\limits_j \mathrm{h}(\theta_j)$;

$\Theta_6 = \max\limits_j \mathrm{h}(\theta_j) - \min\limits_j \mathrm{h}(\theta_j)$;

$\Theta_8 = \min\limits_j \mathrm{h}(\theta_j)$;

$\Theta_{10} = \frac{\sum_{j=1}^n \mathrm{h}(\theta_j)}{n}$.

Проведенный эксперимент показал, что построенные триплетные признаки обладают достаточно высокой степенью инвариантности к масштабным преобразованиям трехмерных текстур.

Заключение

Существует обширный класс задач медицинской, технической диагностики, где ключевая информация заключена в зрительных образах. В настоящей статье рассмотрена задача формирования признаков трехмерных текстур. Для решения данной проблемы предложен новый подход, основанный на аппарате стохастической геометрии и функционального анализа. Его суть заключается в формировании триплетных признаков двух типов: характеризующих особенности проекции минимумов трехмерных текстур на плоскость ; характеризующих особенности высот трехмерных текстур. Таким образом, построенная группа признаков позволит более полно описать трехмерные текстуры без предварительного их упрощения. Благодаря трехкомпонентной структуре триплетных признаков возможна генерация большого их количества, что позволяет увеличить гибкость, универсальность и надежность распознавания. Причем, как показывают проведенные эксперименты, при определенном выборе функционалов, входящие в структуру триплетного признака, формируемые характеристики приобретают свойства инвариантности к группе движений и масштабным преобразованиям.

Литература

[1] *Федотов Н. Г.* Методы стохастической геометрии в распознавании образов. М.: Радио и связь, 1990. 144 с.

[2] *Федотов Н. Г.* Теория признаков распознавания образов на основе стохастической геометрии и функционального анализа. М.а: Физматлит, 2009. 304 с.

[3] *Kendall W. S., Molchanov I.* New perspectives in stochastic geometry. Oxford, UK: Oxford University Press, 2010.

[4] *Kadyrov A. A., Saveleva M. V., Fedotov N. G.* Image scanning leads to alternative understanding of image // *3rd conference (International) on Automation, Robotics and Computer Vision (ICARCV'94)*. Singapore.

[5] *Fedotov N. G., Kadyrov A. A.* Image scanning in machine vision leads to new understanding of image // *5th Workshop (International) on Digital in Processing and Computer Graphics Proceedings, Proc. International Society for Optical Engineering (SPIE)*. 1995. Vol. 2363. P. 256–261.

[6] *Fedotov N. G., Shulga L. A.* New theory of pattern recognition feature on the basis of stochastic geometriy // *WSCG'2000 Conference Proceedings*. University of West Bohemia, 2000. Vol. 1(2). P. 373–380.

[7] *Kadyrov A. A., Fedotov N. G.* Triple features pattern recognition and image analysis // *Advances in Mathematical Theory and Applications*. 1995. Vol. 5, no 4. P. 546–556.

[8] *Федотов Н. Г., Кольчугин А. С., Смолькин О. А., Моисеев А. В., Романов С. В.* Формирование признаков распознавания сложноструктурированных изображений на основе стохастической геометрии // *Измерительная техника*. 2008. № 2. С. 56–61.

[9] *Vidal M., Amigo J. M.* Pre-processing of hyperspectral images. Essential steps before image analysis // *Chemometrics and Intelligent Laboratory Systems*. 2012. Vol. 117, no. 1. P. 138–148.

[10] *Федотов Н. Г., Никифорова Т. В.* Техническая дефектоскопия на основе новой теории распознавания образов // *Измерительная техника*. 2002. № 12. С. 27–31.

[11] *Fedotov N. G., Shulga L. A., Roy A. V.* Visual mining for biomatrical system based on stochastic geometry // *Conference (International). Pattern Recognition and Image Analysis Proceedings (PRIA-7-2004) Proceedings*. 2004. Vol. 2. P. 473–475.

[12] *Shin B.-S., Cha E.-Y., Kim K.-B., Cho K.-W., Klette R., Young W. W.* Effective feature extraction by trace transform for insect footprint recognition // *3rd Conference (International) on Bio-Inspired Computing: Theories and Applications (BICTA 2008): IEEE South Australia Section*. Adelaide, NT. 2008. P. 97–102.

[13] *Fooprateepsiri R., Kurutach W.* Highly robust approach face recognition using hamming–trace combination // *IADIS Conference (International) Intelligent Systems and Agents 2010 Proceedings, IADIS Conference (European) on Data Mining 2010 Proceedings, Part of the MCCSIS 2010*. P. 83–90.

[14] *Fedotov N. G., Shulga L. A., Kol'chugin A. S., Smol'kin O. A., Romanov S. V.* Triple features of ultrasonic image recognition // *8th Conference on Pattern Recognition and Image Analysis (PRIA-8-2007) Proceedings*. Yoshkar-Ola, Russia. 2007. Vol. 1. P. 299–300.

[15] *Fedotov N. G., Mokshanina D. A.* Recognition of halfton textures from the stochastic geometry and functional analysis // *Pattern Recognition and Image Analysis*. 2010. Vol. 20, no 4. P. 551–556.

Решение задач анализа данных, основанное на линейной комбинации деформаций*

Дьяконов А. Г.
djakonov@mail.ru
Московский государственный университет имени М.В. Ломоносова;
Вычислительный центр им. А.А. Дородницына РАН

Дан обзор некоторых теоретических результатов представления функций и алгоритмов в специальном виде: линейной комбинации «деформации» линейных функций/алгоритмов. В теории интерполяции подобные результаты отталкиваются от работ А.Н. Колмогорова, а в теории классификации — от работ Ю.И. Журавлева. Показано, что идеи подобного представления можно успешно использовать на практике. Описаны решения нескольких прикладных задач в рамках крупных международных конкурсов.

Ключевые слова: *интерполяция, деформация, алгебраический подход, представление функций, регрессия, рекомендация, классификация, прикладные задачи.*

Data mining problems solving using linear combinations of deformations*

D'yakonov A.G.
Lomonosov Moscow State University; Dorodnicyn Computing Centre of RAS

A review of some theoretical results on function representation by linear combination of "deformations" of linear functions is presented. The results are started from Kolmogorov's papers on interpolation and Zhuravlev's papers on algebraic approach. It is shown that the idea of such representation can be useful in practice. Some solutions of real problems from international data mining competitions are described.

Keywords: *interpolation, calibrating, algebraic approach, function representation, regression, recommendation, classification, applied problems.*

Введение

Работа посвящена теоретическим обоснованиям и практическим приложениям поиска решения задачи регрессии в виде

$$c_1\varphi(B_1) + \cdots + c_r\varphi(B_r), \tag{1}$$

где $B_1, \ldots, B_r$ – решения отдельных регрессионных алгоритмов (либо очень простых, либо представимых в виде линейной комбинации «простых»), а φ – «функция деформации», которая выбирается заранее (исходя из специфики задачи) или специально строится. Поскольку практически любой алгоритм классификации сначала получает некоторые значения, которые естественно назвать «оценками принадлежности к классам», а затем на их основе классифицирует, речь в работе пойдёт не только о регрессии, а о более широком спектре приложений: задачи классификации, рекомендации и т.п. В алгебраическом подходе к распознаванию Ю.И. Журавлева вводятся операции над алгоритмами класси-

Работа выполнена при финансовой поддержке РФФИ, проект №12-07-00187-а.

фикации, которые, по сути, индуцируются операциями над ответами так называемых распознающих операторов. Поэтому с точки зрения этого подхода выражение (1) определяет вид алгоритма.

Сначала в работе дается обзор теоретических результатов, отдельно приводятся схемы доказательств некоторых теорем, затем описываются реальные прикладные задачи и методы их решения, основанные на представлении (1).

Обзор теоретических результатов

Теория интерполяции

В теории интерполяции хорошо известна следующая теорема А.Н. Колмогорова [1], которую постоянно приводят в учебниках по теории нейронных сетей (см., например, [2]) как фундаментальный результат, обосновывающий концепцию нейросетей.

Теорема 1. *Любую непрерывную функцию f, заданную на единичном кубе n-мерного пространства, можно представить в виде*

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} h_q \left(\sum_{p=1}^{n} \varphi_{p,q}(x_p) \right), \tag{2}$$

где h_q, $\varphi_{p,q}$ – непрерывные функции, кроме того, $\varphi_{p,q}$ не зависят от выбора функции f.

Естественно, выражение (2) в теореме лишь внешним видом напоминает функциональность нейронной сети, поскольку на практике используются достаточно простые функции активации (в (2) это функции h_q и $\varphi_{p,q}$). В конце XX века удалось получить несколько результатов, которые в большей степени оправдывают использование нейронных сетей на практике. В русскоязычной литературе чаще приводят следующую теорему 1998 г. [3].

Теорема 2. *Пусть X – компактное пространство, $E \subseteq C(X)$ – замкнутое линейное подпространство в $C(X)$, $1 \in E$, функции из E разделяют точки в X и E замкнуто относительно нелинейной унарной операции $\varphi \in C(\mathbb{R})$, тогда $E = C(X)$.*

Здесь и далее $C(X)$ – множество непрерывных функций со значениями из $\mathbb{R}$, определенных на X. Под замкнутостью относительно φ понимается, что для любой функции $g \in E$ выполняется вложение $\varphi g \in E$ (где $\varphi g(x) = \varphi(g(x))$). Разделение точек означает, что

$$\forall x_1, x_2 \in X \;\; \exists g \in E : \; g(x_1) \neq g(x_2).$$

Из теоремы следует, что любую непрерывную вещественную функцию $f(x_1, \ldots, x_n)$ можно приблизить нейросетью, у которой в качестве функций активации (у нейронов) используются фиксированная функция φ (любая удовлетворяющая условиям теоремы), и тождественная функция. Этот результат был получен значительно раньше (в 1991 г.) В. Крейновичем [4]. Нам потребуется следующий очень интересный результат, полученный в 1993 г. А. Пинкусом и соавторами [5], [6] (см. также обзор [7]):

Теорема 3. *В топологии равномерной сходимости на компактах*

$$C(\mathbb{R}) = \overline{L}\left(\{\varphi(ax + b) \mid a, b \in \mathbb{R}\}\right),$$

где $\varphi \in C(\mathbb{R})$ – не полином.

Здесь $L(X)$ – множество конечных линейных комбинаций элементов из X, верхняя черта – замыкание. Изначально [5], теорема была доказана при более сильных предположениях, однако в дальнейшем – для класса непрерывных функций, обобщена на многомерный случай, кроме того, сформулирована в виде критерия (см. предположение 6.4 работы [7]):

Теорема 4. *Для любой функции $\varphi \in C(\mathbb{R})$ справедливо равенство*

$$C(\mathbb{R}^n) = \overline{L}\left(\left\{ \varphi\left(\sum_{i=1}^{n} a_i x_i + b\right) \,\middle|\, (a_1, \ldots, a_n) \in \mathbb{R}^n,\, b \in \mathbb{R} \right\}\right). \tag{3}$$

тогда и только тогда, когда φ не является полиномом.

Таким образом, любая непрерывная функция может быть приближена двухслойной нейронной сетью, причем в первом слое – фиксированная функция активации φ (непрерывная, неполиномиальная), а во втором – тождественная (открытым остается вопрос о числе нейронов для достижения заданного приближения). Результат А.Н. Колмогорова часто интерпретируют как «отсутствие функций большого числа переменных», т.е. существуют только сумма и функции одного переменного, а все остальные функции – их суперпозиции (речь идет, естественно, о непрерывных функциях). У результата А. Пинкуса и соавторов более сильная интерпретация. С точностью до приближения существуют только линейные комбинации и ровно одна функция одного переменного! Причем на роль такой единственной функции φ подходит любая непрерывная, кроме полинома. Кроме того, при приближении любой непрерывной функции не надо использовать «вложенных суперпозиций», т.е. вида

$$\varphi(\ldots\varphi(\ldots)\ldots). \tag{4}$$

В данной работе мы покажем, что эту смелую интерпретацию часто удается эксплуатировать на практике при решении реальных прикладных задач анализа данных.

Деформации в алгебраическом подходе к решению задач классификации

Как уже было сказано, в алгебраическом подходе вводятся операции над алгоритмами. Подробнее о подходе можно узнать в работах [8],[9]. Пусть алгоритм A должен классифицировать объекты $S_1, \ldots, S_q$ (контрольную выборку) на l, вообще говоря, пересекающихся класса, т.е. получить бинарную $q \times l$-матрицу, ij-й элемент которой является ответом алгоритма на вопрос «принадлежит ли j-му классу объект S_i». В алгебраическом подходе постулируется, что любой алгоритм классификации A можно представить в виде суперпорзиции распознающего оператора B и решающего правила C (в [8] есть даже теорема, доказывающая это утверждение, но она немного искусственная). Распознающий оператор получает вещественную $q \times l$-матрицу оценок $\Gamma[B] = ||\gamma_{ij}||_{q\times l}$: её ij-й элемент интерпретируется как оценка принадлежности j-му классу объекта S_i. Решающее правило переводит матрицу оценок в бинарную матрицу, например, так:

$$C(||\gamma_{ij}||) = ||\alpha_{ij}||,\ \alpha_{ij} = \begin{cases} 1, & \gamma_{ij} \geqslant c, \\ 0, & \gamma_{ij} < c \end{cases} \tag{5}$$

(пороговое решающее правило). В алгебраическом подходе решающее правило фиксируется, а операции над алгоритмами индуцируются операциями над распознающими операторами, которые вводятся следующим образом:

$$\Gamma[B_1 + B_2] = \Gamma[B_1] + \Gamma[B_2],$$

$$\Gamma[c \cdot B_1] = c \cdot \Gamma[B_1],$$

$$\Gamma[B_1 \cdot B_2] = \Gamma[B_1] \cdot \Gamma[B_2],$$

здесь B_1, B_2 – распознающие операторы, умножение «$\cdot$» – поэлементное.

С помощью введенных операций можно строить полиномы над алгоритмами (распознающими операторами).

Весь смысл введения операций объясняется в основной теореме алгебраического подхода [8] (см. также [9]): при необременительных ограничениях на постановку задачи можно с помощью полиномов над некорректными алгоритмами (которые допускают ошибки на контрольной выборке) строить корректные алгоритмы.

Оказывается (мы сформулируем соответствующую теорему в следующем параграфе), что корректные полиномы можно строить в виде (1), где φ – полином специального вида. Но «корректность» начинается с определённой степени полинома. Если же взять в качестве φ неполиномиальную операцию, при этом считаем, что

$$\Gamma[\varphi(B)] = \varphi(\Gamma[B]),$$

где функция φ действует поэлементно, то также корректный алгоритм представим в виде (1). Первоначально автор доказал это для деформации $\varphi(x) = \frac{1}{x}$ [10], но после того как ему стали известны результаты А. Пинкуса (о справедливости которых он уже сам догадался), общий случай стал очевидным (см. доказательство ниже).

Определение 1. *Линейным замыканием модели (множества матриц оценок) называется множество всевозможных линейных комбинаций операторов модели (матриц оценок).*

Отметим, что если линейное замыкание модели алгоритмов пополнить специальными операциями нормировки [10], которые часто применяются на практике, то корректный алгоритм можно не получить, но все, что получается, можно представить в виде (1). Таким образом, опять при пополнении линейного замыкания вложения вида (4) можно не использовать.

Доказательства некоторых результатов алгебраического подхода

Этот параграф не является историческим обзором, и его можно пропустить читателю без потери нити рассуждения. Здесь мы сформулируем и докажем результаты, которые описали раньше. В формулировках будет фигурировать модель алгоритмов B^* (точнее, это фиксированное множество операторов). На нее накладываются следующие условия:

1. В линейном замыкании пространства матриц оценок (операторов модели) есть константная матрица (все элементы которой равны ненулевой константе).
2. В линейном замыкании пространства матриц оценок есть базис, состоящий из бинарных матриц.
3. В линейном замыкании пространства матриц оценок есть матрица с попарно различными элементами.

Этим требованиям удовлетворяет, например, модель операторов вычисления оценок (см. [9]).

При отказе от «экзотического» требования (2) формулировки теорем немного поменяются; на самом деле, многие модели этому требованию удовлетворяют. Требование (3) является скорее свойством задачи, а не модели (ясно, что если $S_1 = \ldots = S_q$, то вряд ли оно выполняется в «разумных моделях»). Оно следует из «непротиворечивости» исходной информации.

Теорема 5. *Любую матрицу оценок, которую можно получить полиномом степени не выше k над операторами модели B^* можно получить полиномом вида* (1)*, где φ – полином k-й степени специального вида (например, подойдет $\varphi(x) = x^k$), $B_1, \ldots, B_r$ – линейные комбинации операторов модели (причём они не зависят от получаемой матрицы и степени k).*

Доказательство. Приведем лишь схему доказательства, которое основано на технике, похожей на обоснование применения полиномиального ядра в SVM. По этой схеме и работе [9] без труда восстанавливается полное доказательство.

Распознающий оператор можно отождествить с его $q \times l$-матрицей оценок, а её – с ql-мерным вектором, достаточно «вытянуть» матрицу в вектор:

$$||\gamma_{ij}||_{q\times l} \;\to\; (\gamma_{11},\ldots,\gamma_{1l},\gamma_{21},\ldots,\gamma_{ql}).$$

Из ограничения на модель следует, что в линейном замыкании M таких ql-мерных векторов есть базис $\tilde{x}_1,\ldots,\tilde{x}_s$ из бинарных векторов и вектор $\tilde{1}=(1,\ldots,1)^{\mathrm{T}}$.

Запишем базисные векторы $\tilde{x}_i$ и их инверсии $\tilde{1}-\tilde{x}_i$ по столбцам в матрицу X. Нетрудно показать, что линейное замыкание $L(X)$ столбцов матрицы X (а это и есть линейное замыкание M) совпадает с линейным замыканием $L(XX^{\mathrm{T}})$ столбцов матрицы Грама XX^{T}.

Пусть теперь матрица X' состоит из всевозможных мономов степени не выше k (мономы x_ix_j и x_jx_i считаем разными) над столбцами матрицы X. Оказывается, что матрица $X'X'^{\mathrm{T}}=\varphi(XX^{\mathrm{T}})$, где функция $\varphi(x)=x^k$ применяется поэлементно. При этом $L(X')=$ $=L(XX')$. Таким образом, в пространстве полиномов степени не выше k над векторами исходного пространства есть база векторов $\varphi(y_j)$, где y_j пробегает столбцы матрицы XX^{T}, т.е. является линейной комбинацией столбцов матрицы X.

Переходя теперь от ql-мерных векторов к соответствующим операторам получаем справедливость утверждения теоремы.

■

Эта теорема описывает полиномиальные замыкания модели, т.е. пространства полиномов степени не выше k (впервые были полностью описаны в [9]). С прикладной точки зрения удручает тот факт, что, вообще говоря, в теореме 5 $r=ql$ (понижение r возможно лишь в частных случаях).

Рассмотрим теперь неполиномиальные операции:

Теорема 6. *Любую $q\times l$-матрицу можно получить оператором вида* (1) *где $B_1,\ldots,B_r$ – операторы из линейного замыкания модели B^*, $\varphi\in C(\mathbb{R})$ – не полином.*

Доказательство. Как и в предыдущем доказательстве отождествим оператор с ql-мерным вектором. В линейном замыкании множества векторов, которое соответствует модели, есть вектор $\tilde{1}=(1,\ldots,1)^{\mathrm{T}}$ и вектор $\tilde{x}=(x_1,\ldots,x_{ql})^{\mathrm{T}}$ с попарно различными элементами. Рассмотрим выражение вида

$$\sum_i c_i\varphi(a_i\tilde{1}+b_i\tilde{x})=\sum_i c_i\begin{bmatrix}\varphi(a_i+b_ix_1)\\ \vdots\\ \varphi(a_i+b_ix_{ql})\end{bmatrix} \tag{6}$$

(сумма конечная). Из теоремы 3 следует, что любой ql-мерный вектор можно представить с любой наперед заданной точностью вектором вида (6) и из векторов такого вида можно составить базис ql-мерного пространства. Действительно, это следует из того, что непрерывную функцию одного аргумента, которая равна 1 в точке x_j, а в остальных x_t, $t\neq j$, равна нулю, можно сколь угодно точно приблизить выражением

$$\sum_i c_i\varphi(a_i+b_ix).$$

Линейная комбинация полученных базисных векторов ql-мерного пространства также будет иметь вид (6). Переходя теперь от векторов к соответствующим операторам получаем справедливость утверждения теоремы. ■

Решение прикладных задач

Приведем примеры реальных прикладных задач, которые удалось успешно решить, используя описанные выше результаты. Причём не просто успешно, а существенно лучше многих классических методов, поскольку решения стали победителями в рамках крупных международных конкурсов. Все они получены алгоритмами, которые имеют вид (1).

Технология LENKOR

Технология разрабатывалась для решения задач с разнородной информацией и изначально применялась для построения рекомендательных систем и прогнозирования деятельности научных коллективов. В [11] описано решение задачи Международного конкурса «ECML/ PKDD Discovery Challenge 2011 (VideoLectures.Net Recommender System Challenge)» [12] по созданию рекомендательной системы ресурса VideoLectures.net. Ниже мы опишем принципы решения этой задачи, чтобы продемонстрировать связь с поиском решения в виде (1).

Для некоторого множества лекций заданы их описания:

— идентификационный номер лекции,
— язык лекции (английский, французский, русский и т.п.),
— категория лекции (например «Computer Science» или «Biology»),
— число просмотров лекции,
— дата публикации на сайте,
— автор лекции (а также подробная информация о нём: e-mail, сайт в Интернете и т.д.),
— название лекции,
— описание лекции (небольшой текст).

Аналогичные данные имеются по событиям, к которым относятся лекции (конференции, на которых они прочитаны, школы-семинары, циклы лекций и т.д., см. подробнее [12]).

Множество описанных лекций разбито на два непересекающихся подмножества: 6983 «старые лекции», которые были опубликованы на сайте до 1 июля 2009 г. (по ним известна вся статистика) и 1122 «новые лекции», которые выложены на сайте после этой даты (по ним не известна информация о просмотрах). Есть также контрольная выборка – это подмножество множества старых лекций. Для каждой лекции из контрольной выборки необходимо предложить список рекомендуемых 30 новых лекций, таким образом, если новый пользователь (интересы которого нам не известны) заходит на сайт и начинает смотреть лекцию – мы можем рекомендовать ему новинки (упорядоченный список из 30 недавно закачанных на сайт лекций).

Качество рекомендации лекций с номерами $a_1, \ldots, a_{30}$ по лекции из контрольной выборки оценивалась по формуле

$$\frac{1}{6} \sum_{z \in \{5,10,15,20,25,30\}} \frac{|\{y_1, \ldots, y_{\min(z,s)}\} \cap \{a_1, \ldots, a_{\min(z,s)}\}|}{\min(z, s)}, \tag{7}$$

где $y_1, \ldots, y_s$ – номера лекций, которые действительно просматривались после рассматриваемой «старой» лекции (идут по уменьшению популярности).

Применение технологии «LENKOR» для решения подобных задач состоит в следующих этапах:

1. Выделение различных видов информации, описание способов вычислений близости по каждому виду.
2. Формирование линейной комбинации функций близости, настройка коэффициентов (методом покоординатного спуска).
3. «Деформирование» комбинации (попытка построить нелинейную формулу решения путем перебора различных алгебраических выражений), настройка коэффициентов (методом покоординатного спуска).

В этой задачи на первом этапе были разработаны способы сравнения похожести заголовков лекции, их тематики, авторов и т.д. Например, при сравнении заголовков применяется стандартный подход к обработке текстов: стемминг, TF·IDF-преобразование и косинусная мера сходства [13]. «Вид информации» здесь понимается в широком смысле, например, бралась вся текстовая информация (заголовок лекции, описание, текст из слайдов), объединялась, и сравнивались такие общие тексты.

На втором этапе применение очень простого оптимизационного метода – покоординатного спуска – оправдано тем, что приходится максимизировать нестандартный функционал качества (7) (он максимизировался напрямую). Замечено, что автоматически определяются «ненужные» виды информации: те, которые не нужно учитывать в решении – им соответствуют нулевые веса. Важность этого наблюдения состоит в том, что деформация не сделает их «нужными», т.е. если на втором этапе некоторый коэффициент нулевой, то он будет нулевым и при настройке коэффициентов на третьем этапе (поэтому сразу лишние слагаемые можно удалить). Естественно, это наблюдение теоретически не обосновано.

На третьем этапе происходит перебор различных φ в выражении вида (1) (из некоторого списка функций). В качестве B_i здесь брались подкомбинации линейной комбинации построенной на первых двух этапах. Конкретно в этой задаче наилучшей оказалась деформация вида $\varphi(x) = \sqrt{x}$.

В результате строится функция близости двух лекций вида (1). Все параметры в ней настроены так, чтобы близкими считались лекции, которые смотрят вместе. Для «старой» лекции находим 30 новых с наибольшими значениями близостей к ней (в соответствующем порядке их и выдаем).

Видно, что идея описанной технологии соответствует теоретическим результатам, описанным в обзоре. Разработанный алгоритм занял первое место среди 62 участников с результатом 35.857% и достаточно солидным отрывом от второго места (30.743%). Более подробное описание алгоритма можно найти в [11].

Задача классификации биомедицинских статей

На Международном соревновании «Topical Classification of Biomedical Research Papers» [14] была представлена задача классификации биомедицинских научных статей на $l = 83$, вообще говоря, пересекающихся класса. Каждый класс описывал принадлежность к определенной теме («хирургия», «стоматология» и т.п.). Каждая статья описывалась значениями $n = 25640$ признаков, природа признаков участникам соревнования не разглашалась, но скорее всего это были числа вхождений определенных терминов («трансплантация», «зуб» и т.п.). В обучающей выборке было $m = 10000$ статей, которые вручную раскласифицированы экспертами, в контрольной, по которой оценивалось качество алгоритмов, также было $q = 10000$ статей (их классификация участникам не была известна).

Функционалом качества решения была F-мера: качество ответа алгоритма $(y_1, \ldots, y_l)$ (бинарный вектор, i-й элемент которого – ответ на вопрос, принадлежит ли рассматрива-

емая статья i-му классу) при правильном ответе $(\alpha_1, \ldots, \alpha_l)$ вычислялась как

$$\frac{\sum_{i=1}^{l} \alpha_i y_i}{\sum_{i=1}^{l} \alpha_i + \sum_{i=1}^{l} y_i}.$$

Для определения победителя бралось среднее арифметическое значений F-меры для всех объектов контрольной выборки.

Достаточно качественные решения поставленной задачи (могли бы войти в 10 сильнейших на соревновании) можно получить следующим образом. Поскольку на вход алгоритму дается матрица $X_{m\times n}$, а на выходе – матрица $Y_{m\times l}$, то логично искать матрицу $W_{n\times l}$ такую, что

$$XW = Y. \tag{8}$$

Ясно, что при этом определяется слишком большое число nl неизвестных (параметров модели), что вызывает эффект переобучения (качество на тесте существенно меньше качества на обучении). Поэтому уравнение (8) «упрощается»: вместо матрицы X используют матрицу X', которая получается в результате неполного сингулярного разложения (SVD):

$$X \approx X'_{m\times k}\Lambda_{k\times k}U_{k\times n}.$$

Используя парадигму алгебраического подхода находят C:

$$C(X'W) = Y,$$

где C – пороговое решающее правило (5) (мы в любом случае вынуждены бинаризовать ответ). Наилучшее качество достигалось при использовании $k = 700$ максимальных собственных значений в сингулярном разложении, т.е. матрица X' имела размеры 10000×700.

Следуя универсальной формуле (1), необходимо было подобрать преобразование φ. Из логики решения и с помощью перебора было установлено, что лучше всего подходят преобразования типа нормировки. Наилучшее качество показала нормировка

$$\varphi(||\gamma_{ij}||) = ||\theta_{ij}||,\ \theta_{ij} = \frac{\gamma_{ij}}{\sqrt{\max\limits_{t}(\gamma_{it})}}.$$

Коэффициенты в (1) настраивались методом покоординатного спуска, напрямую оптимизируя F-меру решения. В качестве алгоритмов B_i в (1) были взяты

— Описанный выше линейный алгоритм над SVD-разложением.
— Два LIBSVM-алгоритма [15] (отличались предварительной нормировкой матрицы X, в первом все элементы делились на максимальный элемент в строке, во втором — в столбце).
— Алгоритм k ближайших соседей с косинусной мерой сходства (может быть исключён без сильной потери качества).

При решении задачи оказалось, что порог в пороговом решающем правиле (5) также можно выбирать, повышая качество классификации. Оптимальное значение порога – для i-й строки матрицы оценок $||\gamma_{ij}||$

$$c(i) = \min(\max_{j}(\gamma_{ij}), 0.345)$$

(необходимо было гарантировать отсутствие нулевых строк в матрице-ответе).

Таким образом, в этой задаче также удалось получить решение в виде (1). Здесь B_i брались из разных семейств алгоритмов, в остальном применение формулы (1) стандартное (каждый оператор B_i получал вещественную $q \times l$-матрицу, матрицы «деформировались» нормировками и складывались с определенными коэффициентами). Итоговое решение показало результат 0.53242 (по F-мере) и заняло 3 место на соревновании среди 126 участников (0.53579 — результат команды победителей). Подробнее о решении написано в [16].

Задача предсказания поведения клиентов сети супермаркетов

На Международном соревновании «Dunnhumby's Shopper Challenge» [17] была представлена задача прогнозирования даты следующего визита и суммы покупок каждого клиента сети супермаркетов. Техника решения самой задачи немного выходит за рамки данной статьи, поскольку ответ алгоритма для данного клиента считался верным, если точно угадывался день следующего визита и с точностью до 10$ угадывалась сумма. Таким образом, задача разбивалась на две разные подзадачи, которые требовалось решить согласованно (например, траты клиента могут зависеть от «типичности дня визита»).

Для нас интересно решение первой подзадачи — прогнозирования даты визита. Пусть у нас есть матрица визитов конкретного клиента

$$||v_{ij}||_{d\times 7},$$

в которой $v_{ij} = 1$, если был визит в j-й день i недель назад (если сейчас конец воскресения, то v_{12} соответствует вторнику этой недели, а v_{25} — пятнице прошлой), d — число недель, за которое есть статистика.

Есть несколько различных способов оценки вероятности следующего прихода в каждый конкретный день. Первый способ — оценить вероятности визитов

$$p_j = \frac{1}{d}\sum_{i=1}^{d} v_{ij}, \tag{9}$$

а затем пересчитать их в искомые вероятности

$$\tilde{p}_j^1 = p_j \prod_{k=1}^{j-1}(1 - p_k). \tag{10}$$

Второй способ — оценивать вероятности непосредственно:

$$\tilde{p}_j^2 = \frac{|\{i \in \{1, 2, \ldots, d\} \mid v_{ij} = 1,\ v_{i1} = \cdots = v_{i,j-1} = 0\}|}{d}.$$

Если от матрицы визитов $||v_{ij}||_{d\times 7}$ перейти к матрице первых визитов $||v'_{ij}||_{d\times 7}$, оставив только первые ненулевые элементы в каждой строке, то формула для второго метода запишется по аналогии с (9):

$$\tilde{p}_j^2 = \frac{1}{d}\sum_{i=1}^{d} v'_{ij}.$$

Это наблюдение будет в дальнейшем использовано следующим образом. Разумно комбинацией двух методов назвать метод, который вычисляет вероятности так:

$$\tilde{p}_j = \alpha\tilde{p}_j^1 + (1 - \alpha)\tilde{p}_j^2,\ \alpha \in [0, 1].$$

Однако, поскольку в решении будет использована деформация (которая должна «одинаково действовать на алгоритмы»), учитывая схожесть вычислений p_j и $\tilde{p}_j^2$, мы будет оценивать вероятности визитов по формуле

$$p_j^{\text{new}} = \alpha p_j + (1-\alpha)\tilde{p}_j^2,$$

а затем пересчитывать их в вероятности первых визитов (10). Это не совсем правильно с точки зрения теории, но дало выигрыш на практике. В качестве ответа выдается день, в который следующий визит максимально вероятен:

$$\arg\max_j \tilde{p}_j.$$

Далее логика решения простая. Надо определиться с «деформацией». Поскольку статистика дана за длительный период, надо учесть «устаревание» данных (элементы из первых строк матриц $||v_{ij}||$, $||v'_{ij}||$ более полезны для решения, чем элементы последних).

Используя простейшую весовую схему $w_1 \geqslant w_2 \geqslant \ldots \geqslant w_d \geqslant 0$, $\sum_{i=1}^d w_i = 1$, получаем оценку вероятности визита

$$p_j^{\text{new}} = \alpha \sum_{i=1}^d w_i v_{ij} + (1-\alpha) \sum_{i=1}^d w_i v'_{ij} = \sum_{i=1}^d w_i(\alpha v_{ij} + (1-\alpha)v'_{ij}).$$

В этой задаче в качестве весовой схемы была выбрана «степенная»:

$$w_i = \frac{(d-i+1)^r}{\sum_{k=1}^d k^r}$$

(хотя возможен и выбор других схем). Параметры r, α были выбраны так, чтобы максимизировать число верных прогнозов на последней неделе: статистика «обрезалась» 7 дней назад и предсказывались первые визиты каждого клиента (если они были) на последней неделе.

В результате был построен алгоритм, который занял первое место на соревновании среди 279 команд [17]. Алгоритм способен предсказывать день следующего визита с точностью 43%, но финальная версия показывала лишь 41.79%, поскольку параметры настраивались с учётом второй подзадачи (предсказывался не самый вероятный визит, поскольку учитывалась дисперсия цен покупок в эти дни недели).

Деформация ответов

«Деформации ответов» алгоритмов, т.е. применение к ответам некоторой функции φ часто используются прикладниками. При этом они называются по-разному («деформация» — авторский термин), например, «calibrating» [18]. Часто они бывают полезными при подобных функциях ошибки:

$$-\frac{1}{q}\sum_{i=1}^q \begin{cases} \log(1-a_i), & y_i = 0, \\ \log(a_i), & y_i = 1 \end{cases} \tag{11}$$

(Logarithmic Loss — логарифм функции правдоподобия распределения Бернулли), где y_i — верный ответ в задаче классификации с двумя классами $\{0,1\}$ на i-м объекте контрольной выборки из q объектов, а $a_i \in [0,1]$ — ответ алгоритма (он может лежать в отрезке).

Особенность функции (11) в том, что «она не прощает ошибок»: если $a_i = 1 - y_i$, то штраф равен бесконечности.

Если про объект известно, что с вероятностью p он принадлежит классу 1, тогда матожидание ошибки на нем

$$(1-p)\log(1-a) + p\log(a).$$

Приравняв к нулю производную, получаем, что $a = p$ — оптимальный ответ с точки зрения рассматриваемого функционала ошибки.

Таким образом, в случае функционала (11) надо выдавать вероятность. Поскольку алгоритмы не всегда вычисляют именно вероятность, они могут проигрывать простым наивным байесовским техникам по функционалу (11). Выход — перевод ответа в вероятность. Часто срабатывает такой прием: для $a \in [0,1]$ берутся все объекты из отложенного контроля (на них не обучался наш алгоритм), такие, что ответы нашего алгоритма на них попали в отрезок $[a-\varepsilon, a+\varepsilon]$. Функция φ в точке a полагается равной среднему арифметическому меток классов этих объектов (или взвешенной линейной комбинации, с весами, зависящими от расстояния). Подробнее об улучшении качества таким способом в реальных задачах анализа данных можно узнать из дипломной работы «Задачи анализа данных с нестандартным функционалом качества» дипломницы автора Ермушевой А.А. [19].

Таким образом, в случае некоторых функций ошибок, деформацию удается строить в явном виде (задавая значения в каждой точке), а потом использовать стандартную схему, описанную в этой работе. Интересно, что часто деформация похожа на сигмоиду (т.е. в нейронных сетях используется «правильная» функция активации — такое искривление ответов, как правило, и нужно для получения приемлемого качества).

Заключение

Мы описали универсальную форму представления функций и алгоритмов, а также способы её реализации на практике. Естественно, эти способы не претендуют на создание универсального алгоритма решения задач классификации, регрессии или построения рекомендаций. Тем не менее несколько побед на крупных соревнованиях по анализу данных автору удалось одержать благодаря такому «правильному построению алгоритма». При этом решались совершенно разные задачи: рекомендации, классификации и прогнозирования.

Доклад по схожей тематике впервые делался автором в 2012 г. на научной школе КРОМШ, статья с аналогичным теоретическим обзором (меньшим по объему и без доказательств) была подана в журнал «Spectral and Evolution Problems» (также содержала лишь один пример прикладной задачи).

В данной работе было представлено несколько реальных прикладных задач, впервые опубликовано решение задачи о визитах клиентов, предложены схемы доказательств теорем о представлении алгоритмов в виде (1) (для строгих доказательств потребовалось бы погружение в формалистику алгебраического подхода).

Когда работа была уже подписана в печать, с автором связался профессор А. Н. Горбань и сообщил, что результат, из которого следует, что для приближения непрерывной функции можно использовать нейросети с произвольной нелинейной функцией активацией (и тождественной) был получен еще в статье [20] (почти на 30 лет раньше В. Крейновича [4]). Автор выражает благодарность А. Н. Горбаню за внимание к этой работе.

Литература

[1] *Колмогоров А. Н.* О представлении непрерывных функций нескольких переменных в виде суперпозиции непрерывных функций одного переменного // *Докл. АН СССР*. 1957. Т. 114, №. 5. С. 953–956.

[2] *Хайкин С.* Нейронные сети. Полный курс. — М.: Вильямс, 2006.

[3] *Горбань А. Н.* Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей // *Сибирский журнал вычислительной математики*. 1998. Т. 1, №. 1. С. 12–24.

[4] *Kreinovich V. Y.* Arbitrary nonlinearity is sufficient to represent all functions by neural networks: A theorem // *Neural Networks*. 1991. Vol. 4, no. 3. C. 381–383.

[5] *Leshno M., Lin V. Ya., Pinkus A., Schocken S.* Multilayer feedforward networks with a nonpolynomial activation function can approximate any function // *Neural Networks*. 1993. No. 6. P. 861–867.

[6] *Pinkus A.* TDI-subspaces of $C(\mathbb{R}^d)$ and some density problems from neural networks // *Journal of Approximation Theory*. 1996. Vol. 85. P. 269–287.

[7] *Pinkus A.* Density in approximation theory // *Surveys in Approximation Theory*. 2005. No. 1. P. 1–45.

[8] *Журавлёв Ю. И.* Корректные алгоритмы над множествами некорректных (эвристических) алгоритмов. I–II // *Кибернетика*. 1977. №. 4. С. 5–7; 2011. №. 6. С. 21–27.

[9] *Дьяконов А. Г.* Теория систем эквивалентностей для описания алгебраических замыканий обобщенной модели вычисления оценок // *Ж. вычисл. матем. и матем. физ.* 2010. Т. 50, №. 2. С. 388–400; 2011. Т. 51, №. 3. С. 529–544.

[10] *Дьяконов А. Г.* Алгебра над алгоритмами вычисления оценок: нормировка и деление // *Ж. вычисл. матем. и матем. физ.* 2007. Т. 47, №. 6. С. 1099–1109.

[11] *Дьяконов А. Г.* Алгоритмы для рекомендательной системы: технология LENKOR // *Бизнес-Информатика*. 2012. №. 1(19). С. 32–39.

[12] *Antulov-Fantulin N., Bošnjak M., Šmuc T., Jermol M., Žnidaršič M., Grčar M., Keše P., Lavrač N.* ECML/PKDD 2011 — Discovery challenge: "VideoLectures.Net Recommender System Challenge." 2012. `http://tunedit.org/challenge/VLNetChallenge`.

[13] *Маннинг К. Д., Рагхаван П., Шютце Х.* Введение в информационный поиск. М.: Вильямс, 2011.

[14] Международное соревнование "JRS 2012 Data Mining Competition: Topical Classification of Biomedical Research Papers." 2012. `http://tunedit.org/challenge/JRS12Contest`.

[15] Библиотека для решения задач классификации и регрессии на базе метода SVM. 2012. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

[16] *D'yakonov A.* A blending of simple algorithms for topical classification // *Rough Sets and Current Trends in Computing, Lecture Notes in Computer Science*. 2012. Vol. 7413. P. 432–438.

[17] Международное соревнование "dunnhumby's Shopper Challenge." 2012. `http://www.kaggle.com/c/dunnhumbychallenge`.

[18] *Bostrom H.* Calibrating random forests // *Proceedings of the 2008 7th International Conference on Machine Learning and Applications*. 2008. P. 121–126.

[19] Страница спецсеминара Дьяконова А. Г. на факультете ВМК МГУ. 2012. `http://www.machinelearning.ru/wiki/index.php?title=Алгебра_над_алгоритмами_и_эвристический_поиск_закономерностей`.

[20] *Leeuw K., Katznelson Y.* Functions that operate on non-self-adjoint algebras // *J. d'Anal. Math.* 1963. Vol. 11. P. 207–219.

Интеллектуальный анализ данных и знаний по стентированию коронарных артерий*

Янковская А. Е.[1,2]***, Китлер С. В.***[3]
ayyankov@gmail.com, svkitler@gmail.com
[1]Томский государственный университет; [2]Томский государственный архитектурно-строительный университет; [3]Томский государственный университет систем управления и радиоэлектроники

Статья посвящена интеллектуальному анализу данных и знаний по стентированию коронарных артерий. Излагаются основные этапы интеллектуального анализа данных и знаний по стентированию коронарных артерий. Описывается подход к анализу данных и знаний по стентированию коронарных артерий, реализованный в интеллектуальной системе. Интеллектуальная система разработана в виде динамически подключаемых модулей к интеллектуальному инструментальному средству ИМСЛОГ, на базе которого конструируются прикладные интеллектуальные системы. Приводятся результаты исследования интеллектуальной системы.

Intelligent Analysis of Data and Knowledge for Stenting of Coronary Artery*

Yankovskaya A. E.[1,2]***, Kitler S. V.***[3]
[1]Tomsk State University; [2]Tomsk State University of Architecture and Building; [3]Tomsk State University of Control Systems and Radioelectronics

The paper is devoted to intelligent analysis of data and knowledge for stenting of coronary artery. The basic stages of the intelligent analysis of data and knowledge for stenting of coronary artery are described. An approach for analysis of data and knowledge for stenting of coronary artery is presented. Software realization of the approach is implemented in the intelligent system. The intelligent system is developed as dynamical plug-ins in the intelligent instrumental software (IIS) IMSLOG. The applied intelligent systems are constructing on the base of the IIS IMSLOG. Aprobation results of the intelligent system is given.

Введение

В настоящее время весьма актуально развитие технологии интеллектуального анализа данных и знаний [1–5]. Необходимость интеллектуальной обработки и управления знаниями в слабоструктурированных областях, таких как медицина, психология, социология, геология, биомедицина обусловлена увеличением объема обрабатываемых данных и знаний, используемых при решении диагностических задач [1].

Одной из наиболее значимых медицинских областей несомненно является кардиология, включающая такие важные задачи как диагностика ишемической болезни сердца, диагностика окклюзий и рестенозов. Для их решения разработаны и внедрены различные системы [6–8]. Системы «Fuzzy expert system approach for coronary artery disease screening using clinical parameters» [6] и «A multilayer perceptron-based medical decision support system for heart disease diagnosis» [8] основаны на мягких вычислениях и предназначены для обнаружения и прогнозирования ишемической болезни сердца на ранней стадии и для диагно-

Работа частично выполнена при финансовой поддержке РФФИ, проекты №13-07-00373, №13-07-98037-р_сибирь_а, №12-07-31109_мол-а и РГНФ №13-06-00709.

стики сердечно-сосудистых заболеваний соответственно, система «Knowledge Management System for Clinical Decision Support – Application in Cardiology» [7] основана на статистических методах и используется для поддержки принятия решения по лечению больных с острым коронарным синдромом.

В отличие от вышеперечисленных систем в настоящем исследовании предлагается интеллектуальная система, основанная на логико-комбинаторных методах распознавания образов и предназначенная для диагностики окклюзий и рестенозов коронарных артерий.

Интеллектуальный анализ данных и знаний по стентированию коронарных артерий включает в себя формирование базы данных и знаний по исследуемой проблемной области, подход к анализу данных и знаний и его реализацию в интеллектуальной системе.

Следует отметить, что заменив данные и знания из рассматриваемой области на данные и знания из другой проблемной области, предлагаемая интеллектуальная система будет ориентирована на анализ данных и знаний из вновь заполненной проблемной области, т.е. система не является узконаправленной.

В статье приводятся основные понятия и определения, излагается подход к выявлению различного рода закономерностей в данных и знаниях, описывается интеллектуальная система анализа данных и знаний по стентированию коронарных артерий, приводятся результаты её исследования.

Основные понятия и определения

Для дальнейшего изложения воспользуемся понятиями и определениями, введенными в публикациях [1, 9].

Формирование базы данных и знаний осуществляется на основе матричной модели представления данных и знаний [9], включающей целочисленную матрицу описаний ($\mathbf{Q}$), задающую описание объектов в пространстве k–значных характеристических признаков $z_1, \ldots, z_m$ и целочисленную матрицу различий ($\mathbf{R}$), задающую разбиение объектов на классы эквивалентности по каждому механизму классификации. Если значение признака несущественно для объекта, то данный факт отмечается прочерком («–») в соответствующем элементе матрицы $\mathbf{Q}$. Для каждого признака z_j ($j \in \{1, 2, \ldots, m\}$) задаётся интервал изменения его значений.

Множество всех неповторяющихся строк матрицы $\mathbf{R}$ сопоставлено множеству выделенных образов, представленных одностолбцовой матрицей $\mathbf{R}'$, элементами которой являются номера образов.

На рисунке 1 приведена матричная модель представления данных и знаний.

$$\boldsymbol{Q}=\begin{array}{c}\begin{matrix}z_1 & z_2 & z_3 & z_4 & z_5 & z_6 & z_7 & z_8 & z_9 & z_{10} & z_{11}\end{matrix}\\ \begin{bmatrix}1 & 4 & 6 & 3 & 2 & 2 & 1 & 2 & 3 & 4 & 1\\ 4 & 4 & 5 & 2 & 3 & 2 & 7 & 8 & 3 & 4 & 1\\ 3 & 4 & 5 & 3 & 3 & 2 & 4 & 5 & 3 & 4 & 1\\ 3 & 4 & 4 & 1 & 4 & 4 & 2 & 3 & 1 & 5 & 1\\ 2 & 4 & 2 & 1 & 6 & 3 & 4 & 5 & 2 & 3 & 1\\ 2 & 4 & 5 & 1 & 3 & - & 4 & 3 & 1 & 2 & 1\\ 1 & 4 & 3 & 2 & 5 & 2 & 1 & 2 & 3 & 4 & 1\\ 3 & 4 & 2 & 2 & 6 & 2 & 2 & 3 & 3 & 2 & 1\\ 5 & 4 & 2 & 2 & 6 & 3 & 5 & 6 & 2 & 4 & 1\\ 4 & 4 & 6 & 1 & 2 & 5 & 5 & 6 & 1 & 4 & 2\end{bmatrix}\end{array}\begin{matrix}1\\2\\3\\4\\5\\6\\7\\8\\9\\10\end{matrix}\quad \boldsymbol{R}=\begin{array}{c}\begin{matrix}k_1 & k_2\end{matrix}\\ \begin{bmatrix}1 & 2\\ 1 & 2\\ 1 & 2\\ 2 & 1\\ 2 & 1\\ 2 & 1\\ 1 & 3\\ 1 & 3\\ 3 & 2\\ 3 & 2\end{bmatrix}\end{array}\quad \boldsymbol{R'}=\begin{bmatrix}1\\1\\1\\2\\2\\2\\3\\3\\4\\4\end{bmatrix}$$

Рис. 1: Матричная модель представления данных и знаний.

Данная модель позволяет представлять не только данные, но и знания экспертов, поскольку одной строкой матрицы **Q** можно задавать в интервальной форме подмножество объектов, для которых характерны одни и те же итоговые решения, задаваемые соответствующими строками матрицы **R**.

Диагностическим тестом (ДТ) называется совокупность признаков, различающих любые пары объектов, принадлежащих разным образам.

Диагностический тест называется безызбыточным (тупиковым [1]), если содержит безызбыточное количество признаков.

Безызбыточный безусловный диагностический тест (ББДТ) характеризуется одновременным предъявлением всех входящих в него признаков исследуемого объекта при принятии решений.

Для представления диагностических тестов будем использовать двоичную матрицу тестов (**T**) [10], столбцы которой сопоставлены столбцам матрицы **Q**, строки — диагностическим тестам. Единичными значениями в каждой строке матрицы **T** отмечены признаки, вошедшие в соответствующий данной строке диагностический тест.

На рисунке 2 приведен пример построенной двоичной матрицы отказоустойчивых диагностических тестов **T**, то есть устойчивых к заданному числу t ошибок измерения (занесения) значений характеристических признаков исследуемых объектов [11].

$$\boldsymbol{T} = \begin{array}{c} \begin{matrix} z_1 & z_2 & z_3 & z_4 & z_5 & z_6 & z_7 & z_8 & z_9 & z_{10} & z_{11} \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \end{matrix} \end{array}$$

Рис. 2: Двоичная матрица отказоустойчивых диагностических тестов.

Закономерности в данных и знаниях

Понятие закономерностей в исследованиях различных школ [1, 3] отличается от терминологии, предложенной А. Е. Янковской [9] и используемой в настоящем исследовании.

Под закономерностями понимаются подмножества признаков с определенными легко интерпретируемыми свойствами, влияющими на различимость объектов из разных образов, устойчиво наблюдаемыми для объектов из обучающей выборки и проявляющимися на других объектах той же природы, а также весовые коэффициенты признаков, характеризующие их индивидуальный вклад [9] в различимость объектов и информационный вес, определяемый в отличие от [12] на подмножестве тестов, используемых для принятия итогового решения [9].

К упомянутым подмножествам будем относить константные (принимающие одно и то же значение для всех образов), устойчивые (константные внутри образа, но не являющиеся константными), неинформативные (не различающие ни одной пары объектов), альтернативные (в смысле включения в ДТ), зависимые (в смысле включения подмножеств

различимых пар объектов), несущественные (не входящие ни в один безызбыточный ДТ), обязательные (входящие во все безызбыточные ДТ), псевдообязательные (не являющиеся обязательными, но входящие во все ББДТ, участвующие в принятии решений) признаки, подмножества сигнальных признаков первого (различающие исследуемые объекты из двух образов) и второго рода (различающие исследуемые объекты, принадлежащие одному образу, с объектами, принадлежащими другому образу), а также все минимальные и все (либо часть при большом признаковом пространстве) безызбыточные различающие подмножества признаков, являющиеся, по сути, соответственно минимальными и безызбыточными ДТ [13]. Отказоустойчивые ББДТ также входят в число закономерностей [11].

Для выявления различного рода закономерностей при построении ББДТ применяется процедура построения матрицы импликаций ($\mathbf{U}$) [9], задающей различимость объектов из разных образов.

Матрица $\mathbf{U}$ представляет собой целочисленную матрицу, столбцы которой сопоставлены столбцам матрицы $\mathbf{Q}$, а строки — всевозможным парам объектов v, l, соответственно из разных образов (классов) a, b; $v \in \{1, 2, \ldots, \sigma(\mathbf{Q}^a)\}$, $l \in \{1, 2, \ldots, \sigma(\mathbf{Q}^b)\}$, где $\sigma(\mathbf{Q}^a)$ $(\sigma(\mathbf{Q}^b))$ — количество строк в подматрице $\mathbf{Q}^a$ $(\mathbf{Q}^b)$ матрицы $\mathbf{Q}$. Строка $\mathbf{U}_i$ матрицы $\mathbf{U}$ представляет собой значение целочисленной вектор-функции различения, j-я ($j \in \{1, 2, \ldots, m\}$) компонента u_{ij} которой вычисляется по формуле:

$$u_{i,j} = |q^a_{v,j} - q^b_{l,j}|, \tag{1}$$

где $q^a_{v,j}$ $(q^b_{l,j})$ — значение признака z_j для объекта v (l); а $u_{i,j}$ вычисляется для каждой пары образов,

$$i = \prod_{r=1}^{a} \sigma(\mathbf{Q}^r) \sum_{s=r+1}^{b-1} \sigma(\mathbf{Q}^s) + \sum_{r=1}^{v-1} \sigma(\mathbf{Q}^b) + vl.$$

Будем говорить, что строка $\mathbf{U}_d$ поглощает строку $\mathbf{U}_\rho$ $(\mathbf{U}_d \succ \mathbf{U}_\rho)$, если

$$(\mathbf{U}_d \succ \mathbf{U}_\rho) \leftrightarrow \forall i \in I(u_{di} \geqslant u_{\rho i}).$$

Безызбыточной матрицей импликаций называется матрица $\mathbf{U}'$, в которой отсутствуют поглощающие строки.

Далее воспользуемся понятиями константных, устойчивых, альтернативных и зависимых признаков, приведенными в публикации [10].

Если элементы в рассматриваемом столбце матрицы $\mathbf{U}'$ принимают одинаковое значение, то признак, соответствующий этому столбцу, является константным. Иными словами константный признак соответствует столбцу матрицы $\mathbf{Q}$ с одинаковыми значениями элементов.

Устойчивый признак соответствует столбцу матрицы $\mathbf{Q}$, содержащему одинаковые элементы внутри одного образа, но не являющийся константным.

Альтернативные признаки соответствуют одинаковым столбцам матрицы $\mathbf{U}'$.

Зависимые признаки соответствуют столбцам матрицы $\mathbf{U}'$, которые поглощаются другим столбцом (другими столбцами) матрицы $\mathbf{U}'$.

Одной из наиболее важных проблем при выявлении закономерностей и принятии решений является анализ данных и знаний, включающий выявление наиболее значимых признаков и оценивание величины их значимости [9].

Приведём формулы для вычисления значения весового коэффициента k–значного j-го признака (w_j) и вычисления значения веса i-го теста (W_i), данные в статье [14], и являющиеся модификацией формул, приведённых в публикации [9] для булевых признаков:

w_j — весовой коэффициент целочисленного признака z_j ($j \in \{1, 2, \ldots, m\}$);

q_{aj} — значение признака z_j для объекта из образа a;

δ^j_{ab} — величина j-й компоненты вектор-функции различения между объектами из разных образов a, b ($a \neq b$), вычисляемая по формуле:

$$\delta^j_{a,b} = |q_{a,j} - q_{b,j}|;$$

N_a — число строк в описании a-го образа;

K — число выделенных образов;

S_j — интервал изменения j-го признака, сопоставленного j-му столбцу матрицы $\mathbf{Q}$;

p_l — коэффициент повторения l-ой строки, задаваемый извне;

$$d^-_{lj} = \begin{cases} 1, & \text{если } q_{lj} = \text{«–»} \\ 0, & \text{если } q_{lj} \neq \text{«–»}; \end{cases}$$

σ_a — количество объектов в образе a ($a = 1, \ldots, K$), вычисляемое по формуле:

$$\sigma_a = \sum_{l=1}^{N_a} p_l \prod_{j}^{m} (S_j)^{d^-_{lj}};$$

r, s ($r \neq s$) и a, b ($a \neq b$) — номера образов;

L_i — множество признаков, входящих в i-ый тест.

Весовой коэффициент w_j целочисленного признака z_j [11] будем определять по его разделяющей способности аналогично вычислению w_j для двоичного признака z_j [9].

$$w_j = \frac{\sum\limits_{r=1}^{K-1} \sum\limits_{s=r+1}^{K} \sum\limits_{a=1}^{N_r} \sum\limits_{b=1}^{N_s} \delta^j_{ab}}{S_j \sum\limits_{a=1}^{K-1} \sum\limits_{b=a+1}^{K} \sigma_a \sigma_b}. \tag{2}$$

Множество обязательных признаков будем называть ядром всех диагностических тестов, поскольку исключение любого признака из ядра нарушает свойство каждого из тестов быть тестом [9].

Каждому i-му тесту соответствует вес теста W_i [9], вычисляемый по формуле:

$$W_i = \sum_{j \in L_i} w_j.$$

Основные этапы интеллектуального анализа данных и знаний

Интеллектуальный анализ данных и знаний сводится к выявлению различного рода закономерностей.

Дано: матрицы $\mathbf{Q}$ и $\mathbf{R}$ (рис. 1) в пространстве k–значных характеристических признаков и диапазон изменения их значений.

Необходимо: построить двоичную матрицу $\mathbf{U}''$, найти все ее безызбыточные столбцовые покрытия и соответствующие множества характеристических признаков, вошедших в k-значные ББДТ.

Кратко изложим основные этапы интеллектуального анализа данных и знаний [9, 11, 15].

1. Формирование матрицы $\mathbf{R}'$ по матрице $\mathbf{R}$.
2. Упорядочивание строк матриц $\mathbf{Q}$ и $\mathbf{R}'$ по принадлежности к образам.
3. Построение по матрицам $\mathbf{Q}$ и $\mathbf{R}'$ матрицы $\mathbf{U}'$ (формула 1) с одновременным удалением поглощающих строк и вычислением весовых коэффициентов признаков по формуле 2.
4. Построение двоичной матрицы $\mathbf{U}''$ путём замены значений всех элементов, отличных от нуля, в матрице $\mathbf{U}'$ на значение «1».
5. Выявление различного рода закономерностей по матрице $\mathbf{U}''$.
6. Построение всех безызбыточных столбцовых покрытий матрицы $\mathbf{U}''$ либо части при превышении числа безызбыточных столбцовых покрытий в процессе их построения наперед заданной величины g.
7. Построение по найденным безызбыточным столбцовым покрытиям матрицы $\mathbf{U}''$ всех ББДТ по алгоритму, приведенному в публикации [16].

Описание интеллектуальной системы анализа данных и знаний

Интеллектуальная система анализа данных и знаний предназначена для выявления различного рода закономерностей. Интеллектуальная система разработана в виде динамически подключаемых модулей к интеллектуальному инструментальному средству (ИИС) ИМСЛОГ [17], на базе которого конструируются прикладные интеллектуальные системы.

Архитектура ИИС ИМСЛОГ является открытой и представляет собой иерархическую систему программных модулей, разработанных с использованием средств Builder C++.

Один модуль реализован как резидентный, имеет встроенную систему команд, выполняет функции координирующего центра и называется ядром. Все остальные модули являются динамически подключаемыми, называются плагинами и подразделяются на функциональные модули, модули системных данных и базовый модуль интеллектуального пользовательского интерфейса.

Интеллектуальная система анализа данных и знаний по стентированию коронарных артерий (рис. 3) представляет собой связанный набор функциональных модулей.

Первый сверху модуль предназначен для работы с базой данных и знаний. Выходными данными модуля являются структура базы знаний, объекты базы знаний, название признаков и номера признаков.

Второй модуль предназначен для выбора признаков, включаемых в матрицы $\mathbf{Q}$ и $\mathbf{R}$.

Третий модуль, получая на вход номера целочисленных характеристических признаков, номера целочисленных классификационных признаков, структуру базы знаний, номера выбранных обучающих объектов и соответствующие им описания объектов из базы знаний, строит матрицы $\mathbf{Q}$ и $\mathbf{R}$. Также выходными данными этого модуля являются вектор номеров выбранных характеристических признаков, минимальные и максимальные значения этих признаков.

Четвертый сверху модуль предназначен для построения матрицы $\mathbf{U}'$. В этом модуле используются следующие входные параметры: целочисленные матрицы $\mathbf{Q}$ и $\mathbf{R}$, вектор номеров характеристических признаков, а также вектора минимальных и максимальных значений этих признаков. В результате исполнения модуля формируются выходные значения, содержащие вещественный вектор весовых коэффициентов характеристических признаков и двоичную матрицу $\mathbf{U}''$, являющуюся входными параметрами для модуля поиска различного рода закономерностей. Выходными параметрами 4-го модуля являются: матрица $\mathbf{U}''$, неинформативные, обязательные, альтернативные, зависимые характеристические признаки, а также весовые коэффициенты всех характеристических признаков.

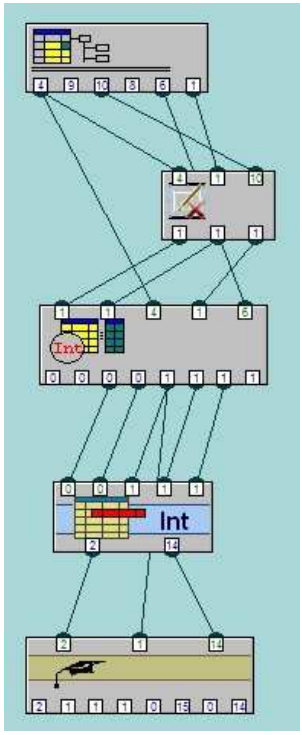


Рис. 3: Интеллектуальная система анализа данных и знаний.

Отметим, что в настоящее время не реализована процедура выявления минимальных подмножеств сигнальных признаков.

Исследование интеллектуальной системы

Исследование интеллектуальной системы связано с возможным сокращением признакового пространства. Сокращение признакового пространства позволит сократить число выявляемых значений характеристических признаков для новых исследуемых объектов (обследуемых пациентов), а также сократить временные и стоимостные затраты на исследование и построение диагностических тестов и принятие решений на их основе.

Первые исследования интеллектуальной системы по выявлению различного рода закономерностей по стентированию коронарных артерий описаны в публикации [18]. Исследования проводились по классификационным признакам: наличие окклюзии стента и срок окклюзии стента, наличие рестеноза за всё время наблюдения и срок рестеноза, смерть пациента и время смерти.

Система «A multilayer perceptron-based medical decision support system for heart disease diagnosis» [8] включает 40 характеристических признаков и обучающую выборку, состоящую из 352 обучающих объектов, разбитых на 5 классов (используется только один механизм классификации). При этом размер матрицы $\mathbf{Q}$ равен 352x40 (14080), а размер матрицы $\mathbf{R}$ равен 352x1. В отличие от вышеупомянутой системы, предлагаемая интеллектуальная система включает обучающую выборку, количество обследуемых в которой составило 381 человек. Количество характеристических признаков в описании исследуемых объектов (обследуемых) равно 165. Таким образом, размер матрицы $\mathbf{Q}$ равен 381x165 (62865). Описание же самого признакового пространства приведено в публикации [18].

В настоящей статье приводятся два исследования по выявлению различного рода закономерностей. В первом исследовании использовались следующие механизмы классификации: 1) наличие окклюзии стента (3 класса); 2) смерть пациента (2 класса). При этом

размер матрицы $\mathbf{R_1}$ равен 381x2 (762); Во втором исследовании использовались механизмы классификации: 1) наличие рестеноза за всё время наблюдения (3 класса); 2) смерть пациента (2 класса). Размер матрицы $\mathbf{R_2}$, как и в первом исследовании, равен 381x2 (762). Число образов по каждому из исследований равно 6.

В результате по первому исследованию была построена матрица $\mathbf{U}''$ размером 27352x165 и были выявлены следующие закономерности: признак 155 (дата второй повторной операции) является неинформативным; признак 134 (длина второго стента правой коронарной артерии) зависим от признака 139 (длина первого стента диагональной артерии); признак 30 (количество установленных стентов с лекарственным покрытием) зависим от признака 28 (применение баллона со сжатым воздухом); признак 23 (наличие стентирования при прогрессировании атеросклероза) зависим от признака 31 (вид вмешательства). Альтернативных и обязательных признаков в данном исследовании обнаружено не было.

После удаления зависимых и неинформативных признаков признаковое пространство сократилось до 161 признака.

В результате по второму исследованию была построена матрица $\mathbf{U}''$ размером 26149x165 и были выявлены следующие закономерности: признак 155 (дата второй повторной операции) является неинформативным; признак 30 (количество установленных стентов с лекарственным покрытием) зависим от признака 28 (применение баллона со сжатым воздухом); признак 23 (наличие стентирования при прогрессировании атеросклероза) зависим от признака 31 (вид вмешательства); признак 33 (установка стента, покрытого лекарством «сиролимус») зависим от признака 37 (установка стента, покрытого лекарством «эверолимус»). Альтернативных и обязательных признаков в данном исследовании обнаружено не было.

После удаления зависимых и неинформативных признаков, как и в первом исследовании, признаковое пространство сократилось до 161 признака.

Заключение

Дан кратко обзор литературы, посвященный созданным системам по диагностике ишемической болезни сердца. Показано их отличие от предложенной интеллектуальной системы анализа данных и знаний по стентированию коронарных артерий.

Приведены основные этапы интеллектуального анализа данных и знаний по стентированию коронарных артерий, реализованные в интеллектуальной системе.

Описана интеллектуальная система анализа данных и знаний и приведены результаты её исследования. Результаты исследования показали, что в первом и во втором исследованиях признаковое пространство сократилось на 4 признака, что в ряде случаев существенно снижает временные затраты на построение диагностических тестов и принятие решений на их основе при диагностике рестенозов и окклюзий коронарных артерий.

Применение интеллектуального анализа данных и знаний по стентированию коронарных артерий позволит снизить риски возникновения неблагоприятных сердечно-сосудистых событий и оптимизировать лечебную тактику при длительном наблюдении за больными, подвергшимся инвазивным вмешательствам на коронарных артериях.

Дальнейшие исследования связаны с реализацией процедуры выявления минимального подмножества сигнальных признаков и алгоритма построения отказоустойчивых тестов [11] в предложенной интеллектуальной системе; с применением подхода к анализу данных и знаний в других проблемных областях, например, для диагностики психических забо-

леваний [19], а также с созданием гибридной интеллектуальной обучающе-тестирующей системы [20] с применением вышеопривёдённого подхода к анализу данных и знаний.

Коллектив авторов выражает благодарность профессору д.м.н. заслуженному деятелю наук РФ научному руководителю отделения сердечной недостаточности ФГБУ НИИ кардиологии СО РАМН А. Т. Теплякову и к.м.н. старшему научному сотруднику отделения сердечной недостаточности ФГБУ НИИ кардиологии СО РАМН Е. В. Граковой за предоставление данных по стентированию коронарных артерий и консультации по вопросам стентирования.

Литература

[1] *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* Распознавание. Математические методы. Программная система. Практические применения. — Москва: Фазис, 2006. —176 с.

[2] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Н-ск: ИМ СО РАН, 1999. — 270 с.

[3] *Закревский А. Д.* Логика распознавания. Изд. 2-е, доп. — М.: Едиториал УРСС, 2003. — 144 с.

[4] *Финн В. К.* Об интеллектуальном анализе данных. // Новости Искусственного интеллекта. — 2004. — № 3. — С. 3–18.

[5] *Янковская А. Е., Гедике А. И.* Интеллектуальный анализ информации на базе инструментального средства ИСМЛОГ // Новости искусственного интеллекта. —2005. —№ 1. — С. 36–47.

[6] *Pal D., Mandana K. M., Pal S., Sarkar D., Chakraborty C.* Fuzzy expert system approach for coronary artery disease screening using clinical parameters // Knowledge-Based Systems. — 2012. — Vol. 36. — Pp. 162–174.

[7] *Kostic P., Vasiljevic Z., Pavlovic S., Milosavljevic I., Milanovic J. G.,Blesic S. and Milanovic S.* Knowledge Management System for Clinical Decision Support – Application in Cardiology // Proceedings of 19th Telecommunications Forum (TELFOR), 2011. — Pp. 1261-1264.

[8] *Hongmei Yana, Yingtao Jiangb, Jun Zhenge, Chenglin Pengc, Qinghui Li* A multilayer perceptron-based medical decision support system for heart disease diagnosis // Expert Systems with Applications. — 2005. — Vol. 30, № 2. — Pp. 272–281.

[9] *Янковская А. Е.* Логические тесты и средства когнитивной графики // Издательский Дом: LAP LAMBERT Academic Publishing. — 2011. — 92 с.

[10] *Янковская А. Е., Петелин А. Е.* Развитие алгоритма многокритериального выбора оптимального подмножества диагностических тестов // Математические методы распознавания образов (ММРО-14). Сборник докладов 14-ой Всероссийской конференции. — М.: МАКС Пресс, 2009. — С. 212–215.

[11] *Yankovskaya A. E., Kitler S. V.* Parallel Algorithm for Constructing k–Valued Fault-Tolerant Diagnostic Tests in Intelligent Systems // Pattern Recognition and Image Analysis. — 2012. — Vol. 22, № 3. — Pp. 473–482.

[12] *Кренделев Ф. П., Дмитриев А. Н., Журавлев Ю. И.* Сравнение геологического строения зарубежных месторождений докембрийских конгломератов с помощью дискретной математики // Доклады АН СССР. — Т. 173, № 5. — 1967. — С. 1149–1152.

[13] *Yankovskaya A. E.* New Kinds of Regularities in Knowledge and Algorithms of Their Revealing // 7th Open German/Russian Workshop on Pattern Recognition and Image Understanding. August 20-23, 2007 Ettlingen, Germany.

[14] *Янковская А. Е., Китлер С. В.* Принятие решений на основе параллельных алгоритмов тестового распознавания образов // Искусственный интеллект. — 2010. — № 3. — С. 151–159.

[15] *Yankovskaya A. E., Semenov M. E.* To the problem about the intelligent extension construction of the geoinforma-tional systems // Pattern Recognition and Image Understanding (OGRW-8-

11). Proceedings of 8th Open German-Russian Workshop. — Nizhny Novgorod: Nizhny Novgorod Lobachevsky State University, 2011. — Pp. 349–352.

[16] *Гедике А. И., Янковская А. Е.* Построение всех безызбыточных безусловных диагностических тестов в интеллектуальном инструментальном средстве ИМСЛОГ // Интеллектуальные системы (AIS'05), Интеллектуальные САПР (CAD-2005). Труды Между-народных научно-технических конференций. Том 1. — Москва: Физматлит, 2005. — С. 209–214.

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

[18] *Янковская А. Е., Китлер С. В., Аметов Р. В., Гракова Е. В.* Информационная технология выявления закономерностей в данных и знаниях по стентированию коронарных артерий // Труды пятой международной конференции «Системный анализ и информационные технологии» САИТ-2013. — 2013 (в печати).

[19] *Янковская А. Е., Китлер С. В., Изюмов А. А., Кривдюк Н. М., Силаева А. В.* Основы создания комплекса интеллектуальных систем диагностики и профилактики депрессии // Приложение к журналу «Вестник Московского университета имени С.Ю. Витте. Серия 1: Экономика и управление», 2013. — С. 81–84.

[20] *Yankovskaya A. E., Semenov M. E.* Application Mixed Diagnostic Tests in Blended Education and Training // Proceedings of the IASTED International Conference Web–based Education (WBE 2013) February 13–15, 2013 Innsbruck, Austria. — 2013. — Pp. 935-939.

Определение точной границы зрачка*

Чинаев Н. Н., Матвеев И. А.
nikolaj.chinaev@phystech.edu
Московский физико-технический институт, Вычислительный Центр им. А. А. Дородницына РАН

Предлагается метод определения точной границы зрачка на монохромном изображении глаза. Метод основан на бинаризации изображения с последующим поиском зрачка как одной из компонент связности. Граница зрачка определяется как граница или часть границы компоненты связности. Для отделения зрачка в случае его объединения в одну компоненту связности с другими объектами, а также для проверки правдоподобия решения используется преобразование Хафа. Приведены статистические результаты, показывающие точность работы метода; в качестве тестовых данных использованы изображения из открытой базы данных.

Ключевые слова: *выделение радужки, поиск границы зрачка, преобразование Хафа, обработка изображений.*

Precise pupil border locating*

Chinaev N. N., Matveev I. A.
Moscow institute of physics and technology, Dorodnitsyn Computer Centre of RAS

Method of locating precise pupil border in monochrome images of eye is proposed. The method is based on image binarisation followed by detection of pupil as one of the connected components. Pupil border is estimated as a boundary or part of the boundary of the connected component. Hough transform is employed to separate pupil in case of its merging to one component with other objects and to verify the plausibility of the detection. Statistical results illustrating performance of the method are presented. Images from public domain database were used for tests.
Keywords: *iris segmentation, pupil boundary detection, Hough transform, image processing.*

Введение

Задача автоматического поиска границы зрачка на фотографии глаза возникает на ранних стадиях обработки входных изображений в системах безопасности и медицинской диагностики. Существует несколько подходов к решению задачи: одним из пионеров в этой области является Даугман [1], предложивший интегро-дифференциальный оператор, который может быть рассмотрен как разновидность преобразования Хафа. Метод работает надёжно, но требует больших вычислительных затрат, а также может работать ошибочно, если на изображении присутствуют шумы, например, такие как блики. Различные виды преобразования Хафа также нашли широкое применение в решении этой задачи [2, 3, 4]. Суть этого преобразования заключается в следующем: при поиске на изображении некоторого объекта определённого вида (окружность, прямая и т.д.) принимается в рассмотрение аккумуляторное пространство — пространство параметров, которыми определяется искомый объект. Для окружности произвольного радиуса аккумуляторное пространство трёхмерно, так как окружность задаётся тремя параметрами, например двумя координатами центра и радиусом. Производится процедура голосования: каждой точке изобра-

Работа выполнена при финансовой поддержке РФФИ (проект 12-07-00778)

жения ставится в соответствие некоторое множество из аккумуляторного пространства, состоящее из комбинаций параметров объектов, которым эта точка могла бы принадлежать. В качестве ответа выдаётся точка из аккумуляторного пространства, попавшая в наибольшее число таких множеств. При поиске зрачка использовалось также двумерное аккумуляторное пространство [2].

Часто применяется бинаризация изображения — устанавливается некоторый порог интенсивности таким образом, чтобы все точки, обладающие меньшей интенсивностью, образовывали связное множество небольшого размера. Такой подход связан с предположением, что зрачок — наиболее тёмная область на фотографии. Это предположение не всегда оправдывается, поскольку часто после установки порога бинаризации остаётся несколько компонент связности, так как яркость бровей или ресниц может мало отличаться от яркости зрачка или даже быть ниже. В работе [5] предложены локальные статистические признаки, вычисляемые в областях небольшого размера на изображении, которые имеют характерно разные значения для небольших областей, перекрывающих границу зрачка или радужки, и областей, лежащих внутри зрачка, радужки или склеры. Используется метод активного контура, где вводится некоторое множество точек, являющихся вершинами замкнутой ломаной, которая может деформироваться и перемещаться по изображению под действием вычисляемых «внешних» и «внутренних» сил [6, 7].

Постановка задачи

Входные данные задачи составляет растровое изображение глаза $\boldsymbol{M}$ размера $W \times H$ (данная работа проиллюстрирована результатами обработки изображений 640×480), каждый пиксель которого описывается одним байтом, то есть может соответствовать одной из двухсот пятидесяти шести градаций серого. Требуется определить границу зрачка на этом изображении в виде списка пикселей. Эта граница должна представлять собой непрерывный замкнутый контур, где непрерывность понимается в том смысле, что будучи соединёнными восьмисвязными рёбрами граничные пиксели образовывали бы связное множество. Два пикселя (x_1, y_1) и (x_2, y_2) соединены восьмисвязным ребром, если $\max(|x_1 - x_2|, |y_1 - y_2|) = 1$. Если построить такую границу невозможно в силу того, что зрачок перекрыт веком (как на рис. 1 (а)), то требуется определить ту её часть, которая представлена на изображении.

Описание алгоритма

Алгоритм исполняется как набор шагов.

— Изображение бинаризуется с некоторым порогом бинаризации. Предусмотрено использование нескольких порогов для оптимального отделения зрачка от фона.
— На полученном бинарном изображении строятся компоненты связности тёмных точек. Изображение обычно содержит несколько компонент связности, одна из которых включает зрачок. Граница или часть границы такой компоненты представляет собой линию, близкую к окружности.
— Компоненты связности и их границы поочерёдно рассматриваются, для границ применяется модификация преобразования Хафа с целью выделения центра гипотетической окружности. Аккумуляторы преобразований Хафа анализируются на наличие характерного максимума, возникающего при условии, что граница или часть границы области связности близки к окружности.
— При наличии такого максимума строится гистограмма расстояний от точки, в которой достигается максимум, до всех точек границы. Дополнительная проверка того, что компонента содержит зрачок, проводится на основе анализа этой гистограммы.

— В случае положительного ответа радиус окружности определяется по пику на гистограмме. Точками контура зрачка считаются точки, лежащие на расстоянии от центра, заметаемом пиком гистограммы.

Если зрачок не был распознан ни для одной компоненты, то устанавливается следующий порог бинаризации. Если не удалось получить решение ни при одном значении порога (из априорно заданных четырёх), то выдаётся решение, что зрачок не найден.

Бинаризация Применяется стандартная процедура бинаризации, состоящая в том, что устанавливается порог Θ, после чего всем пикселам изображения $\boldsymbol{M}$, значения которых меньше Θ, присваивается значение 0 (чёрный цвет), а оставшимся, где значение больше Θ, присваивается 255 (белый).

Выделение компонент связности Изображение $\boldsymbol{M}$ рассматривается как неориентированный граф, вершинами которого являются пиксели чёрного цвета, с четырёхсвязными рёбрами — (x_1, y_1) и (x_2, y_2) соединены ребром, если $\boldsymbol{M}[x_i, y_i] = 0$ и $|x_1 - x_2| + |y_1 - y_2| = 1$. Граф имеет несколько компонент связности, соответствующих отдельным тёмным областям на изображении. Для каждой их них находятся следующие характеристики:

— $\boldsymbol{representer}$ — некоторая точка, принадлежащая границе компоненты;
— *weight* — общее число точек, содержащихся в компоненте;
— $\boldsymbol{Boundary}$ — внешняя граница, заданная как список пикселей.

Точка компоненты является граничной, если она имеет белых соседей в $\boldsymbol{M}$, или лежит на границе изображения.

В списке $\boldsymbol{Boundary}$ для каждой точки помимо её координат хранится также ссылка на соседние граничные точки, таким образом, этот список представляет собой связное множество граничных точек.

Выделение компонент производится при помощи поиска в ширину. Как одна из главных целей преследуется построение списка $\boldsymbol{Boundary}$ указанной структуры для каждой компоненты. Создётся матрица $\boldsymbol{Visit}$ тех же размеров, что и $\boldsymbol{M}$, в которую будет записываться информация о степени изученности точек графа. Значение в $\boldsymbol{Visit}$ для каждого пикселя изображения может соответствовать одному из трёх возможных состояний: 0 — пиксель ещё не обрабатывался; 1 — пиксель обнаружен и является граничной точкой некоторой компоненты; 2 — пиксель обнаружен, при этом он не является граничной точкой или принадлежит внешней границе компоненты, обработка которой завершена. Все элементы $\boldsymbol{M}$ по очереди просматриваются, и если в некоторый момент найдётся не обнаруживавшийся ранее чёрный пиксель $(x, y) : \boldsymbol{M}[x, y] = 0, \boldsymbol{Visit}[x, y] = 0$, то это будет свидетельствовать об обнаружении новой компоненты связности; она обрабатывается, и просмотр $\boldsymbol{M}$ продолжается.

Обработка новой компоненты G происходит следующим образом: в качестве представителя $\boldsymbol{representer}$ берётся первая обнаруженная точка (x, y), которая, как важно заметить, также будет принадлежать внешней границе $\boldsymbol{Boundary}$, далее применяется поиск в ширину (известный также как «метод лесного пожара») с небольшими изменениями. Для его работы создётся очередь Q, изначально содержащая единственную точку $\boldsymbol{representer}$. Далее, на очередном шаге из Q извлекается элемент (x_0, y_0), находящийся в голове очереди, и в её конец помещаются все чёрные точки (x_i, y_i), соединённые с (x_0, y_0) четырёхсвязными рёбрами, не обнаруживавшиеся ранее ($\boldsymbol{Visit}[x_i, y_i] = 0$). Сама точка (x_0, y_0) помечается как просмотренная, значение *weight* при этом увеличивается на единицу. Если (x_0, y_0) является граничной, то по завершении просмотра соседей ей ставится в

соответствие значение 1 в $\boldsymbol{Visit}$, иначе — 2. В итоге обработки компоненты, то есть когда очередь Q пуста, всем граничным точкам G соответствует значение 1 в матрице **Visit**, а внутренним — значение 2.

После первого прохода, при котором вычисляется значение *weight* и размечается матрица $\boldsymbol{Visit}$, осуществляеся второй, при котором заполняется список $\boldsymbol{Boundary}$. Из точки $\boldsymbol{representer}$ поиск в ширину запускается во второй раз теперь по точкам, которым соответствует значение 1 в матрице $\boldsymbol{Visit}$, соединённым восьмисвязными ребрами. В качестве соседней для каждой точки (x, y) в список $\boldsymbol{Boundary}$ записывается та, из которой она была обнаружена в процессе нового поиска, эта запись происходит в момент помещения (x, y) в очередь Q. По окончании обработки (x, y) после её извлечения из очереди производится запись $\boldsymbol{Visit}[x, y] = 2$.

В оптимальном случае зрачок выделяется в одну компоненту связности, представляющую собой округлый объект, возможно, имеющий несколько пробелов, то есть пустых областей, окружённых точками компоненты. Следует отметить, что внутренние границы не попадают в список $\boldsymbol{Boundary}$ и не участвуют в дальнейшей обработке.

На большей части изображений выделяется несколько компонент связности. Они обрабатываются только если *weight* превосходит порог, что позволяет отсечь шумы, имеющие обычно малый размер области. В целях оптимизации компоненты обрабатываются в порядке убывания признака:

$$\Omega = \frac{\mathbf{weight} - |\mathbf{Boundary}|}{|\mathbf{Boundary}|^2}.$$

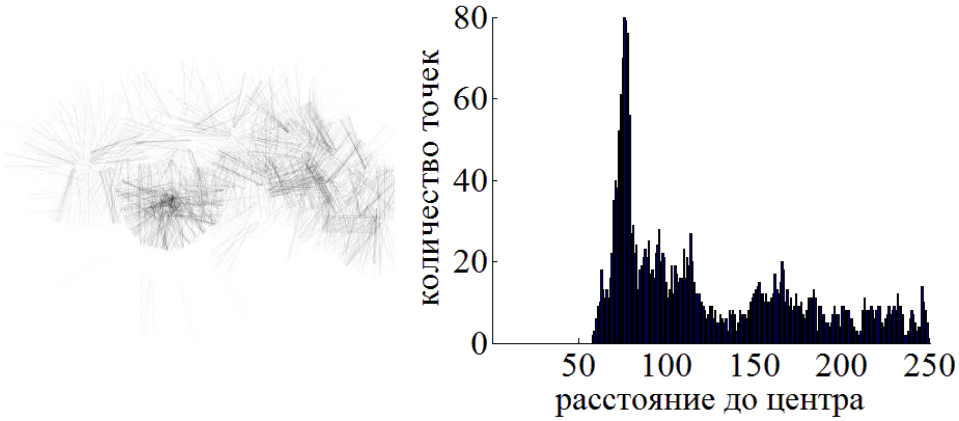
Такой признак обусловлен тем фактом, что среди всех фигур равной площади круг имеет минимальный периметр.

(а) изображение глаза

(б) бинаризованное изображение

Рис. 1: Бинаризация

Выделение центра компоненты связности Достаточно часто на изображении зрачок частично перекрыт веком, и при бинаризации он остаётся связан с ресницами (рис. 1). В этом случае и сама компонента, и форма её границы существенно отличаются от круглых. Для выявления округлой части компонент используется разновидность преобразования Хафа. Строится аккумуляторное пространство $\boldsymbol{A}$, которое представляет собой матрицу того же размера, что и $\boldsymbol{M}$. Изначально оно инициализировано нулевыми значениями. Проводится процедура голосования: строятся внутренние нормали в каждой точке из $\boldsymbol{Boundary}$, на луче выбирается отрезок, ограниченный некоторыми значениями



(а) преобразование Хафа (б) гистограмма расстояний

Рис. 2

R_{min} и R_{max}, после чего значения элементов аккумуляторного пространства в точках отрезка увеличиваются на единицу. Построение нормали происходит следующим образом: в окрестности граничной точки $\boldsymbol{r_0}$ выбирается элемент границы, состоящий из пятнадцати точек, и аппроксимируется касательное направление $\boldsymbol{\tau}$:

$$\boldsymbol{\tau} = \sum_{i=-7}^{-1} \frac{\boldsymbol{r_0} - \boldsymbol{r_i}}{|\boldsymbol{r_0} - \boldsymbol{r_i}|} + \sum_{i=1}^{7} \frac{\boldsymbol{r_i} - \boldsymbol{r_0}}{|\boldsymbol{r_i} - \boldsymbol{r_0}|}$$

Из двух возможных нормальных направлений выбирается то, которое направлено в сторону чёрной точки, с которой соприкасается $\boldsymbol{r_0}$. Если возникает неоднозначность выбора направления, то берется любое — важно найти правильное направление только для точек, принадлежащих фактической границе зрачка, а каждая из них лежит на стыке чёрной и белой областей, поэтому проблема выбора для них не возникнет. Отрезок голосования строится на растре путем применения алгоритма Брезенхема. По окончании процедуры голосования значения в аккумуляторном пространстве сглаживаются фильтром низкой частоты.

На рис. 2 (а) представлен результат описанного преобразования Хафа для наибольшей из компонент связности изображения рис. 1 (б), которая содержит зрачок. Это единственная обработанная компонента, остальные были отбракованы по порогу *weight*.

Оценка центра компоненты связности На рис. 2 (а) виден резкий максимум вблизи фактического центра зрачка. Наличие такого максимума является также признаком принадлежности зрачка компоненте. Следует оценить достоверность такого предположения, для того, чтобы из всех рассматриваемых компонент связности выбрать наиболее соответствующую зрачку. Если зрачок содержится в компоненте связности, то, как правило, сигнальные отрезки, проведенные из точек его границы, дают глобальный максимум, который существенно превышает остальные значения. Значения максимумов аккумулятора для компонент связности, содержащих зрачок и не содержащих его (шумовых), были вычислены для базы данных [8]. В таблице 1 приведены доли компонент, содержащих зрачок, и шумовых, имеющих максимум аккумулятора в заданных интервалах.

Видна хорошая разделимость истинных и шумовых компонент по значению максимума. Таким образом, считается, что компонента может содержать зрачок, если в аккумуляторном пространстве преобразования Хафа этой компоненты содержится точка, на-

Таблица 1

Диапазон значений максимума	< 30	30 —40	40 —50	50 —70	> 70
Доля компонент, содержащих зрачок			5%	25 %	70 %
Доля шумовых компонент	80 %	10 %	10 %		

бравшая хотя бы 50 голосов. Та небольшая часть компонент, содержащих зрачок, которая отсеивается этим порогом, на самом деле не создаёт ошибки ложного пропуска, поскольку при другом пороге бинаризации эта компонента изменяет форму, становясь более круглой, порождает максимум выше порога и корректно определяется как зрачок.

Построение гистограмм радиусов Строится гистограмма $H(r)$ расстояний от найденного центра C области до всех точек $\boldsymbol{Boundary}$:

$$H(r) = |(x,y) : (x,y) \in \boldsymbol{Boundary},\ r - 0.5 < \rho\,(C,(x,y)) < r + 0.5|$$

Гистограмма в случае принадлежности зрачка компоненте имеет специфический вид. Например, на рис. 2 (б), дана гистограмма для компоненты рис. 1 (б). Виден резкий пик, соответствующий радиусу зрачка, присутствующего на изображении: $R = arg\max\limits_{r} H(r)$. Пик может и не быть настолько резким, если зрачок имеет овальную форму.

Окончательный вывод относительно того, содержится ли зрачок в компоненте, делается на основе анализа $H(r)$. Вычисляется масса гистограммы в окне шириной 10 с центром в точке R, где наблюдается максимальное среди всех значение гистограммы:

$$L = \sum_{r=R-5}^{R+5} H(r)$$

Полученное значение представляет собой оценку длины контура зрачка и сравнивается с R, домноженным на некоторый коэффициент α. Ширина окна и α в работе подобраны субъективно, требуется их уточнение в более подробном исследовании.

Результаты обработки изображений

Для тестирования были использованы изображения открытой базы [8]. Эти изображения были размечены экспертами — на каждом была построена окружность радиуса r_0 с центром в точке (x_0, y_0), которой приближалась граница зрачка. Функционал качества задавался как

$$Q_1 = |x_0 - x| + |y_0 - y| + |r - r_0|,$$

где величины без индексов представляют собой координаты центра, полученного в результате преобразования Хафа, и радиус, соответствующий пику на гистограмме.

На диаграмме рис. 3 представлено процентное распределение по величинам отклонения, в интервал на графике попали $98,6$ % точек. На диаграммах рис. 4 отдельно показаны отклонения центров и радиусов, определяемые функционалами

$$Q_2 = \sqrt{(x - x_0)^2 + (y - y_0)^2},$$

$$Q_3 = |r - r_0|\,.$$

Оценка времени обработки изображений на компьютере с ОЗУ 2 Гб и процессором частотой $2,8$ ГГц имеет следующий вид: 56 % изображений обрабатывались не дольше одной

секунды, и 80 % обрабатывались не дольше двух секунд. На некоторых изображениях время обработки могло доходить до 15-ти секунд. Такое замедление связано с тем, что первые несколько порогов бинаризации не позволяли определить зрачок как компоненту связности.

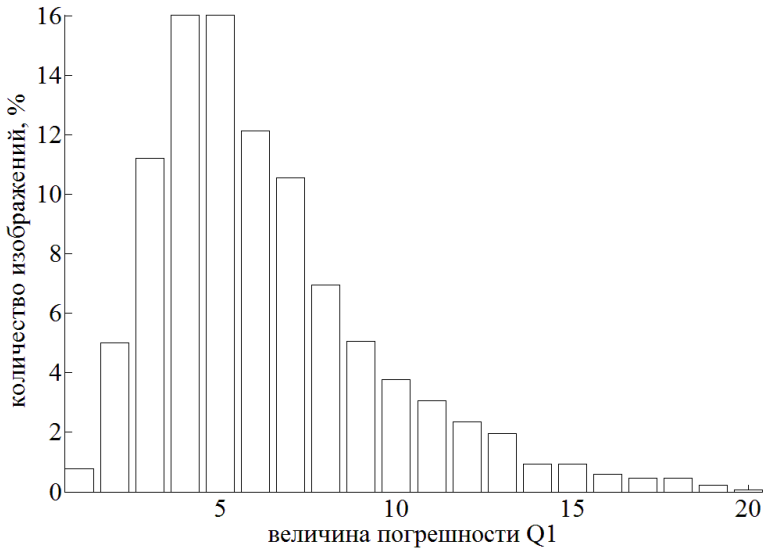
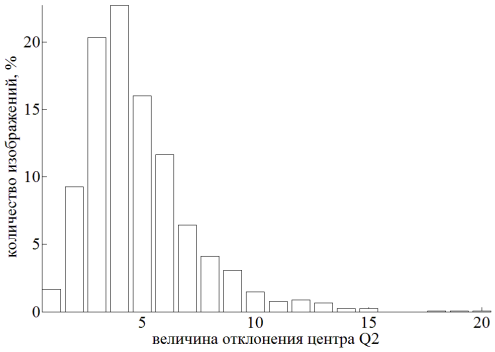
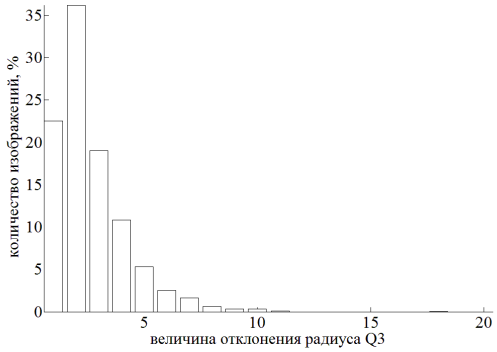


Рис. 3: диаграмма отклонений



(а) отклонения центров



(б) отклонения радиусов

Рис. 4

Заключение

Предложен метод выделения контура зрачка на изображении глаза. Контур строится из точек, лежащих на границе области зрачка, полученной при бинаризации изображения как компонента связности множества тёмных точек. Для выбора правильной (наиболее близкой к окружности) компоненты (или её части) используется преобразование Хафа для граничных точек, приводящее к выделению центра её округлой части. Радиус окружности определяется как максимум гистограммы расстояний граничных точек до выделенного центра. Сами точки контура зрачка затем выбираются из всего множества точек границы

как имеющие близкое к этому радиусу расстояние до центра. Проведена апробация работы алгоритма на наборе изображений из открытой базы данных BATH.

Использование преобразования Хафа позволяет корректно выделять истинные центр, радиус и контур зрачка даже при наличии больших помех, при условии лишь частичной видимости контура зрачка на изображении. Недостатком метода является потенциально большое время работы, возникающее при переборе многих компонент связности и нескольких порогов бинаризации. Этот недостаток частично скомпенсирован введением признака качества компоненты связности.

Литература

[1] *Daugman J.* How iris recognition works // *Proc. IEEE Trans. Circ. Syst. Video Technol.* 2004. Vol. 14. Pp. 21–30.

[2] *Матвеев И. А.* Поиск центра радужки на изображении методом Хафа с двумерным пространством параметров // *Известия РАН. Теория и системы управления.* 2012. Vol. 5. Pp. 44–51.

[3] *Li Ma, Tan T., Wang Y.* Iris recognition using circular symmetric filters // *Proceedings of the 16th International Conference on Pattern Recognition.* 2002. Pp. 414–417.

[4] *Wildes R. P.* Iris recognition an emerging biometric technology // *Proceedings of the IEEE.* 1995. Vol. 85.

[5] *Kennell L., Ives R. W., Gaunt R. M.* Binary morphology and local statistics applied to itis segmentation for recognition // *Proc. of the 2006 IEEE International Conference on Image Processing, Atlanta, GA.* 2006. Pp. 293–296.

[6] *Ritter M.* Location of the pupil-iris border in slit-lamp images of the cornea // *Proceedings of the International Conference on Image Analysis and Processing.* 1999.

[7] *Ritter M., Cooper J. R.* Locating the iris: A first step to registration and identification // *Proc. 9th IASTED International Conference on Signal and Image Processing* . 2003. Pp. 507–512.

[8] *Monro D. M., Rakshit S., Zhang D.* Universtiy of Bath, U.K. Iris Image Database.

Применение методов распознавания образов для синтеза кусочно-линейных систем квазиинвариантного управления*

Теклина Л. Г., Котельников И. В., Гельфер И. С.
neymark@pmk.unn.ru
Нижегородский государственный университет им. Н. И. Лобачевского

Работа посвящена дальнейшему развитию нового подхода к синтезу систем квазиинвариантного управления, основанному на постановке и решении задачи синтеза методами распознавания образов с активным экспериментом. Расширение новой методики синтеза линейных систем на область нелинейных систем управления связано с преодолением главного недостатка линейных систем: больших значений функции управления в переходном процессе.

Application of the pattern recognition methods for the synthesis of a piecewise-linear systems of the quasi-invariant control*

Teklina L. G., Kotel'nikov I. V., Gel'fer I. S.
Lobachevsky State University of Nizhni Novgorod

The work is devoted to the further development of a new approach to the synthesis of the quasi-invariant control systems. This approach is based on the formulation and solution of the synthesis problem using methods of pattern recognition with an active experiment. Extension of a new method of synthesis for the linear systems to the field of nonlinear control systems is related with overcoming the major drawback of linear systems: large values of a control function in the transition process.

Введение

Подавление внешних возмущений традиционно считается трудной задачей в теории управления. Проблеме синтеза систем управления, которые должны не устранять возникающие ошибки, а предотвращать их, т.е. сделать объект управления невосприимчивым (инвариантным) к внешним воздействиям, много лет. Идея синтеза таких систем прошла сложный путь: от полного отрицания возможности их создания (дискуссия по поводу инвариантного регулятора Г.В. Щипанова в 1939 г.) до попыток отыскать зерно истины (введение понятия $\varepsilon-$ инвариантности) и найти методы расчета таких систем. Об истории этого вопроса можно прочитать в книге [1]. Поиск шел в привычном направлении: сведение к задаче оптимизации путем введения различных мер неинвариантности, или реактивности, системы [2]. Лишь в конце XX, начале нашего XXI века появились циклы работ Якубовича В.А. [3,4] и Неймарка Ю.И. [5-7], посвященных инвариантному управлению, где было доказано, что при отсутствии знаний о помехе идеальный инвариантный регулятор действительно не реализуем, но реализуемы линейные системы *квазиинвариантного* управления. Показано, что при надлежащем выборе управляющей функции существует множество значений параметров синтезируемого квазиинвариантного управления, при которых управляемый объект удовлетворяет требованию квазиинвариантности. Но вот

Работа выполнена при финансовой поддержке РФФИ, проект №11-01-00379.

методов поиска этих значений не существует. Квазиинвариантное управление не является оптимальным, и для его расчета не применимы классические методы оптимального управления. Для решения этой проблемы предлагается использовать методы интеллектуального анализа данных, получаемых в процессе решения поставленной задачи как задачи распознавания с активным экспериментом.

Задача синтеза систем квазиинвариантного управления

В работах [5-7] доказана возможность синтеза и изучены некоторые свойства линейных систем квазиинвариантного управления (и стабилизации, и слежения) для произвольного объекта, описываемого математической моделью вида

$$\begin{cases} A_n(p)x = -B_m(p)(u + \xi(t)) \\ \mu u = D_{n-m-1}(p)(x - f(t)) \end{cases} \tag{1}$$

где $p = \frac{d}{dt}$, $A_n(p)$, $B_m(p)$, $D_{n-m-1}(p)$ - действительные полиномы степеней n, m, $n-m-1$ соответственно, $\xi(t)$ - ограниченное неизвестное возмущение, $f(t)$ - заданная функция отслеживания ($f(t) = 0$ отвечает системе стабилизации), x и u - одномерные переменные объекта и управления. Для этой модели установлены условия реализации квазиинвариантного управления, а именно:

- $n > m$ для кусочно-гладких ограниченных $u(t)$, и $\xi(t)$;
- $x(t)$ и $f(t)$ - кусочно-гладкие функции, имеющие $n - m$ производных;
- полином $\mu A_n(p) + B_m(p)D_{n-m-1}(p)$ Гурвицев, т. е. все его корни λ_i имеют отрицательные реальные части ($Re\lambda_i < 0$).

Показано, что при выполнении этих условий существует множество значений параметров - неизвестных коэффициентов $(\mu, d_1, \ldots, d_{n-m-1}) = (\mu, \mathbf{d})$ управляющей функции во втором уравнении системы (1), при которых система удовлетворяет требованию квазиинвариантности управления, когда надлежащим выбором параметров ошибка управления в установившемся режиме может быть сделана сколь угодно малой, т. е. $|x(t) - f(t)| < \varepsilon$ при $t > T_{tp}$, где T_{tp} - длительность переходного процесса.

Кроме того синтезируемая система управления должна обеспечивать и требуемое качество переходного процесса, а именно:

- ограничение на длительность переходного процесса $T_{tp} \leqslant T_{max}$;
- ограничение на величину ошибки управления в переходном процессе $|x(t) - f(t)| \leqslant x_{max}$;
- ограниченность значений функции управления $|u(t)| \leqslant u_{max}$.

Синтез системы управления заключается в поиске таких значений неизвестных параметров μ, d, при которых синтезируемая система отвечала бы всем предъявляемым к ней требованиям по квазиинвариантности управления и качеству переходного процесса для неизвестного, но ограниченного по величине внешнего возмущения $\xi(t)$. Причем ставится задача отыскания и описания не всех возможных значений неизвестных параметров, а хотя бы некоторого их подмножества достаточно простой конфигурации (совокупность параллелепипедов, сфер, эллипсоидов и т.п.), но определяемого с заданной высокой степенью статистической достоверности и отвечающего условию достаточности меры робастной устойчивости по определяемым параметрам. Рассмотрим возможности синтеза такой системы методами распознавания при условии, что описание объекта управления (первое уравнение системы (1)) известно и заданы начальные условия $\mathbf{x}^* = (x^*, \dot{x}^*, \ldots, x^{(n-1)*})$ в фазовом пространстве системы.

Постановка задачи синтеза в качестве проблемы распознавания образов

Рассматривается задача управления объектом, описываемым известным линейным дифференциальным уравнением n - ого порядка, с помощью управления, задаваемого также дифференциальным уравнением, но с неизвестными $k = n - m$ параметрами

$$\boldsymbol{\omega} = (\omega_1, \ldots, \omega_k) = (\mu, d_1, \ldots, d_{n-m-1}).$$

Следовательно, синтезируемый объект описывается двумя группами переменных: фазовыми переменными $\mathbf{x} = (x, \dot{x}, \ldots, x^{(n-1)})$ в n - мерном фазовом пространстве $\mathbf{X}$ и параметрами $\boldsymbol{\omega}$ в k - мерном пространстве параметров $\boldsymbol{\Omega}$. В фазовом пространстве известно состояние равновесия $\mathbf{x_0}$ и заданы начальные условия $\mathbf{x}^*$. Дополнительно задаются определенные требования к функционированию синтезируемой системы в виде l неравенств, характеризующих некоторое множество $\mathbf{Y}^*$ в l - мерном пространстве «характеристик» системы управления $\mathbf{Y}$. Классическими характеристиками y_i систем управления являются требования, перечисленные в предыдущем разделе:

- $y_1 = |x(t) - f(t)| < \varepsilon$ при $t > T_{tp}$;
- $y_2 = T_{tp} \leqslant T_{max}$;
- $y_3 = |u(t)| \leqslant u_{max}$;
- $y_4 = |x(t) - f(t)| \leqslant x_{max}$ при $t \leqslant T_{tp}$.

При задании вектора $\boldsymbol{\omega}$ и начальных условий $\mathbf{x}^*$ в фазовом пространстве при интегрировании системы (1) получаем некоторое решение $x(t)$, которое имеет характеристики $\mathbf{y} = \Gamma(\mathbf{x}^*, \boldsymbol{\omega})$, представляющие собой численные оценки $\mathbf{y} = (y_1, \ldots, y_4)$ качества переходного процесса и установившегося режима. Синтезировать систему - значит найти такое множество значений $\boldsymbol{\omega} \in \boldsymbol{\Omega}^* \subset \boldsymbol{\Omega}$, для которых $\mathbf{y} = \Gamma(\mathbf{x}^*, \boldsymbol{\omega}) \in \mathbf{Y}^*$.

Для постановки задачи в виде проблемы распознавания образов в качестве пространства признаков выбираем пространство неизвестных параметров $\boldsymbol{\Omega}$. За распознаваемый образ принимается область $\boldsymbol{\Omega}^*$ в пространстве признаков. Решение задачи распознавания в пространстве $\boldsymbol{\Omega}$ - построение в этом пространстве локального решающего правила достаточно простого вида (набор параллелепипедов, сфер, эллипсоидов и т. п.), описывающих искомую область параметров $\widetilde{\boldsymbol{\Omega}^*} \subseteq \boldsymbol{\Omega}^*$, при условии минимизации числа ошибок второго рода для распознаваемого образа и максимизации меры робастной устойчивости для множества параметров $\widetilde{\boldsymbol{\Omega}^*}$. С целью ускорения процесса решения ищется решение не оптимальное, а близкое к оптимальному, удовлетворяющее требуемой надежности распознавания P_0 и необходимой мере робастной устойчивости по определямым параметрам $R(\widetilde{\boldsymbol{\Omega}^*}) \geqslant R_0$. За оценку достоверности правила распознавания принимается величина $P = 1 - \frac{N_0}{N} (P > P_0)$, где N_0 — число ошибочных ответов на контрольной выборке объема , а за меру робастной устойчивости для области $\boldsymbol{\Omega}^*$ — радиус робастной устойчивости $R(\widetilde{\boldsymbol{\Omega}^*}) = \max\limits_{\boldsymbol{\omega} \in \widetilde{\boldsymbol{\Omega}^*}} (\min\limits_{\boldsymbol{\omega} \in \widetilde{\boldsymbol{\Omega}}} \rho(\boldsymbol{\omega}, \widetilde{\widetilde{\boldsymbol{\Omega}^*}}))$, где ρ — эвклидово расстояние от точки внутри области до ее границы $\widetilde{\widetilde{\boldsymbol{\Omega}^*}}$.

Для построения решающего правила формируется обучающая выборка с использованием решающего правила в пространстве $\mathbf{Y}$:

$$\boldsymbol{\omega} \in \boldsymbol{\Omega}^* \quad \text{если} \quad \Gamma(\mathbf{x}^*, \boldsymbol{\omega}) \in \mathbf{Y}^*,$$

которая выполняет роль учителя. Все обучающие выборки формируются в процессе решения задачи, т. е. решается задача распознавания с активным экспериментом. Для формирования таких выборок решается задача планирования и проведения эксперимента (выбор

параметров $\boldsymbol{\omega}$ - построение $x(t)$ - отыскание характеристик $\mathbf{y}$) с целью получения представительных обучающих выборок. В задаче планирования эксперимента удобнее использовать решающую функцию, построенную на базе решающего правила в пространстве $\mathbf{Y}$:

$$F(\boldsymbol{\omega}) = \rho(\Gamma(\mathbf{x}^*, \boldsymbol{\omega}), \mathbf{Y}^*),$$

где ρ- функция, представляющая собой расстояние в пространстве $\mathbf{Y}$ от точки $\mathbf{y} = \boldsymbol{\Gamma}(\mathbf{x}^*, \boldsymbol{\omega})$ (технические характеристики объекта) до множества $\mathbf{Y}^*$. Функция $F(\boldsymbol{\omega})$ не только указывает на принадлежность точки $\boldsymbol{\omega}$ к распознаваемому образу ($F(\boldsymbol{\omega}) = 0$ для $\boldsymbol{\omega} \in \boldsymbol{\Omega}^*$), но и может служить мерой близости $\boldsymbol{\omega}$ к $\boldsymbol{\Omega}^*$ при $F(\boldsymbol{\omega}) \neq 0$.

Синтез линейной системы управления

Решение задачи синтеза реализуется в два этапа путем решения двух задач распознавания:

I. Отыскание области значений в пространстве параметров $\boldsymbol{\Omega_0} \subset \boldsymbol{\Omega}$, при которых синтезируемая система устойчива, из условия, что полином $\mu A_n(p) + B_m(P)D_{n-m-1}(p)$ Гурвицев, т. е. все его корни имеют отрицательные действительные части. Множество $\boldsymbol{\Omega_0}$ - распознаваемый образ. В общем случае область устойчивости – область невыпуклая и несвязная, но мы ставим задачу выделения и описания хотя бы части $\widetilde{\boldsymbol{\Omega_0}} \subseteq \boldsymbol{\Omega_0}$ этой области, полагая ее связной и выпуклой.

II. Построение области параметров $\widetilde{\boldsymbol{\Omega}^*} \subseteq \widetilde{\boldsymbol{\Omega_0}}$, удовлетворяющей целевому условию управления. На этом этапе распознаваемым образом является искомое множество $\boldsymbol{\Omega}^*$.

Каждый этап реализуется последовательным выполнением трех процедур:

1) поиск точки с координатами значений параметров, удовлетворяющих целевому условию этапа;

2) формирование обучающей выборки на основе выбранной точки и исходя из гипотезы компактности искомого множества;

3) построение решающего правила, отвечающего заданной степени статистической достоверности.

На первом этапе процедура (1) реализуется путем решения задачи минимизации $\min\limits_{\boldsymbol{\omega}} \Lambda(\boldsymbol{\omega})$, где $\Lambda(\boldsymbol{\omega}) = \max\limits_{i}(Re\lambda_i)$, λ_i - корни характеристического полинома, причем процесс минимизации заканчивается, как только выполнится неравенство $\Lambda(\boldsymbol{\omega}) < 0$.

На втором этапе процедура (1) реализуется путем решения задачи минимизации

$$\min_{\boldsymbol{\omega} \in \widetilde{\boldsymbol{\Omega_0}}} F(\boldsymbol{\omega}) = \min_{\boldsymbol{\omega} \in \widetilde{\boldsymbol{\Omega_0}}} \rho(\Gamma(\mathbf{x}^*, \boldsymbol{\omega}), \mathbf{Y}^*).$$

Процедура (2) формирует обучающую последовательность путем случайного выбора параметров на основе равномерного распределения из некоторой области G, представляющей собой окрестность найденной точки. Требования к этой области состоят в том, что она должна включать как точки со значениями параметров, удовлетворяющих цели этапа, так и точки, не удовлетворяющие поставленной цели. Это сделать не сложно простым расширением области G.

По найденным обучающим выборкам строятся решающие правила распознавания (процедура 3). При этом проводится оценка надежности построенного решающего правила на независимой контрольной выборке. В случае несоответствия результата проверки заданной степени статистической достоверности P_0 обучающая выборка пополняется новыми данными, и процесс обучения продолжается. Мы в своих исследованиях строили синдромальные решающие правила [8], выбирая из полученных синдромов (параллелепипедов)

для описания распознаваемого образа синдромы максимального объёма с целью максимизации меры робастной устойчивости.

Задача синтеза решена, если построенное множество $\widetilde{\mathbf{\Omega_0}}$ удовлетворяет заданной степени статистической достоверности, а мера робастной устойчивости достаточна для синтезируемой системы. В противном случае существуют два пути модификации управляющей функции: изменение пространства признаков и использование нелинейных стратегий управления. Второй способ особенно актуален при превышении допустимых значений функции управления. И теоретические [6,7], и экспериментальные исследования показывают, что для линейных систем управления самой большой неприятностью являются большие значения функции управления в переходном процессе, причем эти значения тем больше, чем выше требования к точности управления. Зачастую преодолеть этот недостаток линейных систем можно путем выбора простейшей нелинейной стратегии управления, а именно: через построение кусочно-линейной управляющей функции.

Алгоритм построения кусочно-линейной управляющей функции

Цель введения кусочно-линейного управления – выполнение требования ограниченности значений функции управления $|u(t)| \leqslant u_{max}$ во время переходного процесса как при запуске системы, так и при сбое в системе управления. Алгоритм построения кусочно-линейной функции базируется на знании области устойчивости системы управления $\widetilde{\mathbf{\Omega_0}}$, найденной для линейной системы. И теоретически [5-7], и экспериментально установлена большая роль параметра μ при синтезе квазиинвариантного управления с заданными свойствами. Именно поэтому при построении кусочно-линейной управляющей функции в первую очередь изменяется параметр μ. Для того, чтобы области изменения параметров μ ($\mu \in \mathbf{M}$) и $\mathbf{d}$ ($\mathbf{d} \in \mathbf{H}$) не зависели друг от друга, область $\widetilde{\mathbf{\Omega_0}}$ должна представлять собой прямую сумму двух множеств $\mathbf{M_0} = \{\mu\}$ и $\mathbf{H_0} = \{\mathbf{d}\}$, т. е. $\widetilde{\mathbf{\Omega_0}} = \mathbf{M_0} \oplus \mathbf{H_0}$, или

$$\widetilde{\mathbf{\Omega_0}} = \{\boldsymbol{\omega} = (\mu, \mathbf{d}) \quad / \quad \mu \in \mathbf{M_0} \quad \& \quad \mathbf{d} \in \mathbf{H_0}\}.$$

Простейшим типом такой области является параллелепипед.

Из области $\widetilde{\mathbf{\Omega_0}}$ выделяются множества $\mathbf{M_0} = \{\mu\}$ и $\mathbf{H_0} = \{\mathbf{d}\}$. В $\mathbf{M_0}$ выбираем некоторые значения, близкие к минимальному и максимальному по модулю (μ_{min} и μ_{max} соответственно), и разбиваем весь интервал изменения $|\mu|$ на k значений:

$$|\mu_i| = \mu_{min} + \Delta\mu(i-1), \quad i = 1, \ldots, k, \quad \Delta\mu = \frac{\mu_{max} - \mu_{min}}{k-1}.$$

Начиная с $\mathbf{x} = \mathbf{x}^*$ (начальные условия) для произвольного значения $\mathbf{d}$ из области устойчивости $\mathbf{H_0}$ при интегрировании системы на каждом шаге в качестве μ выбираем минимальное (для достижения минимальной ошибки управления) из ряда полученных значений:

$$|\mu| = \begin{cases} \mu_{min} & \text{если} \quad \hat{D}(t) \leqslant \mu_1 u_{max} \\ \mu_i & \text{если} \quad \mu_{i-1} u_{max} < \hat{D}(t) \leqslant \mu_i u_{max} \\ \text{изменение } \mathbf{d} & \text{если} \quad \mu_k u_{max} < \hat{D}(t) \end{cases},$$

где $\hat{D}(t) = |D(p)(x(t) - f(t))|$. Параметры $\mathbf{d}$ изменяются (временно!) лишь в случае невозможности достижения требуемой величины $u(t)$ только за счет изменения μ.

Изменение вектора $\mathbf{d}$ проводится путем решения задачи

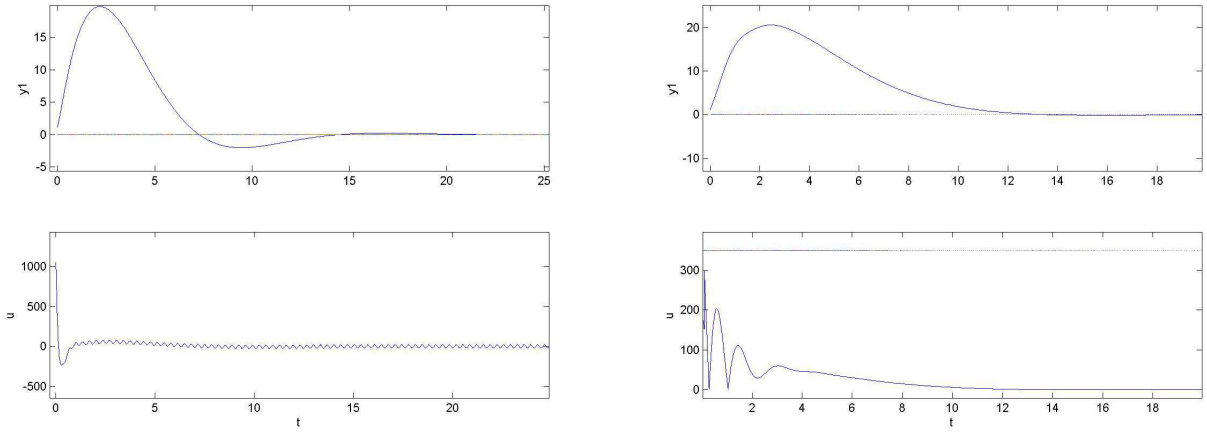
$$\min_{\mathbf{d} \in \mathbf{H_0}} |D(p)(x(t) - f(t))|,$$

причем процесс минимизации можно закончить, как только $|D(p)(x(t)-f(t))| \leqslant \mu_{max} u_{max}$, а можно продолжить с целью выбора меньших значений μ. При возвращении к нормальному режиму работы системы управления восстанавливается исходное значение вектора $\mathbf{d}$, которое можно выбрать не произвольным образом, а исходя, например, из меры робастной устойчивости по этим параметрам.

Если требуемого сокращения функции управления достичь не удается, необходимо рассматривать иные нелинейные стратегии управления.

Как выбрать число k кусочно-линейных интервалов? Эксперименты показали, что с ростом k в среднем снижается, но незначительно, длительность переходного периода, но возрастает число переключений. Мы в своих исследованиях начинали с $k = 2$ и увеличивали число интервалов лишь в случае, если не достигали требуемых результатов.

Для иллюстрации возможностей кусочно-линейных систем управления в сравнении с линейными приведем пример динамической системы, описывающей двухзвенный перевернутый маятник, управляемый перемещением точки опоры в условиях наличия неизвестного внешнего возмущения[9]. На рис.1 представлены характеристики переходного процесса: ошибка управления и управляющая функция $u(t)$.



(а) линейная $u(\mathbf{x})$ (б) кусочно-линейная $u(\mathbf{x})$

Рис. 1: Переходный процесс для линейной и кусочно-линейной управляющей функции

Для линейной системы управления (рис.1а) функция управления в переходном процессе достигает значений $|u(t)| > 1000$ при ограничении $u_{max} \leqslant 300$. Требуемого ограничения удается добиться при переходе к кусочно-линейной управляющей функции, для которой общий вид переходного процесса представлен на рис.1б.

Уточнение области изменения параметров для синтеза системы управления с заданными свойствами

В выбранной области $\widetilde{\mathbf{\Omega_0}}$ существует линейное квазиинвариантное управление, но существование нелинейного (и в частности, кусочно-линейного) требует дополнительной проверки. Кроме того, есть еще и дополнительные требования, предъявляемые к переходному процессу. Синтез системы управления заключается в поиске таких значений неизвестных параметров $\boldsymbol{\omega}^* \in \mathbf{\Omega}^* \subseteq \mathbf{\Omega_0}$, при которых синтезируемая система отвечала бы

всем предъявляемым к ней требованиям для неизвестного, но ограниченного по величине внешнего возмущения $\xi(t)$.

Поскольку при синтезе кусочно-линейных систем квазиинвариантного управления значение параметра μ определяется текущими значениями $\mathbf{d}$ и $\mathbf{x}$, а область изменения μ - это область устойчивости синтезируемой системы управления, то $\mathbf{\Omega}^* = \mathbf{M_0} \oplus \mathbf{H}^*$, где $\mathbf{H}^* \subseteq \mathbf{H_0}$.

При задании вектора $\mathbf{d}$ и начальных условий $\mathbf{x}^*$ в фазовом пространстве при интегрировании системы получаем некоторое решение $x(t)$, которое имеет характеристики $\mathbf{y} = \Gamma(\mathbf{x}^*, \mathbf{d})$. Сказанное выше означает, что

$$\mathbf{H}^* = \{\mathbf{d} \quad / \quad \mathbf{y} = \Gamma(\mathbf{x}^*, \mathbf{d}) \in \mathbf{Y}^*\}.$$

Причем опять же ставится задача отыскания и описания не всех возможных значений неизвестных параметров, а хотя бы некоторого их подмножества $\widetilde{\mathbf{H}^*} \subseteq \mathbf{H}^*$ достаточно простой конфигурации, но определяемого с заданной высокой степенью статистической достоверности. Теперь за распознаваемый образ принимается область $\mathbf{H}^*$ в подпространстве признаков $\mathbf{H}$. Решение задачи распознавания в подпространстве $\mathbf{H}$ - построение в этом подпространстве решающей функции, которое ведется аналогично построению решающего правила для $\widetilde{\mathbf{\Omega}^*}$ в пространстве $\mathbf{\Omega}$, но с использованием следующей решающей функции в пространстве $\mathbf{Y}$:

$$\mathbf{d} \in \mathbf{H}^* \quad \text{если} \quad \Gamma(\mathbf{x}^*, \mathbf{d}) \in \mathbf{Y}^*,$$

при этом $F(\mathbf{d}) = \rho(\Gamma(\mathbf{x}^*, \mathbf{d}), \mathbf{Y}^*)$, где ρ- функция, представляющая собой расстояние в пространстве $\mathbf{Y}$ от точки $\mathbf{y} = \mathbf{\Gamma}(\mathbf{x}^*, \mathbf{d})$ до множества $\mathbf{Y}^*$.

Аналогично поиску области $\widetilde{\mathbf{\Omega}^*}$ построение области $\widetilde{\mathbf{H}^*} \subseteq \mathbf{H}^* \subseteq \mathbf{H_0}$ методами распознавания образов складывается из последовательного решения трех задач:

1. Отыскание хотя бы одной точки $\mathbf{d}^*$, принадлежащей распознаваемому образу. $\mathbf{d}^* \in \mathbf{H}^*$ находится из условия

$$F(\mathbf{d}^*) = \min_{\mathbf{d} \in \mathbf{H_0}} F(\mathbf{d}) = \min_{\mathbf{d} \in \mathbf{H_0}} \rho(\Gamma(\mathbf{x}^*, \mathbf{d}), \mathbf{Y}^*).$$

2. Формирование обучающей выборки $\Xi_{\mathbf{H}^*}$ на основе гипотезы компактности и выбранной точки $\mathbf{d}^*$. Если при построении $x(t)$ $\mathbf{d}$ изменяется, то все полученные значения $\mathbf{d}$ включаются в $\Xi_{\mathbf{H}^*}$.

3. Построение решающего правила распознавания по сформированной обучающей выборке $\Xi_{\mathbf{H}^*}$ с оценкой его надежности на независимой контрольной выборке.

В случае, если $\widetilde{\mathbf{H}^*}$ найдено и удовлетворяет заданной степени статистической достоверности, а мера робастной устойчивости достаточна для синтезируемой системы, задача синтеза решена, и $\widetilde{\mathbf{\Omega}^*} = \mathbf{M_0} \oplus \widetilde{\mathbf{H}^*}$.

При решении упомянутой выше задачи управления двухзвенным перевернутым маятником при требованиях $|x(t)| < 10^{-3}$ при $t > T_{tp}$, $|u(t)| \leqslant 300$ и $P_0 = 0.99$ уже при $k = 2$ была получена достоверность $P = 0.9911$ на всей области $\mathbf{H_0}$, т. е. $\widetilde{\mathbf{\Omega}^*} = \widetilde{\mathbf{\Omega}_0}$, но, подчеркнем, лишь для выдвинутых для этой системы управления требований.

Заключение

Предлагаемая работа имеет целью показать, что проблема синтеза систем управления может плодотворно рассматриваться как задача распознавания образов и что на этом пути возможно существенное продвижение в ее решении. Синтез кусочно-линейных систем

квазиинвариантного управления – это лишь первый шаг на пути решения очень сложной проблемы: синтеза нелинейных систем управления с заданными свойствами. И полученные результаты позволяют надеяться на ее успешное решение методами интеллектуального анализа данных.

Литература

[1] *Составители Лезина З. М., Лезин В. И.* Щипанов Г.В. и теория инвариантности. — М.: Физматлит, 2004.

[2] *Якубович В. А.* Оптимизация и инвариантность линейных стационарных систем управления. // Автоматика и телемеханика. — 1982. — № 8. — С. 5–45.

[3] *Якубович В. А.* Синтез стабилизирующих регуляторов, обеспечивающих независимость выходной переменной системы управления от внешнего воздействия. // ДАН. — 2001. — Т. 380,№ 1. — С. 27–30.

[4] *Проскурников А. В., Якубович В. А.* Синтез стабилизирующих регуляторов, обеспечивающих независимость выходной переменной системы управления от внешнего воздействия. // ДАН. — 2003. — Т. 389,№ 6. — С. 742–746.

[5] *Неймарк Ю. И.* О парадоксе и идеальном регуляторе Щипанова // Вестник ННГУ им. Н. И. Лобачевского. Математическое моделирование и оптимальное управление. — 2006. — Вып. 3(32). — С. 83–88.

[6] *Неймарк Ю. И.* О квазиинвариантном управлении. // Дифференциальные уравнения.— 2007. № 11. — С. 1–6.

[7] *Неймарк Ю. И.* Синтез и функциональные возможности квазиинвариантного управления // Автоматика и телемеханика. — 2008. № 10. — С. 48–56.

[8] *I. V. Kotel'nikov.* A Syndrome Recognition Method Based on Optimal Irreducible Fuzzy Tests. // Pattern Recognition and Image Analysis. — 2001. — Vol. 11, No. 3. — Pp. 553–559.

[9] *Неймарк Ю. И.* Математические модели в естествознании и технике. — Нижний Новгород: Издательство Нижегородского университета, 2004. — 401 с.

О метрической коррекции матриц парных сравнений*

***Двоенко С. Д.*[1], *Пшеничный Д. О.*[2]**
[1]dsd@tsu.tula.ru, [2]denispshenichny@yandex.com
[1,2]Тульский государственный университет

В задачах интеллектуального анализа экспериментальные данные часто сразу представлены результатами парных сравнений объектов между собой. В отсутствие исходного признакового пространства условием корректного погружения данного множества объектов в метрическое пространство является неотрицательная определенность матрицы парных близостей элементов множества друг к другу. В этом случае близости интерпретируются как скалярные произведения, а соответствующие различия — как расстояния. В работе рассмотрены условия возникновения метрических нарушений и предложен подход к коррекции метрических нарушений в матрицах парных сравнений за счет минимальных изменений значений некоторых их элементов.

Ключевые слова: *метрика, расстояние, близость, собственные числа, детерминант, парные сравнения.*

On metric correction of matrices of pairwise comparisons*

***Dvoenko S. D.*[1], *Pshenichny D. O.*[2]**
[1,2]State University of Tula

In machine learning and data mining, the experimental results are often immediately represented as matrices of pairwise comparisons between set elements. The condition of the correct immersion of the given set of objects into a metric space is a nonnegative definiteness of the pairwise similarity matrix. In this case, similarities are interpreted as scalar products and dissimilarities as distances, respectively. In this paper, the metric violation conditions are under investigation. The approach to correct metric violations in pairwise comparison matrices is developed based on an idea to minimize distortions of some elements.

Keywords: *metrics, distance, similarity, eigenvalues, determinant, pairwise comparisons.*

Введение

Условием корректного погружения множества в метрическое пространство является неотрицательная определенность матрицы парных близостей его элементов [1]. В этом случае, применяя соответствующие «беспризнаковые» версии алгоритмов машинного обучения и кластер-анализа [2, 3] к матрице сходства (скалярные произведения или неотрицательные величины — близости) или к соответствующей ей матрице различий (расстояния), мы получим математически корректный результат обработки.

В данной работе рассматривается подход к регулируемой коррекции матриц парных сравнений с целью устранения метрических нарушений. Предполагается, что элементы множества, представленные матрицей парных сравнений, требуется погрузить в некоторое метрическое пространство, например, евклидово.

Как известно, в общем случае метрические нарушения на данном множестве возникают при нарушении неравенства треугольника на некоторых тройках его элементов. Более

Работа выполнена при финансовой поддержке РФФИ, проект №13-07-00010.

жестким условием является невыполнение теоремы косинусов, когда углы в треугольнике не соответствуют длинам его сторон. В этом случае соответствующая матрица скалярных произведений тройки элементов оказывается неположительно определенной.

Следует отметить, что развиваемый нами подход имеет свой аналог в задаче экспертного оценивания, когда ранжируют элементы множества. В такой задаче часто требуется представить результат экспертизы в виде парных сравнений, причем от экспертов принято не требовать транзитивности их индивидуальных мнений. При этом исправление нарушений возлагается на метод построения ранжировки. Ранжирование означает проведение измерений в т.н. менее мощных ранговых шкалах.

Ограниченность допустимых операций в ранговых шкалах позволяет ввести простые условия транзитивности отношений, чтобы получить «сверхтранзитивную» матрицу парных сравнений, для которой выполняются условия Льюса [4, 5]. В нашем же случае измерения проводятся в более мощных, чем ранговые или интервальные, шкалах, что предполагает возможность выполнения всех обычных преобразований результатов измерений.

С другой стороны, также следует отметить, что развиваемый нами подход отличается задачи шкалирования [6, 7]. В задаче шкалирования требуется восстановить содержательно интерпретируемое пространство «стимулов». Дополнительно важным требованием обычно является минимизация его размерности. В этой связи в задаче шкалирования возникает проблема «аддитивной константы». Требуется добавить к результатам парных сравнений такую константу, чтобы матрица парных сравнений в наибольшей степени походила на матрицу скалярных произведений и имела, по возможности, наибольшее число нулевых и, если есть, то небольших по модулю отрицательных собственных значений [6]. Тогда размерность пространства стимулов определяется на основе дискретного разложения Карунена-Лоэва, которое сохраняет до 80% дисперсии исходных данных [8].

Заметим, в методе Карунена-Лоэва степень изменения значений элементов матрицы парных сравнений никак не контролируется. Кроме того, в нашем случае не требуется явного восстановления неизвестного нам признакового пространства.

Следует также отметить, что, просто восстанавливая метрическую конфигурацию элементов множества в соответствии с принципом наименьших искажений [9], мы неизбежно получаем матрицы парных сравнений с близким к нулю положительным детерминантом. Поэтому, вообще говоря, мы рискуем дополнительно получить и плохо обусловленную матрицу. Проблема плохо обусловленных матриц хорошо известна. Для обращения таких матриц разработаны весьма продвинутые вычислительные алгоритмы. В нашем подходе предполагается, что регулируемая коррекция некоторых элементов матрицы парных сравнений позволит достичь не только ее положительной определенности, но и некоторого (при более глубокой коррекции) допустимого уровня «плохой определенности».

В данной работе показано, что метрические нарушения возникают не только на тройках, но и на подмножествах, содержащих больше элементов. Такие нарушения во взаимном расположении элементов относительно друг друга также приводят к неположительной определенности матрицы парных сравнений, даже если на всех тройках элементов метрических нарушений нет. Предложены алгоритмы коррекции матриц парных сравнений. Проблема допустимого уровня плохой определенности здесь не рассматривается.

Коррекция метрических нарушений в треугольнике методом восстановления по элементам строки

Пусть множество объектов представлено парными сравнениями в виде симметричной матрицы $S(n,n)$ с элементами $s_{ij}, i = 1,\dots,n,\ \ j = 1,\dots,n$, где n – число объектов.

Недиагональные элементы принимают значения $-1 < s_{ij} < 1$ или $0 \leqslant s_{ij} < 1$. Если данная матрица положительно определена, то ее элементы рассматриваются как нормированные скалярные произведения между соответствующими векторами в метрическом пространстве и представлены косинусами углов между ними. Тогда в тройке векторов с индексами i, j и k сходство первого с самим собой $s_{ii} = 1$, с двумя остальными $s_{ij} = \cos\alpha$ и $s_{ik} = \cos\beta$.

Значение s_{ij} определяет все положения вектора с индексом j относительно вектора с индексом i как гиперконус, опирающийся на гиперокружность с центром на оси вектора с индексом i. Аналогичный смысл имеет и значение s_{ik}. Тогда относительно фиксированного вектора с индексом i возможные значения s_{jk} определены всеми парами векторов с индексами j и k, концы которых расположены на соответствующих гиперокружностях. Эти векторы наиболее близки друг к другу $c_1^{(ijk)} = s_{jk} = \cos(\beta - \alpha)$, когда расположены на одной линии по одну сторону от вектора с индексом i , и наименее близки $c_2^{(ijk)} = s_{jk} = \cos(\beta + \alpha)$, когда расположены от него по разные стороны. Если величина сходства понимается как значение неотрицательной функции близости, то $\beta + \alpha \leqslant \pi/2$. По формулам преобразования косинуса разности и суммы аргументов получим $c_{1,2}^{(ijk)} = s_{ij}s_{ik} \pm \sqrt{(1 - s_{ij}^2)(1 - s_{ik}^2)}$. Метрическое нарушение означает нарушение диапазона $c_2^{(ijk)} \leqslant s_{jk} \leqslant c_1^{(ijk)}$.

Коррекция методом восстановления по элементам строки заключается в следующем. Рассмотрим в матрице близостей S строку i, которая определяет значения сходства $s_{ij}, j = 1, \ldots, n$ вектора с индексом i с другими векторами. Относительно данного вектора все остальные образуют соответствующие гиперконусы, опирающиеся на соответствующие гиперокружности. Эти гиперокружности определяют все возможные диапазоны значений парных близостей остальных векторов множества между собой относительно данного вектора с индексом i. Поэтому по элементам каждой строки i матрицы близостей можно восстановить матрицы близостей $S^{(i)}, i = 1, \ldots, n$ с элементами в диапазонах $c_2^{(ijk)} \leqslant s_{jk}^{(i)} \leqslant c_1^{(ijk)}$.

Матрицы $S^{(i)}$ являются симметричными, с единичными главными диагоналями. В каждой из них строка i совпадает с такой же строкой в исходной матрице S . Если в матрице $S^{(i)}$ все элементы определены в соответствии с их диапазонами, то любая тройка ее элементов удовлетворяет теореме косинусов.

В общем случае для каждого значения s_{jk} можно определить не более n диапазонов для значений $s_{jk}^{(i)}$, попадание в которые не нарушит метрические соотношения. Сформируем матрицу близостей S^*, в которой для каждого ее элемента s_{jk}^* определен единственный диапазон $c_2^{*(jk)} \leqslant s_{jk}^* \leqslant c_1^{*(jk)}$, где $c_1^{*(jk)} = \min\limits_i c_1^{(ijk)}$ и $c_2^{*(jk)} = \max\limits_i c_2^{(ijk)}$. Для минимальной коррекции нужно выбрать, например, то значение $c_1^{*(ij)}$ или $c_2^{*(ij)}$, к которому исходное значение s_{ij} ближе всего. В итоге получим скорректированную матрицу $\tilde{S}$, в которой нет метрических нарушений на тройках элементов исходной матрицы близостей S.

Тем не менее скорректированная матрица $\tilde{S}$ снова может оказаться неположительно определенной. Это означает, что метрические нарушения также возникли на подмножествах, содержащих более трех элементов.

Коррекция нормированной матрицы близостей

Ранее нами была предложена процедура регулируемой коррекции нормированной матрицы парных близостей для устранения ее отрицательной определенности [9]. Для нормированной матрицы было показано, что метрические нарушения возникают при невыпол-

нении теоремы косинусов, как на тройках, так и на подмножествах, содержащих большее число элементов.

В последнем случае элементы множества поочередно погружаются в специальным образом построенное координатное пространство, где корректное положение очередного добавляемого элемента определяется относительно всех ранее добавленных элементов радиусом соответствующей гиперсферы всех возможных его положений. Метрическое нарушение означает невозможность построения такой гиперсферы из-за того, что ее радиус оказывается комплексным числом.

В такой процедуре главный минор нормированной матрицы парных близостей уменьшается, начиная с единицы, оставаясь положительным, при добавлении очередного элемента множества. Если на множестве элементов возникают метрические нарушения, то главный минор текущего размера оказывается знакопеременным, постепенно уменьшаясь по модулю. Отрицательность очередного главного минора означает, что очередной добавленный элемент множества внес метрическое нарушение. Его следует исправить, корректируя значения парных сравнений нового элемента с предыдущими элементами до получения неотрицательного главного минора текущего размера.

Коррекция матрицы близостей может быть выполнена двумя способами. В первом случае корректируются все парные сравнения объекта, внесшего метрическое нарушение (вектор парных сравнений), с другими объектами. Во втором случае в матрице близостей корректируется подходящий одиночный элемент, соответствующий только одному сравнению объекта, внесшего метрическое нарушение, с некоторым другим объектом.

Коррекция вектора парных сравнений. Элементы множества просматриваются в некотором порядке, формируя множество уже просмотренных элементов, представленных на k–ом шаге главным минором $S_k = S(k,k)$, $k = 1, \ldots, n$ матрицы парных сравнений $S(n,n)$. Если минор S_k отрицателен, то считается, что именно k–й объект внес метрическое нарушение. Этот объект представлен своими парными сравнениями с предыдущими $k-1$ объектами и формирует k–ю строку и k–й столбец главного минора S_k.

Замена вектора сравнений этого элемента с предыдущими на нулевой вектор-орт единичной длины не изменит предыдущего значения детерминанта $S_k^{ort} = S_{k-1} > 0$. Действительно, в миноре S_k^{ort} такой элемент представлен нулевыми строкой и столбцом с единицей на главной диагонали. Вычисление этого минора на основе разложения по элементам последней строки заключается лишь в вычислении положительного минора S_{k-1}. Но орт-вектор парных сравнений слишком далек от вектора, вызвавшего нарушение.

Поэтому орт-вектор следует развернуть в направлении исходного вектора до положения, когда новый вектор сравнений будет не слишком сильно отличаться от исходного, а минор S_k окажется положительным.

Процедура корректировки заключается в следующем. Для заданного порога $\varepsilon > 0$ отклонения от нуля детерминант S'_k варьируется в пределах от $P_1 = S_k < 0$ до $P_2 = S_k^{ort} > 0$, где $S'_k = (P_1 + P_2)/2$. Если $S'_k \leqslant 0$, то $P_1 = S'_k$. Иначе, если $S'_k > 0$, то $P_2 = S'_k$. Если $0 \leqslant S'_k \leqslant \varepsilon$, то остановить процедуру.

Коррекция одиночных элементов. Очевидно, что одиночные коррекции уменьшают искажения, вносимые в результаты исходных парных сравнений. Далее, там, где это не вызывает затруднений, для представления значения некоторого минора будем применять такое же обозначение, как и для самого минора.

Раскроем главный минор S_k по элементам $k-$й строки

$$S_k = \sum_{p=1}^{k}(-1)^{k+p}s_{kp}\,(M_k)_p^k = \sum_{p=1}^{k-1}(-1)^{k+p}s_{kp}\,(M_k)_p^k + (-1)^{k+k}s_{kk}\,(M_k)_k^k\,,$$

где $(M_k)_p^k$ является дополнительным минором, полученным из главного минора S_k при удалении из него $k-$ой строки и $p-$го столбца, а второе слагаемое является предыдущим главным минором: $(-1)^{k+k}s_{kk}\,(M_k)_k^k = (M_k)_k^k = S_{k-1}$.

Далее раскроем миноры $(M_k)_p^k$ по элементам $k-$го столбца, сохранив индексацию строк и столбцов относительно исходного минора S_k. Получим

$$S_k = S_{k-1} + \sum_{p=1}^{k-1}(-1)^{k+p}s_{kp}\Big(\sum_{q=1}^{k-1}(-1)^{(q+k)-1}s_{qk}\big((M_k)_p^k\big)_k^q\Big).$$

Предполагая, что корректируются симметричные элементы $s_{kl} = s_{lk}$, представим главный минор S_k как функцию от их значений $S_k(s_{kl})$, $l = 1, \ldots, k-1$. Выполнив преобразования, получим условие коррекции каждого из элементов:

$$S_k(s_{kl}) = A_l s_{kl}^2 + B_l s_{kl} + C_l > 0,$$

$$A_l = (-1)^{2k+2l-1}\big((M_k)_l^k\big)_k^l = -\big((M_k)_l^k\big)_k^l,$$

$$B_l = 2\sum_{q=1,\ q\neq l}^{k-1}(-1)^{2k+q+l-1}s_{qk}\big((M_k)_l^k\big)_k^q,$$

$$C_l = S_{k-1} + \sum_{p=1,\ p\neq l}^{k-1}\ \sum_{q=1,\ q\neq l}^{k-1}(-1)^{2k+p+q-1}s_{kp}s_{qk}\big((M_k)_p^k\big)_k^q\,.$$

Из решения следует, что коррекция симметричных элементов $s_{kl} = s_{lk}$ возможна в диапазоне $c_2^{(l)} \leqslant s_{kl} \leqslant c_1^{(l)}$, где $c_{1,2}^{(l)} = \frac{1}{2A_l}\left(-B_l \pm \sqrt{B_l^2 - 4A_lC_l}\right)$ при $B_l^2 - 4A_lC_l > 0$.

Если возможна коррекция несколько элементов, то следует выбрать тот, для которого интервал $c_1^{(l)} - c_2^{(l)}$ оказался наибольшим, и взять значение $s_{kl} = \left(c_1^{(l)} + c_2^{(l)}\right)/2$.

В этом случае будет получено максимальное положительное значение главного минора S_k, а при добавлении очередного элемента главный минор уменьшится $S_{k+1} < S_k$ в наименьшей степени.

Если коррекция одиночных элементов невозможна, то придется корректировать все парные сравнения (вектор) объекта, внесшего метрическое нарушение. В общем случае возникает задача оптимальной коррекции нескольких элементов: пар, троек и т.д. Но здесь мы ее не рассматриваем.

Оптимальная последовательность корректировок. В общем случае элементы данного множества могут быть просмотрены и в другом порядке. Может оказаться и так, что другая последовательность элементов данного множества приведет к меньшей величине суммарных искажений значений элементов исходной матрицы парных сравнений.

Пусть k–й объект внес метрическое нарушение. Пусть его коррекция выполняется одним из способов, описанных выше. Будем размещать его поочередно $q = 1, \ldots, k$ на всех

местах и, выполняя каждый раз корректировку, найдем минор $S'_{k(q)}$, который в наименьшей степени отличается от минора S_k. Приняв $S_k = S'_{k(q)}$, перейдем к следующему $k+1$ объекту.

Заметим, что перестановка объекта, внесшего метрическое нарушение, может вызвать дополнительные нарушения и от других объектов. Поэтому задача поиска наименьших искажений исходной матрицы является комбинаторной, требуя просмотра всех перестановок элементов множества в общем случае.

Локализация отрицательных собственных значений. Известно, что одновременная перестановка строк и столбцов матрицы $S(n,n)$ не изменяет ее собственных значений. Согласно закону инерции квадратичных форм [10] также известно, что число смен знаков при просмотре главных миноров $S_k, k = 1, \ldots, n$ совпадает с числом отрицательных собственных значений матрицы $S(n,n)$.

Определим такой порядок просмотра элементов множества, чтобы смены знаков главных миноров $S_k, k = 1, \ldots, n$ происходили преимущественно в конце последовательности. Пусть матрица $S(n,n)$ ранга n имеет v отрицательных собственных значений. Тогда в идеальном случае соответствующая перестановка элементов множества определит такой порядок просмотра его элементов, при котором главный минор $S_{n-v+1} < 0$ впервые окажется отрицательным, а знаки последующих $v-1$ миноров будут чередоваться. В этом случае нужно будет скорректировать парные сравнения не более чем v объектов. В этом смысле отрицательные собственные значения окажутся локализованными в неположительно определенной матрице парных сравнений.

Будем удалять из матрицы парных сравнений $S(n,n)$ очередные строку и столбец, вычисляя дополнительные миноры $(M_n)_i^i$, $i = 1, \ldots, n$, и возвращать их на место. Определитель S_n матрицы $S(n,n)$ равен произведению ее собственных значений. Если этот определитель отрицателен, то имеется нечетное число отрицательных собственных значений, если положителен — то четное число. Найдем строку и столбец с номером i_1, для которых дополнительный минор $(M_n)_{i_1}^{i_1}$ сменит знак по сравнению с главным минором $M_n = S_n$ и окажется максимальным по модулю их всех возможных при $i = 1, \ldots, n$. Тогда при перемещении строки и столбца с номером i_1 на последнее место с номером n окажется, что последняя смена знака произойдет на последнем миноре $M_n = S_n$.

Продолжим этот процесс, рассматривая главные миноры S_k, $k = n-1, \ldots, 1$ и находя соответствующие дополнительные миноры $(M_k)_{i_q}^{i_q}$, $1 \leqslant i_q \leqslant k$. Для локализации всех отрицательных собственных значений может потребоваться более чем v шагов, так как на некоторых из них смены знака минора может не происходить. В этих случаях в очередном главном миноре S_k просто ищется соответствующий максимальный дополнительный минор $(M_k)_{i_q}^{i_q}$ без смены его знака. Пусть u – число всех дополнительных шагов, когда до локализации всех v собственных значений не происходило смены знака соответствующего главного минора. Это означает, что получена матрица $S(n,n)$, у которой придется корректировать не более $v+u$ последних элементов множества в последовательности.

Таким образом, группировка в конце последовательности объектов, предположительно вносящих метрические искажения, позволяет уменьшить дополнительные нарушения, которые могут возникнуть от других объектов в процессе коррекции.

Коррекция ненормированной матрицы близостей

Пусть множество объектов представлено взаимными ненормированными близостями в виде матрицы $S'(n,n)$ с элементами $s'_{ij}, i = 1, \ldots, n$, $j = 1, \ldots, n$, где n–число объектов.

По теореме косинусов получим $s'_{ij} = (d^2_{0i}+d^2_{0j}-d^2_{ij})/2$, считая элементы матрицы $S'(n,n)$ скалярными произведениями относительно начала координат как объекта с индексом 0, где d_{ij} – расстояние между объектами с индексами i и j, d_{0i} – расстояние объекта с индексом i до начала координат. Тогда $s'_{ii} = d^2_{0i}$. Таким образом, ненормированная матрица близостей отражает конфигурацию элементов множества в метрическом пространстве, представляя их расстояния до начала координат.

Если в матрице $S'(n,n)$ имеются метрические нарушения, то их можно выявить и скорректировать, как описано выше, предварительно приведя матрицу $S'(n,n)$ к нормированному виду преобразованием $s_{ij} = s'_{ij}/\sqrt{s'_{ii}s'_{jj}}$. К сожалению, в нормированной матрице $S(n,n)$ не сохраняется исходная конфигурация объектов в пространстве, так как $s_{ii} = 1$ и «нормированные» объекты расположены на гиперсфере единичного радиуса.

Считая, что расстояния до начала координат не изменились, восстановим после корректировки ненормированную матрицу близостей $S'(n,n)$ с элементами $s'_{ij} = s_{ij}d_{0i}d_{0j}$.

В общем случае некоторые расстояния до начала координат могут измениться. Данный случай здесь рассматривать не будем.

Коррекция матрицы различий

Пусть множество объектов представлено взаимными различиями в виде матрицы $D(n,n)$ с элементами $d_{ij},\ i = 1,\dots,n,\ j = 1,\dots,n$, где n – число объектов.

Будем считать различия расстояниями. Чтобы получить матрицу скалярных произведений $S(n,n)$, нужно назначить начало координат, относительного которого можно вычислить скалярные произведения. Его можно назначить в любом месте координатного пространства, например, в центре «тяжести» множества по методу Торгерсона [7].

Для размещения начала координат в центре тяжести множества объектов по методу Торгерсона получим новый объект с индексом 0, представленный своими расстояниями до остальных объектов с индексами $i = 1,\dots,n$ следующим образом:

$$d^2_{0i} = \frac{1}{n}\sum_{p=1}^{n} d^2_{ip} - \frac{1}{2n^2}\sum_{p=1}^{n}\sum_{q=1}^{n} d^2_{pq}. \tag{1}$$

Но иногда такое размещение начала координат не совсем удобно, особенно, если в исходной матрице парных сравнений имеются метрические нарушения. В этом случае некоторые значения (1) отрицательны, а соответствующие расстояния до начала координат оказываются комплексными.

Нужно так выбрать начало координат как новый объект, чтобы он был представлен своими действительными расстояниями до всех объектов исходного множества. Такой новый объект должен лежать вне выпуклой оболочки, образованной данным множеством объектов. Тогда, считая его началом координат, можно определить относительно него все скалярные произведения.

По теореме косинусов элементы матрицы нормированных скалярных произведений $S(n,n)$ множества объектов относительно нового объекта с индексом 0 как нового центра координат вычисляются как

$$s_{ij} = \frac{1}{2d_{0i}d_{0j}}(d^2_{0i} + d^2_{0j} - d^2_{ij}). \tag{2}$$

Рассмотрим выражение (1). Можно увидеть, что второе слагаемое определяет дисперсию (разброс) элементов как отклонений положения объектов от начала координат в

центре тяжести множества:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} d_{0i}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{p=1}^{n} d_{ip}^2 - \frac{1}{2n^2}\sum_{p=1}^{n}\sum_{q=1}^{n} d_{pq}^2\right) =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{p=1}^{n} d_{ip}^2 - \frac{n}{n}\left(\frac{1}{2n^2}\sum_{p=1}^{n}\sum_{q=1}^{n} d_{pq}^2\right) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{p=1}^{n} d_{ip}^2.$$

Рассмотрим первое слагаемое в (1). Рассмотрим расстояния $d_{ip}, p = 1, \ldots, n$ от объекта с индексом i до остальных объектов с индексами p как компоненты соответствующего вектора в некотором n–мерном метрическом пространстве, которое нам удобно назвать «вторичным». Тогда величина $\sum\limits_{p=1}^{n} d_{ip}^2$ представляет собой квадрат нормы этого вектора, т.е. квадрат расстояния от начала координат, а первое слагаемое из (1) представляет собой средневзвешенный квадрат этой нормы.

Таким образом, начало координат в таком вторичном пространстве будет представлено как объект с индексом 0^0 своими расстояниями до остальных объектов $d_{0^0 i}^2, i = 1, \ldots, n$, инвариантными относительно размера множества.

Тогда начало координат по методу Торгерсона будет представлено как объект с индексом 0 своими расстояниями $d_{0i}^2 = d_{0^0 i}^2 - \Delta$, где $\Delta = \sigma^2$.

Легко увидеть, что изменение константы Δ позволит получить начало координат не в центре тяжести множества.

При $\Delta = 0$ начало координат как объект с индексом 0 будет максимально удалено от центра тяжести множества. Условие $\Delta = 0$ можно понимать как предельно маленький разброс элементов множества по сравнению с расстояниями до начала координат. Очевидно, что в этом случае скалярные произведения будут близки к единице, если получившиеся расстояния до начала координат окажутся значительными. Если $\Delta < 0$, то это свойство только усилится.

При $\Delta > \sigma^2$ обязательно возникнет неметрическая конфигурация, когда не удастся получить, согласно (2), корректный вид матрицы нормированных скалярных произведений $S(n, n)$. Это произойдет, когда некоторые из расстояний $d_{0i}^2 = d_{0^0 i}^2 - \Delta$ до начала координат окажутся нулевыми или отрицательными. Условие $\Delta > \sigma^2$ можно понимать как увеличенный по сравнению с реальным разброс элементов множества. В итоге, допустимые значения величины Δ находятся в интервале $0 \leqslant \Delta \leqslant \sigma^2$.

В общем случае метричность конфигурации определяется наличием неотрицательных собственных чисел матрицы нормированных скалярных произведений относительно назначенного начала координат. Поэтому следует так разместить начало координат, чтобы матрица нормированных скалярных произведений относительно него была бы неотрицательно или положительно определена.

Если в исходной матрице расстояний имелись нарушения метрики, то полученную матрицу скалярных произведений $S(n, n)$ необходимо откорректировать.

После исправления метрических нарушений получим положительно определенную матрицу скалярных произведений $S(n, n)$.

Далее восстановим ненормированную матрицу $S'(n, n)$ скалярных произведений с элементами $s'_{ij} = s_{ij} d_{0i} d_{0j}$ и матрицу расстояний $D(n, n)$, считая, что расстояния от объектов до начала координат не изменились, где $d_{ij}^2 = s'_{ii} + s'_{jj} - 2s'_{ij} = d_{0i}^2 + d_{0j}^2 - 2s_{ij} d_{0i} d_{0j}$.

Эксперименты

Эксперименты были проведены на трех матрицах расстояний между некоторыми населенными пунктами Тульской области. Первая матрица была получена путем непосредственного измерения по карте, в результате которого некоторые расстояния были намеренно искажены:

$$D_C = \begin{pmatrix} & Т & А & Н & Б & Я & Ар & Е \\ Т & 0 & 67 & 62 & 123 & 50 & 105 & 150 \\ А & 67 & 0 & 144 & 149 & 45 & 128 & 218 \\ Н & 62 & 144 & 0 & 180 & 116 & 164 & 135 \\ Б & 123 & 149 & 180 & 0 & 160 & 75 & 231 \\ Я & 50 & 45 & 116 & 160 & 0 & 142 & 186 \\ Ар & 105 & 128 & 164 & 75 & 142 & 0 & 202 \\ Е & 150 & 218 & 135 & 231 & 186 & 202 & 0 \end{pmatrix}.$$

В данной матрице имеются нарушения неравенства треугольника на некоторых тройках элементов. Для трех городов: *Тула*, *Алексин* и *Новомосковск*, суммарное расстояние от *Тулы* до *Алексина* и *Новомосковска* (129 км) оказалось меньше расстояния от *Новомосковска* до *Алексина* (144 км). Также суммарное расстояние от *Тулы* до *Алексина* и *Ефремова* (217 км) оказалось меньше расстояния от *Алексина* до *Ефремова* (218 км). Наконец, суммарное расстояние от *Тулы* до *Новомосковска* и *Ясногорска* (112 км) меньше расстояние от *Новомосковска* до *Ясногорска* (116 км).

Вторая матрица была получена с помощью сервиса Google*:

$$D_G = \begin{pmatrix} & Т & А & Н & Б & Я & Ар & Е \\ Т & 0 & 48 & 42 & 106 & 31 & 81 & 121 \\ А & 48 & 0 & 88 & 99 & 40 & 90 & 165 \\ Н & 42 & 88 & 0 & 140 & 55 & 110 & 104 \\ Б & 106 & 99 & 140 & 0 & 126 & 36 & 150 \\ Я & 31 & 40 & 55 & 126 & 0 & 107 & 150 \\ Ар & 81 & 90 & 110 & 36 & 107 & 0 & 116 \\ Е & 121 & 165 & 104 & 150 & 150 & 116 & 0 \end{pmatrix}$$

и третья матрица была получена с помощью сервиса Yandex:

$$D_Y = \begin{pmatrix} & Т & А & Н & Б & Я & Ар & Е \\ Т & 0 & 50.3 & 49.7 & 107 & 31.8 & 81 & 121 \\ А & 50.3 & 0 & 98.6 & 99.9 & 39.7 & 91.1 & 166 \\ Н & 49.7 & 98.6 & 0 & 144 & 67.2 & 111 & 94.9 \\ Б & 107 & 99.9 & 144 & 0 & 126 & 36.2 & 149 \\ Я & 31.8 & 39.7 & 67.2 & 126 & 0 & 106 & 150 \\ Ар & 81 & 91.1 & 111 & 36.2 & 106 & 0 & 115 \\ Е & 121 & 166 & 94.9 & 149 & 150 & 115 & 0 \end{pmatrix}.$$

В этих матрицах неравенства треугольника выполняются.

*Здесь и далее Т - Тула, А - Алексин, Н - Новомосковск, Б - Белёв, Я - Ясногорск, Ар - Арсеньево, Е - Ефремов (исправления внесены 16/01/15 по запросу автора)

Для всех матриц расстояний были получены матрицы нормированных скалярных произведений и их собственные значения:

$$S_C = \begin{pmatrix} 1 & 0.8611 & 0.8933 & 0.5631 & 0.9174 & 0.6085 & 0.5237 \\ 0.8611 & 1 & 0.3659 & 0.4244 & 0.9354 & 0.5107 & -0.0073 \\ 0.8933 & 0.3659 & 1 & 0.1613 & 0.5615 & 0.2038 & 0.6472 \\ 0.5631 & 0.4244 & 0.1613 & 1 & 0.2981 & 0.8633 & -0.0060 \\ 0.9174 & 0.9354 & 0.5615 & 0.2981 & 1 & 0.3581 & 0.2505 \\ 0.6085 & 0.5107 & 0.2038 & 0.8633 & 0.3581 & 1 & 0.1614 \\ 0.5237 & -0.0073 & 0.6472 & -0.0060 & 0.2505 & 0.1614 & 1 \end{pmatrix},$$

$$\boldsymbol{\lambda}_{S_C}^{\mathsf{T}} = (4.075 \;\; 1.571 \;\; 1.047 \;\; 0.307 \;\; 0.112 \;\; 0.009 \;\; -0.121),$$

$$S_G = \begin{pmatrix} 1 & 0.8468 & 0.8826 & 0.3566 & 0.9454 & 0.4995 & 0.3573 \\ 0.8468 & 1 & 0.5153 & 0.5071 & 0.9002 & 0.4836 & -0.1427 \\ 0.8826 & 0.5153 & 1 & -0.0139 & 0.8083 & 0.2159 & 0.5784 \\ 0.3566 & 0.5071 & -0.0139 & 1 & 0.1843 & 0.9522 & 0.1797 \\ 0.9454 & 0.9002 & 0.8083 & 0.1843 & 1 & 0.2606 & 0.0577 \\ 0.4995 & 0.4836 & 0.2159 & 0.9522 & 0.2606 & 1 & 0.4556 \\ 0.3573 & -0.1427 & 0.5784 & 0.1797 & 0.0577 & 0.4556 & 1 \end{pmatrix},$$

$$\boldsymbol{\lambda}_{S_G}^{\mathsf{T}} = (3.994 \;\; 1.681 \;\; 1.326 \;\; 0.008 \;\; -0.001 \;\; -0.004 \;\; -0.005),$$

$$S_Y = \begin{pmatrix} 1 & 0.8392 & 0.8406 & 0.3568 & 0.9428 & 0.5049 & 0.3487 \\ 0.8392 & 1 & 0.4254 & 0.5115 & 0.9055 & 0.4850 & -0.1493 \\ 0.8406 & 0.4254 & 1 & -0.0349 & 0.7264 & 0.2283 & 0.6547 \\ 0.3568 & 0.5115 & -0.0349 & 1 & 0.1995 & 0.9536 & 0.1870 \\ 0.9428 & 0.9055 & 0.7264 & 0.1995 & 1 & 0.2830 & 0.0530 \\ 0.5049 & 0.4850 & 0.2283 & 0.9536 & 0.2830 & 1 & 0.4572 \\ 0.3487 & -0.1493 & 0.6547 & 0.1870 & 0.0530 & 0.4572 & 1 \end{pmatrix},$$

$$\boldsymbol{\lambda}_{S_Y}^{\mathsf{T}} = (3.94 \;\; 1.639 \;\; 1.418 \;\; 0.007 \;\; 0.002 \;\; -0.001 \;\; -0.005).$$

Все нормированные матрицы были получены при $\Delta = 0$ относительно начала координат, как объекта с индексом 0, вынесенного за пределы выпуклой оболочки данного множества элементов на соответствующие расстояния:

$$\boldsymbol{d}_{0_C}^{\mathsf{T}} = (92.19 \;\; 127.26 \;\; 128.48 \;\; 148.48 \;\; 118.53 \;\; 131.42 \;\; 176.07),$$

$$\boldsymbol{d}_{0_G}^{\mathsf{T}} = (73.16 \;\; 90.060 \;\; 88.680 \;\; 107.11 \;\; 88.970 \;\; 86.980 \;\; 126.003),$$

$$\boldsymbol{d}_{0_Y}^{\mathsf{T}} = (74.32 \;\; 92.320 \;\; 91.630 \;\; 107.94 \;\; 90.010 \;\; 86.980 \;\; 124.860).$$

На всех тройках объектов, представленных парными близостями в матрице S_C, для которых в матрице D_C было нарушено неравенство треугольника:

$$S_{--} = \begin{pmatrix} 1 & 0.8611 & 0.8933 \\ 0.8611 & 1 & 0.3659 \\ 0.8933 & 0.3659 & 1 \end{pmatrix}, \; \boldsymbol{\lambda} = \begin{pmatrix} 2.437 \\ 0.634 \\ -0.071 \end{pmatrix},$$

$$S_{--} = \begin{pmatrix} 1 & 0.8611 & 0.5237 \\ 0.8611 & 1 & -0.0073 \\ 0.5237 & -0.0073 & 1 \end{pmatrix}, \; \boldsymbol{\lambda} = \begin{pmatrix} 2.005 \\ 1.007 \\ -0.011 \end{pmatrix},$$

$$S_{--} = \begin{pmatrix} 1 & 0.8933 & 0.9174 \\ 0.8933 & 1 & 0.5615 \\ 0.9174 & 0.5615 & 1 \end{pmatrix}, \ \boldsymbol{\lambda} = \begin{pmatrix} 2.592 \\ 0.439 \\ -0.030 \end{pmatrix}$$

были скорректированы близости по элементам строк этих подматриц.

Выполним коррекцию близостей по элементам первой строки матрицы S_{--}. В матрице $S^{(1)}$ элементы первой строки не изменятся, элементы второй строки $s_{21}^{(1)}$ и $s_{22}^{(1)}$ также не изменятся из-за симметрии, а для элемента $s_{23}^{(1)}$ определен диапазон $0.5407 \leqslant s_{23}^{(1)} \leqslant 0.9977$:

$$c_{1,2}^{(123)} = s_{12}s_{13} \pm \sqrt{(1-s_{12}^2)(1-s_{13}^2)} = 0.7692 \pm 0.2285 \, .$$

Значение $s_{23} = 0.3659$ за пределами диапазона вносит метрическое искажение относительно элементов первой строки исходной матрицы. Элементы третьей строки уже определены в силу симметрии. Можно убедиться, что для элементов $s_{31}^{(1)}$ и $s_{33}^{(1)}$ их диапазоны содержат исходные значения s_{31} и $s_{33} = 1$:

$$c_{1,2}^{(131)} = s_{13}s_{11} \pm \sqrt{(1-s_{13}^2)(1-s_{11}^2)} = s_{13} \pm 0 = s_{13},$$

$$c_{1,2}^{(133)} = s_{13}s_{13} \pm \sqrt{(1-s_{13}^2)(1-s_{13}^2)} = s_{13}^2 \pm (1-s_{13}^2),$$

а $s_{32} = 0.3659$ также находится за пределами диапазона $0.4443 \leqslant s_{32}^{(1)} \leqslant 0.9939$:

$$c_{1,2}^{(132)} = s_{13}s_{12} \pm \sqrt{(1-s_{13}^2)(1-s_{12}^2)} = 0.7692 \pm 0.2285 \, .$$

В итоге, восстановление по элементам каждой из строк дает матрицы

$$S^{(1)} = \begin{pmatrix} 1 & 0.8611 & 0.8933 \\ 0.8611 & 1 & 0.5407 \div 0,9977 \\ 0.8933 & 0.5407 \div 0.9977 & 1 \end{pmatrix},$$

$$S^{(2)} = \begin{pmatrix} 1 & 0.8611 & 0.1581 \div 0.7883 \\ 0.8611 & 1 & 0.3659 \\ 0.1581 \div 0.7883 & 0.3659 & 1 \end{pmatrix},$$

$$S^{(3)} = \begin{pmatrix} 1 & -0.0914 \div 0,7452 & 0.8933 \\ -0.0914 \div 0.7452 & 1 & 0.3659 \\ 0.8933 & 0.3659 & 1 \end{pmatrix}.$$

Объединение диапазонов дает матрицу без метрических нарушений

$$\tilde{S}_{--} = \begin{pmatrix} 1 & 0.7452 & 0.7883 \\ 0.7452 & 1 & 0.5407 \\ 0.7883 & 0.5407 & 1 \end{pmatrix}, \ \boldsymbol{\lambda} = \begin{pmatrix} 2.388 \\ 0.461 \\ 0.151 \end{pmatrix}.$$

Так же были получены и две другие матрицы, где для повторяющихся пар городов *Тула-Алексин* и *Тула-Новомосковск* были взяты одни и те же значения близостей, так как они попали в соответствующие диапазоны:

$$\tilde{S}_{--} = \begin{pmatrix} 1 & 0.7452 & 0.5021 \\ 0.7452 & 1 & 0.0178 \\ 0.5021 & 0.0178 & 1 \end{pmatrix}, \ \boldsymbol{\lambda} = \begin{pmatrix} 1.907 \\ 0.984 \\ 0.110 \end{pmatrix},$$

$$\tilde{S}_{--} = \begin{pmatrix} 1 & 0.7883 & 0.8735 \\ 0.7883 & 1 & 0.6406 \\ 0.8735 & 0.6406 & 1 \end{pmatrix}, \ \boldsymbol{\lambda} = \begin{pmatrix} 2.539 \\ 0.367 \\ 0.094 \end{pmatrix}.$$

При объединении всех корректировок была получена матрица близостей, в которой нет ни одного нарушения неравенства треугольника. Тем не менее у нее снова оказалось одно отрицательное собственное значение

$$\boldsymbol{\lambda}^{\mathsf{T}}_{\tilde{S}_C} = (4.048 \ \ 1.544 \ \ 1.012 \ \ 0.237 \ \ 0.139 \ \ 0.049 \ \ -0.029).$$

Поэтому дополнительно была выполнена коррекция ее симметричных элементов s_{25} и s_{52} на основе процедуры локализации отрицательных собственных значений. Все измененные значения показаны жирным шрифтом. Полученная матрица

$$\tilde{S}_C = \begin{pmatrix} 1 & \mathbf{0.7452} & \mathbf{0.7883} & 0.5631 & \mathbf{0.8735} & 0.6085 & \mathbf{0.5021} \\ \mathbf{0.7452} & 1 & \mathbf{0.5407} & 0.4244 & \mathbf{0.8729} & 0.5107 & \mathbf{0.0178} \\ \mathbf{0.7883} & \mathbf{0.5407} & 1 & 0.1613 & \mathbf{0.6406} & 0.2038 & 0.6472 \\ 0.5631 & 0.4244 & 0.1613 & 1 & 0.2981 & 0.8633 & -0.0060 \\ \mathbf{0.8735} & \mathbf{0.8729} & \mathbf{0.6406} & 0.2981 & 1 & 0.3581 & 0.2505 \\ 0.6085 & 0.5107 & 0.2038 & 0.8633 & 0.3581 & 1 & 0.1614 \\ \mathbf{0.5021} & \mathbf{0.0178} & 0.6472 & -0.0060 & 0.2505 & 0.1614 & 1 \end{pmatrix}$$

имеет суммарное отличие от исходной 0.6282 и имеет только положительные собственные значения

$$\boldsymbol{\lambda}^{\mathsf{T}}_{\tilde{S}_C} = (4.025 \ \ 1.545 \ \ 0.986 \ \ 0.232 \ \ 0.162 \ \ 0.05 \ \ 0.000005).$$

Восстановленная матрица расстояний имеет вид

$$\tilde{D}_C = \begin{pmatrix} & \mathit{Т} & \mathit{А} & \mathit{Н} & \mathit{Б} & \mathit{Я} & \mathit{Ар} & \mathit{Е} \\ \mathit{Т} & 0 & 84.9 & 79.6 & 123 & 58.8 & 105 & 152.3 \\ \mathit{А} & 84.9 & 0 & 122.6 & 149 & 62.5 & 128 & 215.4 \\ \mathit{Н} & 79.6 & 122.6 & 0 & 180 & 105.1 & 164 & 135 \\ \mathit{Б} & 123 & 149 & 180 & 0 & 160 & 75 & 231 \\ \mathit{Я} & 58.8 & 62.5 & 105.1 & 160 & 0 & 142 & 186 \\ \mathit{Ар} & 105 & 128 & 164 & 75 & 142 & 0 & 202 \\ \mathit{Е} & 152.3 & 215.4 & 135 & 231 & 186 & 202 & 0 \end{pmatrix}.$$

Для матрицы S_C также была выполнена корректировка без предварительного устранения нарушений неравенства треугольника. Коррекция одиночных элементов оказалась невозможна, поэтому была выполнена коррекция вектором на основе процедуры локализации отрицательных собственных значений. Полученная матрица

$$S_C = \begin{pmatrix} 1 & \mathbf{0.7434} & \mathbf{0.7712} & \mathbf{0.4861} & \mathbf{0.7919} & \mathbf{0.5253} & \mathbf{0.4521} \\ \mathbf{0.7434} & 1 & 0.3659 & 0.4244 & 0.9354 & 0.5107 & -0.0073 \\ \mathbf{0.7712} & 0.3659 & 1 & 0.1613 & 0.5615 & 0.2038 & 0.6472 \\ \mathbf{0.4861} & 0.4244 & 0.1613 & 1 & 0.2981 & 0.8633 & -0.0060 \\ \mathbf{0.7919} & 0.9354 & 0.5615 & 0.2981 & 1 & 0.3581 & 0.2505 \\ \mathbf{0.5253} & 0.5107 & 0.2038 & 0.8633 & 0.3581 & 1 & 0.1614 \\ \mathbf{0.4521} & -0.0073 & 0.6472 & -0.0060 & 0.2505 & 0.1614 & 1 \end{pmatrix}$$

имеет суммарное отличие от исходной 0.5971 и собственные значения

$$\boldsymbol{\lambda}^{\mathsf{T}}_{S_C} = (3.861 \ \ 1.569 \ \ 1.047 \ \ 0.314 \ \ 0.112 \ \ 0.096 \ \ 0.0003).$$

Восстановленная матрица расстояний имеет вид

$$D_C = \begin{pmatrix} & Т & А & Н & Б & Я & Ар & Е \\ Т & 0 & 85.2 & 82.1 & 131.3 & 72.4 & 114.2 & 157.6 \\ А & 85.2 & 0 & 144 & 149 & 45 & 129 & 218 \\ Н & 82.1 & 144 & 0 & 180 & 116 & 164 & 135 \\ Б & 131.3 & 149 & 180 & 0 & 160 & 75 & 231 \\ Я & 72.4 & 45 & 116 & 160 & 0 & 142 & 186 \\ Ар & 114.2 & 128 & 164 & 75 & 142 & 0 & 202 \\ Е & 157.6 & 218 & 135 & 231 & 186 & 202 & 0 \end{pmatrix}.$$

Корректировка матриц S_G и S_Y также была выполнена основе процедуры локализации отрицательных собственных значений, их суммарные отличия от исходных матриц составляют 0.0605 и 0.0298. При этом корректировались как одиночные элементы, так и векторы:

$$S_G = \begin{pmatrix} 1 & \mathbf{0.8402} & \mathbf{0.8757} & \mathbf{0.3538} & \mathbf{0.9380} & \mathbf{0.4956} & \mathbf{0.3546} \\ \mathbf{0.8402} & 1 & 0.5153 & 0.5071 & 0.9002 & \mathbf{0.4798} & -0.1427 \\ \mathbf{0.8757} & 0.5153 & 1 & -0.0139 & \mathbf{0.7967} & \mathbf{0.2142} & 0.5784 \\ \mathbf{0.3538} & 0.5071 & -0.0139 & 1 & 0.1843 & \mathbf{0.9448} & 0.1797 \\ \mathbf{0.9380} & 0.9002 & \mathbf{0.7967} & 0.1843 & 1 & \mathbf{0.2585} & 0.0577 \\ \mathbf{0.4956} & \mathbf{0.4798} & \mathbf{0.2142} & \mathbf{0.9448} & \mathbf{0.2585} & 1 & \mathbf{0.4520} \\ \mathbf{0.3546} & -0.1427 & 0.5784 & 0.1797 & 0.0577 & \mathbf{0.4520} & 1 \end{pmatrix},$$

$$\boldsymbol{\lambda}_{S_G}^{\mathsf{T}} = (3.975 \;\; 1.674 \;\; 1.327 \;\; 0.014 \;\; 0.006 \;\; 0.004 \;\; 0.0003),$$

$$S_Y = \begin{pmatrix} 1 & 0.8392 & \mathbf{0.8341} & 0.3568 & 0.9428 & 0.5049 & 0.3487 \\ 0.8392 & 1 & \mathbf{0.4221} & \mathbf{0.5044} & 0.9055 & 0.4850 & -0.1493 \\ \mathbf{0.8341} & \mathbf{0.4221} & 1 & \mathbf{-0.0346} & \mathbf{0.7207} & \mathbf{0.2266} & \mathbf{0.6496} \\ 0.3568 & \mathbf{0.5044} & \mathbf{-0.0346} & 1 & 0.1995 & 0.9536 & 0.1870 \\ 0.9428 & 0.9055 & \mathbf{0.7207} & 0.1995 & 1 & 0.2830 & 0.0530 \\ 0.5049 & 0.4850 & \mathbf{0.2266} & 0.9536 & 0.2830 & 1 & 0.4572 \\ 0.3487 & -0.1493 & \mathbf{0.6496} & 0.1870 & 0.0530 & 0.4572 & 1 \end{pmatrix},$$

$$\boldsymbol{\lambda}_{S_Y}^{\mathsf{T}} = (3.931 \;\; 1.638 \;\; 1.417 \;\; 0.009 \;\; 0.004 \;\; 0.001 \;\; 0.0001).$$

Восстановленные матрицы расстояний имеют вид:

$$D_G = \begin{pmatrix} & Т & А & Н & Б & Я & Ар & Е \\ Т & 0 & 48.9 & 43.1 & 106.2 & 32.5 & 81.3 & 121.2 \\ А & 48.9 & 0 & 88 & 99 & 40 & 90.3 & 165 \\ Н & 43.1 & 88 & 0 & 140 & 56.6 & 110.1 & 104 \\ Б & 106.2 & 99 & 140 & 0 & 126 & 37.9 & 150 \\ Я & 32.5 & 40 & 56.6 & 126 & 0 & 107.1 & 150 \\ Ар & 81.3 & 90.3 & 110.1 & 37.9 & 107.1 & 0 & 116.3 \\ Е & 121.2 & 165 & 104 & 150 & 150 & 116.3 & 0 \end{pmatrix},$$

$$D_Y = \begin{pmatrix} & Т & А & Н & Б & Я & Ар & Е \\ Т & 0 & 50.3 & 50.6 & 107 & 31.8 & 81 & 121 \\ А & 50.3 & 0 & 98.9 & 100.6 & 39.7 & 91.1 & 166 \\ Н & 50.6 & 98.9 & 0 & 144 & 67.9 & 111.1 & 95.5 \\ Б & 107 & 100.6 & 144 & 0 & 126 & 36.2 & 149 \\ Я & 31.8 & 39.7 & 67.9 & 126 & 0 & 106 & 150 \\ Ар & 81 & 91.1 & 111.1 & 36.2 & 106 & 0 & 115 \\ Е & 121 & 166 & 95.5 & 149 & 150 & 115 & 0 \end{pmatrix}.$$

Результаты работы процедуры локализации отрицательных собственных значений представлены в таблице 1.

Таблица 1: Локализация отрицательных собственных значений

Матрица	n	v	Перестановка объектов	Позиция $n-v+1$	Коррекция объекта
$\tilde{S}_C$	7	1	7 4 2 3 6 1 5	7	5
S_C	7	1	7 4 5 3 6 2 1	7	1
S_G	7	3	7 2 4 3 5 1 6	5	5 1 6
S_Y	7	2	7 1 6 2 5 4 3	6	4 3

Все рассмотренные выше матрицы скалярных произведений имеют ранг $n = 7$ и свое число v отрицательных собственных значений. Процедура локализации выявляет оптимальную перестановку объектов и позицию $n - v + 1$ первого отрицательного главного минора. В соответствии с этими позициями оказалось, что в каждой матрице корректировалось минимальное число объектов v, в точности совпадающее с соответствующим числом отрицательных собственных значений.

Заключение

Что есть истина?
(И. 18: 33, 36-38)

В данной работе рассмотрен один из возможных подходов к погружению элементов множества в метрическое пространство и проведены эксперименты по коррекции расстояний на примере некоторых населенных пунктов Тульской области.

Условием корректного погружения является отсутствие метрических нарушений во взаимном расположении элементов множества друг относительно друга.

Следует отметить, что в задаче коррекции нет речи о восстановлении каких-то «истинных» расстояний, так как условием корректировки является только восстановление положительной определенности соответствующей матрицы парных сравнений. Это видно из результата восстановления первой матрицы расстояний, в которую были внесены большие искажения. В восстановленной и теперь корректной матрице эти искажения «истинных» расстояний остались.

Показано, что суть нарушений как на тройках, так и на большем числе элементов множества, заключается, в итоге, в нарушении теоремы косинусов. Это удалось показать в явном виде [9] только для нормированной матрицы парных сравнений, имеющих смысл скалярных произведений.

В этих условиях потребовалась разработка специальных процедур для корректировки произвольных матриц парных сравнений, которые могут быть матрицами расстояний и ненормированными матрицами близостей или скалярных произведений.

Литература

[1] *Young G., Householder A. S.* Discussion of a set of points in terms of their mutual distances // *Psychometrica*. 1938. Vol. 3, no. 1. P. 19–22.

[2] *Двоенко С. Д.* Распознавание элементов множества, представленных взаимными расстояниями и близостями // *ММРО-14*. М.: МАКС Пресс, 2009. С. 112–115.

[3] *Двоенко С. Д.* Кластеризация множества, описанного парными расстояниями и близостями между его элементами // *Сибирский журнал индустриальной математики*. 2009. Т. 12, № 1(37). С. 61–73.

[4] *Киселев Ю. В.* Оценка важности программ методом парных сравнений // *Известия АН СССР. Техническая кибернетика*. 1971. № 3. С. 41–47.

[5] *Миркин Б. Г.* Проблема группового выбора. М.: Наука, 1974. 256 с.

[6] *Терехина А. Ю.* Анализ данных методами многомерного шкалирования. М.: Наука, 1986. 168 с.

[7] *Дэйвисон М.* Многомерное шкалирование. М.: Финансы и статистика, 1988. 254 с.

[8] *Ту Дж., Гонсалес Р.* Принципы распознавания образов. М.: Мир, 1978. 411 с.

[9] *Двоенко С. Д., Пшеничный Д. О.* Об устранении отрицательных собственных значений матриц парных сравнений // *ИОИ-2012*. М.: ТОРУС ПРЕСС, 2012. С. 13–16.

[10] *Гантмахер Ф. Р.* Теория матриц. М.: Наука, 1988. 552 с.

Восстановление симметричности точек на изображениях объектов с отражательной симметрией*

Каркищенко А. Н.[1]***, Мнухин В. Б.***[2]
[1]karkishalex@gmail.com
[1]Северо-Кавказский Федеральный Университет; [2]Южный Федеральный Университет

В работе предлагается несколько алгебраических методов, которые позволяют уточнить положение характерных точек, описывающих какие-либо объекты на изображении, на основе априорно известной информации об их симметричном расположении. Эти методы называются симметризацией характерных точек. Рассмотрена симметризация точек для случая вертикальной и произвольной симметрии с известными параметрами оси симметрии, а также более общий случай симметризации при неизвестных параметрах осевой симметрии. Рассматриваемые методы дают решение задачи осевой симметризации при минимальном изменении положения характерных точек. **Ключевые слова**: *симметризация, отражательная симметрия, характерные точки, биометрическая идентификация, распознавание лиц.*

Recovery of points symmetry in images of objects with reflectional symmetry*

Karkishchenko A. N.[1]***, Mnukhin V. B.***[2]
[1]North Caucasian Federal University, [2]Southern Federal University

In this work, we consider the problem to obtain more accurate information about location of points based on a priori knowledge of their symmetries. Methods to solve this symmetrization problem with respect to vertical and inclined axes of reflectional symmetry are considered jointly with the more general case of reflectional symmetry with respect to an indefinite reflection axis. The methods produce the minimal deformation that enhances approximate symmetries present in a given arrangement of points.

Keywords: *symmetrization, reflectional symmetry, feature points, biometrical identification, face recognition.*

Введение

Симметрия играет значительную роль в правильном восприятии как естественных, так и в большей степени искусственно созданных объектов. В последнее время многочисленные усилия при анализе формы объектов, заданных в оцифрованном виде, были сосредоточены на выявлении симметрии 2D и 3D объектов [1–3]. Информация о симметрии эффективно используется в многочисленных приложениях — в компактном описании моделей [3], обработке сканированных изображений [4], сегментации изображений [5], установлении соответствия форм [1] и др. Как правило, на «атомарном» уровне методы сводятся к исследованию симметричности так называемых характерных точек, смысл которых зависит от решаемой задачи.

В данной работе предлагается ряд методов, которые позволяют уточнить положение характерных точек, описывающих объекты на изображении, на основе известной инфор-

Работа выполнена при финансовой поддержке РФФИ, проекты № 11-07-00591 и № 13-07-00327.

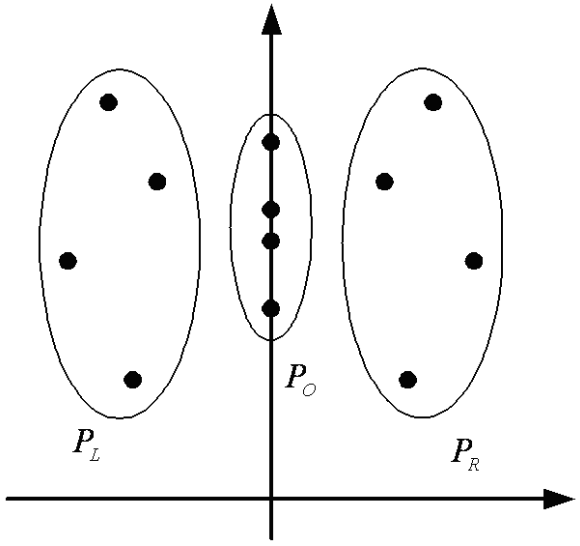


Рис. 1: Симметрия относительно вертикальной оси.

мации об их симметричном расположении. Суть проблемы состоит в том, что локализация точек на изображении всегда осуществляется с некоторой погрешностью, величина которой зависит от многих факторов — используемого алгоритма локализации, качества изображения в целом, зашумления в области, содержащей локализуемую точку и др. Как следствие, координаты обнаруженных точек, про которые априори известно, что они симметричны, в реальности этому условию не удовлетворяют. Поэтому целесообразно воспользоваться известной априорной информацией о симметричности точек для уточнения их положения. При этом само уточнение необходимо осуществлять в некотором смысле наилучшим образом, скажем, симметричность должна достигаться минимальным изменениям их положения.

Одной из наиболее типичных задач, где могут применяться данные методы уточнения «по симметрии», являются задачи биометрического распознавания, в которых правильное определение характерных точек является критическим фактором успешного решения задачи. Здесь достаточно указать на то, что точность детекции и распознавания лиц существенно зависит от точности определения центров зрачков на лице [6]. При этом в задачах идентификации личности, основанных на определении положения характерных точек, заведомо присутствует фактор симметричности, обусловленный симметричностью формы лица анфас [7]. В данных методах определяется, как правило, несколько десятков (в зависимости от метода) характерных точек, большая часть из которых представляет собой пары точек, зеркально симметричных относительно вертикальной оси, либо точки, лежащие непосредственно на оси симметрии.

Постановка и решение задачи при вертикальной осевой симметрии характерных точек

Пусть $P = \{p_1, \ldots, p_n\}$ — множество всех характерных точек, и $p_k = (x_k, y_k)$ — координаты k-той точки. Предположим, что ось симметрии совпадает с осью ординат в прямоугольной системе координат. Условимся нумеровать характерные точки так, что точки $P_R = \{p_1, \ldots, p_m\}$ относятся к правой полуплоскости, $P_L = \{p_{m+1}, \ldots, p_{2m}\}$ — к левой, а точки $P_O = \{p_{2m+1}, \ldots, p_n\}$ лежат на оси симметрии.

Сопоставим с упорядоченным таким образом множеством характерных точек P вектор X размерности $2n$:

$$(p_1, \ldots, p_n) \quad \longleftrightarrow \quad (x_1, \ldots, x_n,\ y_1, \ldots, y_n)^\intercal = X. \tag{1}$$

Множество векторов X образует линейное пространство $\mathbb{R}^{2n}$. Заметим теперь, что если все характерные точки найдены точно, то, с учетом разбиения на классы P_R, P_L, P_O, должны выполняться следующие условия:

$$\begin{array}{rclrcl} x_i & = & -x_{m+i}, & i & = & 1, \ldots, m; \\ y_i & = & y_{m+i}, & i & = & 1, \ldots, m; \\ x_i & = & 0, & i & = & 2m+1, \ldots, n. \end{array}$$

Множество векторов, удовлетворяющих этим условиям, образует n-мерное подпространство в $\mathbb{R}^{2n}$. Обозначим это подпространство R_{Sym}; оно состоит из векторов, соответствующих характерным точкам, расположенным симметрично относительно оси ординат. Заметим, что «симметризация» вектора X характерных точек означает построение в R_{Sym} вектора X_s, наименее отклоняющегося от X по евклидовой норме:

$$X_s = \underset{Z \in R_{Sym}}{\arg\min} \|Z - X\|.$$

Таким образом, задача сводится к нахождению ортогональной проекции X_s вектора X на подпространство R_{Sym}.

Пусть Q — матрица оператора ортогонального проецирования на подпространство R_{Sym}, а A — матрица, столбцами которой служат базисные векторы подпространства R_{Sym}. Тогда, как известно ([8], стр. 165),

$$Q = AA^+ = A\left(A^\intercal A\right)^{-1} A^\intercal,$$

где A^+ — псевдообратная к матрице A. Для нахождения A^+ заметим, что матрица A имеет следующую блочную структуру:

$$A^\intercal = \begin{pmatrix} I_m & -I_m & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_m & I_m & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{n-2m} \end{pmatrix}, \tag{2}$$

где I и 0 — соответственно единичная и нулевая матрицы соответствующих размеров. Тогда, как следует из непосредственных вычислений,

$$Q = \frac{1}{2} I_{2n} + \frac{1}{2} \begin{pmatrix} -S & 0 \\ 0 & S \end{pmatrix}, \tag{3}$$

где S — следующая $n \times n$-матрица:

$$S = \begin{pmatrix} 0 & I_m & 0 \\ I_m & 0 & 0 \\ 0 & 0 & I_{n-2m} \end{pmatrix}. \tag{4}$$

Тем самым, решение поставленной задачи в общем случае имеет вид

$$X_s = QX\ . \tag{5}$$

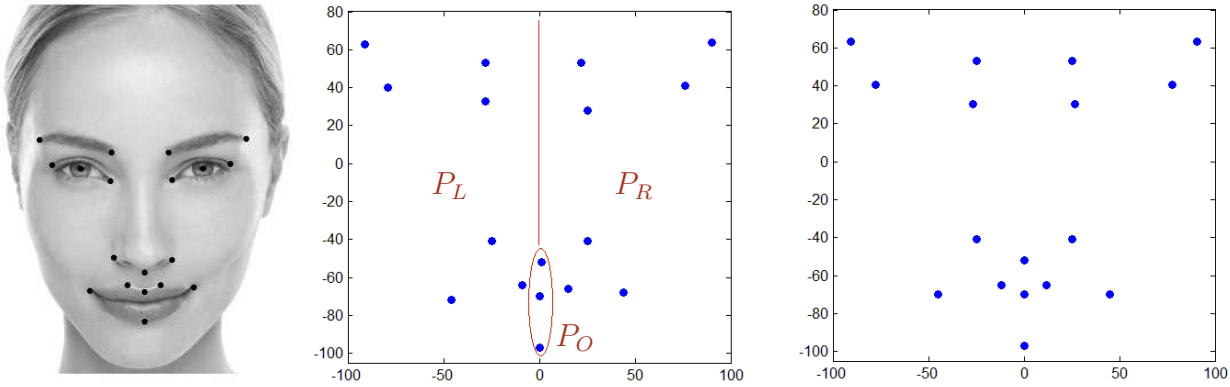


Рис. 2: Пример симметризации относительно вертикальной оси.

Нетрудно видеть, что ортогональное проектирование вектора приводит к простым операциям над координатами характерных точек: в качестве ординаты симметричных точек надо взять среднее арифметическое их ординат, а модули абсцисс равны среднему арифметическому их модулей. Для «осевых» точек их ординаты остаются неизменными, а абсциссы обнуляются.

Результат симметризации системы точек относительно вертикальной оси симметрии с помощью рассмотренного выше метода показан на рис. 2. Его левая часть показывает лицо анфас с отмеченными характерными точками, присутствующее, например, в некоторой базе данных. На средней части рисунка показаны характерные точки, положение которых найдено в результате работы соответствующего алгоритма, например, применяемого в [6]. Как нетрудно заметить, симметрия точек оказывается нарушенной, что затрудняет задачу идентификации образа. В правой части рисунка показан результат симметризации системы характерных точек с помощью рассмотренного выше метода.

Симметризация при произвольной осевой симметрии

Пусть теперь ось симметрии задается уравнением $y = ax + b$, где $a \neq 0$, см. рис. 3. В этом случае задача сводится к предыдущей путем замены координат. А именно, построим новую систему координат $(O'x'y')$, приняв, что ось $O'y'$ совпадает с осью симметрии, начало координат O' совпадает с точкой пересечения оси симметрии и оси Oy, а ось $O'x'$ перпендикулярна оси $O'y'$ и образует правоориентированную систему. Другими словами, система $(O'x'y')$ получается из (Oxy) поворотом последней на некоторый угол φ, как показано на рис. 3. Тогда угловой коэффициент оси симметрии равен $a = \operatorname{tg}(\frac{\pi}{2} + \varphi) = -\operatorname{ctg}\varphi$, а её уравнение при $\varphi \neq 0$ можно записать как $y = b - x\operatorname{ctg}\varphi$.

Пусть

$$\mathscr{R} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}$$

— матрица оператора поворота на угол $-\varphi$, причем её элементы $\sin\varphi$ и $\cos\varphi$ получаются из условия $\operatorname{ctg}\varphi = -a \neq 0$, т.е.,

$$\sin\varphi = \frac{1}{\sqrt{a^2+1}}, \qquad \cos\varphi = -\frac{a}{\sqrt{a^2+1}}.$$

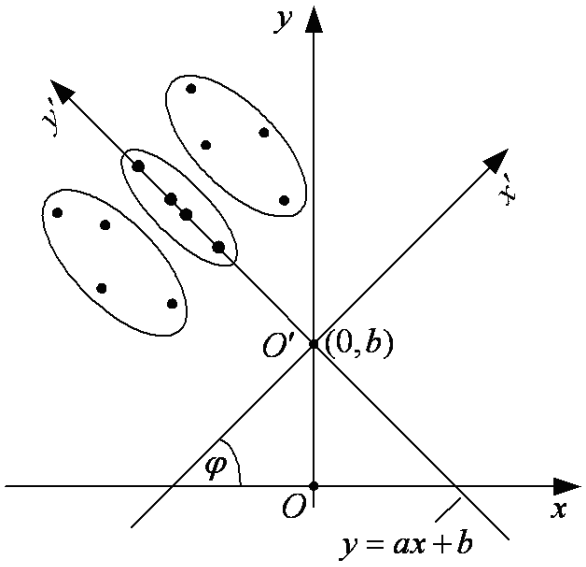


Рис. 3: Симметрия относительно произвольной оси.

Тогда координаты произвольной точки (x, y) в старой системе координат и (x', y') — в новой системе связаны соотношением

$$(x', y')^{\mathsf{T}} = \mathscr{R}(x, y-b)^{\mathsf{T}}\ , \qquad \text{или} \qquad \left\{ \begin{array}{rcrcl} x' & = & x\cos\varphi & + & (y-b)\sin\varphi \\ y' & = & -x\sin\varphi & + & (y-b)\cos\varphi \end{array} \right. .$$

Рассмотрим, как и в предыдущем разделе, вектор

$$X = (x_1, \ldots, x_n,\ y_1, \ldots, y_n)^{\mathsf{T}} \in \mathbb{R}^{2n}$$

составленный из координат характерных точек в системе (Oxy) с учетом разбиения на классы P_R, P_L, P_O, и пусть вектор X' состоит из координат тех же точек в системе $(O'x'y')$. Чтобы установить связь между векторами X и X', рассмотрим матрицу

$$R^{\mathsf{T}} = \mathscr{R} \otimes I_n = \begin{pmatrix} \cos\varphi \cdot I_n & \sin\varphi \cdot I_n \\ -\sin\varphi \cdot I_n & \cos\varphi \cdot I_n \end{pmatrix}, \tag{6}$$

где $\otimes$ означает кронекерово произведение. Из его свойств [9] немедленно вытекает, что матрица R ортогональна, $R^{-1} = R^{\mathsf{T}}$. Тогда

$$X' = R^{\mathsf{T}}(X - bK), \qquad \text{так что} \qquad X = RX' + bK, \tag{7}$$

где K — следующий вектор: $K = (0, \ldots, 0,\ 1, \ldots, 1)^{\mathsf{T}}$.

Преобразование (7) сводит рассматриваемую задачу к решенной ранее. А именно, симметризируя вектор X', и возвращая затем результат в старую систему координат, получаем:

$$X_s = RX'_s + bK = R\big(QR^{\mathsf{T}}(X - bK)\big) + bK\ ,$$

или

$$X_s = GX + b(I - G)K\ , \qquad \text{где} \qquad G = RQR^{\mathsf{T}}\ .$$

Как следует из непосредственных вычислений,

$$G = \frac{1}{2}\, I_{2n} + \widetilde{G}, \tag{8}$$

где

$$\widetilde{G} = \frac{1}{2}\begin{pmatrix} -\cos 2\varphi \cdot S & \sin 2\varphi \cdot S \\ \sin 2\varphi \cdot S & \cos 2\varphi \cdot S \end{pmatrix}, \tag{9}$$

причем матрица S определяется выражением (4). При $b = 0$ и $\varphi = 0$ получаем найденное выше решение (5).

Симметризация с отысканием параметров оси симметрии

Покажем, что на основе рассмотренного метода можно не только осуществлять симметризацию относительно известной оси, но и определять неизвестные параметры оси симметрии таким образом, что симметризация будет достигаться минимальным изменением положения точек.

Пусть, как и ранее, $P = \{p_1, \ldots, p_n\}$ — множество всех характерных точек, заданных своими координатами $p_k = (x_k, y_k)$. Допустим, что на основе некоторой априорной информации, (скажем, того, что рассматриваются характерные точки изображения лица человека анфас), множество P разбито на непересекающиеся классы

$$P_R = \{p_1, \ldots, p_m\}, \qquad P_L = \{p_{m+1}, \ldots, p_{2m}\} \qquad \text{и} \qquad P_O = \{p_{2m+1}, \ldots, p_n\}$$

таким образом, что соответствующие точки $p_i \in P_R$ и $p_{m+i} \in P_L$ первых двух классов должны быть симметричны относительно некоторой заранее неизвестной оси симметрии, а точки класса P_O — лежать на этой оси. Учитывая это разбиение, сопоставим, как и в предыдущем разделе, с множеством $P = P_R \cup P_L \cup P_O$ вектор

$$X = (x_1, \ldots, x_n,\ y_1, \ldots, y_n)^{\mathsf{T}} \in \mathbb{R}^{2n}$$

исходных координат характерных точек. Кроме того, для решения задачи удобно использовать следующие обозначения

$$\sigma = (\underbrace{1, \ldots, 1}_{n}, \underbrace{1, \ldots, 1}_{n})^{\mathsf{T}}, \qquad \sigma_0^1 = (\underbrace{1, \ldots, 1}_{n}, \underbrace{0, \ldots, 0}_{n})^{\mathsf{T}}, \qquad \sigma_1^0 = (\underbrace{0, \ldots, 0}_{n}, \underbrace{1, \ldots, 1}_{n})^{\mathsf{T}}.$$

Будем считать, что ось симметрии задана уравнением $y = ax + b$ с неизвестными пока параметрами $a = -\operatorname{ctg}\varphi \neq 0$ и b. Рассмотрим вектор

$$X'(\varphi, b) = R^{\mathsf{T}}(\varphi)\big(X - b\sigma_1^0\big),$$

где матрица $R^{\mathsf{T}}(\varphi)$ преобразования вращения зависит от угла поворота φ и определяется выражением (6). Вспоминая соотношение (7) предыдущего раздела, заметим, что соответствующая оптимизационная задача может быть записана в виде

$$\Big\| AY - R^{\mathsf{T}}(\varphi)\big(X - b\sigma_1^0\big) \Big\|^2 \xrightarrow{Y,\varphi,b} \min\,,$$

где матрица A задается выражением (2) первого раздела, а $Y \in \mathbb{R}^{2n}$ — переменный вектор. Иными словами, для отыскания оптимальной симметризации X_s необходимо найти φ_s, b_s и вектор Y_s, минимизирующие левую часть предыдущего выражения, после чего вернуть найденный вектор Y_s в исходную систему координат:

$$X_s = R(\varphi_s)Y_s + b_s\sigma_1^0\,.$$

Построенный вектор X_s будет определять оптимальную симметризацию данного множества характерных точек.

Приступая к решению поставленной задачи, обозначим для удобства минимизируемое выражение через

$$F(Y,\varphi,b) \stackrel{\text{def}}{=\!=} \left\| AY - R^{\mathsf{T}}(\varphi)\big(X - b\sigma_1^0\big) \right\|^2$$

и распишем его подробно в матричном виде:

$$\begin{aligned} F(Y,\varphi,b) &= \Big(\; AY - R^{\mathsf{T}}(\varphi)X + bR^{\mathsf{T}}(\varphi)\sigma_1^0, \;\; AY - R^{\mathsf{T}}(\varphi)X + bR^{\mathsf{T}}(\varphi)\sigma_1^0 \; \Big) = \\ &= \Big(\; AY - R^{\mathsf{T}}(\varphi)X + bR^{\mathsf{T}}(\varphi)\sigma_1^0 \; \Big)^{\mathsf{T}} \cdot \Big(\; AY - R^{\mathsf{T}}(\varphi)X + bR^{\mathsf{T}}(\varphi)\sigma_1^0 \; \Big) \\ &= Y^{\mathsf{T}}A^{\mathsf{T}}AY - 2Y^{\mathsf{T}}A^{\mathsf{T}}R^{\mathsf{T}}(\varphi)X + 2bY^{\mathsf{T}}A^{\mathsf{T}}R^{\mathsf{T}}(\varphi)\sigma_1^0 + X^{\mathsf{T}}X - 2bX^{\mathsf{T}}\sigma_1^0 + b^2\big(\sigma_1^0\big)^{\mathsf{T}}\sigma_1^0 \; . \end{aligned}$$

Для решения оптимизационной задачи необходимо решить систему уравнений

$$\frac{\partial F(Y,\varphi,b)}{\partial Y} = 0, \qquad \frac{\partial F(Y,\varphi,b)}{\partial \varphi} = 0, \qquad \frac{\partial F(Y,\varphi,b)}{\partial b} = 0 \; .$$

После нахождения частных производных и очевидных преобразований получаем:

$$\begin{cases} AY = QR^{\mathsf{T}}(\varphi)(X - b\sigma_1^0), \\ \big(\sigma_1^0\big)^{\mathsf{T}}R(\varphi)AY = \big(\sigma_1^0\big)^{\mathsf{T}}(X - b\sigma_1^0), \\ (X - b\sigma_1^0)\dfrac{dR(\varphi)}{d\varphi}AY = 0. \end{cases}$$

Первое (матричное) уравнение дает выражение для симметризованного вектора в зависимости от φ и b. Второе и третье уравнения являются скалярными и позволяют найти неизвестные параметры оси симметрии φ и b. Выпишем эти уравнения отдельно, выразив в них AY через φ и b с помощью первого уравнения:

$$\begin{cases} (X - b\sigma_1^0)^{\mathsf{T}}\dfrac{dR(\varphi)}{d\varphi}QR^{\mathsf{T}}(\varphi)(X - b\sigma_1^0) = 0, \\ \big(\sigma_1^0\big)^{\mathsf{T}}R(\varphi)QR^{\mathsf{T}}(\varphi)(X - b\sigma_1^0) = \big(\sigma_1^0\big)^{\mathsf{T}}(X - b\sigma_1^0). \end{cases}$$

Заметим, что

$$\frac{dR(\varphi)}{d\varphi} = R(\pi/2)R(\varphi)$$

и примем во внимание, что в силу введенных выше обозначений (8) и (9)

$$R(\varphi)QR^{\mathsf{T}}(\varphi) = G = \frac{1}{2}I + \widetilde{G}.$$

Поэтому систему можно переписать в более простом виде:

$$\begin{cases} (X - b\sigma_1^0)^{\mathsf{T}}R\left(\dfrac{\pi}{2}\right)\left(\dfrac{1}{2}I + \widetilde{G}\right)(X - b\sigma_1^0) = 0, \\ \big(\sigma_1^0\big)^{\mathsf{T}}\widetilde{G}(X - b\sigma_1^0) = \dfrac{1}{2}\big(\sigma_1^0\big)^{\mathsf{T}}(X - b\sigma_1^0). \end{cases}$$

Как нетрудно видеть,

$$(X - b\sigma_1^0)^\intercal R\left(\frac{\pi}{2}\right)(X - b\sigma_1^0) = 0,$$

поэтому система упрощается:

$$\begin{cases} (X - b\sigma_1^0)^\intercal R\left(\frac{\pi}{2}\right)\widetilde{G}(X - b\sigma_1^0) = 0, \\ \\ \qquad\qquad (\sigma_1^0)^\intercal \widetilde{G}(X - b\sigma_1^0) = \dfrac{1}{2}(\sigma_1^0)^\intercal (X - b\sigma_1^0). \end{cases}$$

Для нахождения φ и b проанализируем вначале второе уравнение системы. Нам понадобятся величины

$$x_{av} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{и} \qquad y_{av} = \frac{1}{n}\sum_{i=1}^{n} y_i\ ,$$

представляющие собой соответственно усредненную абсциссу и ординату вычисленных характерных точек. Тогда нетрудно видеть, что правая часть второго уравнения равна

$$\frac{1}{2}(\sigma_1^0)^\intercal (X - b\sigma_1^0) = \frac{n}{2}(y_{av} - b). \tag{10}$$

Рассмотрим теперь левую часть уравнения. Непосредственными вычислениями получаем:

$$(\sigma_1^0)^\intercal \widetilde{G}(X - b\sigma_1^0) =$$

$$= \frac{1}{2}\left[-\sin 2\varphi \cdot \underbrace{(\sigma_0^1)^\intercal X}_{nx_{av}} + \cos 2\varphi \cdot \underbrace{(\sigma_1^0)^\intercal X}_{ny_{av}} + b\sin 2\varphi \cdot \underbrace{(\sigma_0^1)^\intercal \sigma_1^0}_{0} - b\cos 2\varphi \cdot \underbrace{(\sigma_1^0)^\intercal \sigma_1^0}_{n}\right] = \tag{11}$$

$$= \frac{n}{2}\Big(-x_{av}\sin 2\varphi + y_{av}\cos 2\varphi - b\cos 2\varphi\Big).$$

Приравнивая выражения (10) и (11), после несложных преобразований находим значение параметра b:

$$b = y_{av} + x_{av}\operatorname{ctg}\varphi\ . \tag{12}$$

Смысл полученного соотношения вполне понятен из рисунка 4.

Рассмотрим теперь первое уравнение последней системы. Для этого разобьем его на три слагаемых, как показано ниже:

$$(X - b\sigma_1^0)^\intercal R\left(\frac{\pi}{2}\right)\widetilde{G}(X - b\sigma_1^0) =$$

$$= \underbrace{X^\intercal R\left(\frac{\pi}{2}\right)\widetilde{G}X}_{1} - b\underbrace{\left[X^\intercal R\left(\frac{\pi}{2}\right)\widetilde{G}\sigma_1^0 + (\sigma_1^0)^\intercal R\left(\frac{\pi}{2}\right)\widetilde{G}X\right]}_{2} + b^2 \cdot \underbrace{(\sigma_1^0)^\intercal R\left(\frac{\pi}{2}\right)\widetilde{G}\sigma_1^0}_{3}\ .$$

Вычислим последовательно каждое слагаемое.

Первое слагаемое. Вспоминая определение (1) вектора X, представим его в блочном виде:

$$X = (x_1, \ldots, x_n,\ y_1, \ldots, y_n)^\intercal = \left(\boldsymbol{x_1}^\intercal \boldsymbol{x_2}^\intercal \boldsymbol{x_3}^\intercal\ \boldsymbol{y_1}^\intercal \boldsymbol{y_2}^\intercal \boldsymbol{y_3}^\intercal\right)\ ,$$

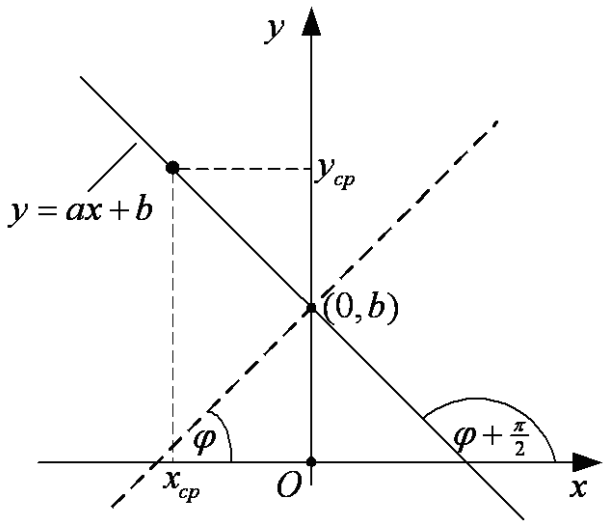


Рис. 4: Нахождение параметра b.

где

$$\boldsymbol{x_1} = (x_1, \ldots, x_m), \qquad \boldsymbol{x_2} = (x_{m+1}, \ldots, x_{2m}), \qquad \boldsymbol{x_3} = (x_{2m+1}, \ldots, x_n),$$
$$\boldsymbol{y_1} = (y_1, \ldots, y_m), \qquad \boldsymbol{y_2} = (y_{m+1}, \ldots, y_{2m}), \qquad \boldsymbol{y_3} = (y_{2m+1}, \ldots, y_n)$$

— векторы x- и y-координат характерных точек, лежащих соответственно правее, левее и на оси симметрии. С учетом этого получаем:

$$X^{\intercal} R\left(\frac{\pi}{2}\right) \widetilde{G} X =$$

$$= -\Big[(\boldsymbol{x_1}, \boldsymbol{y_2}) + (\boldsymbol{x_2}, \boldsymbol{y_1}) + (\boldsymbol{x_3}, \boldsymbol{y_3})\Big] \cos 2\varphi - \Big[(\boldsymbol{y_1}, \boldsymbol{y_2}) + (\boldsymbol{x_1}, \boldsymbol{x_2}) - \tfrac{1}{2}(\boldsymbol{x_3}, \boldsymbol{x_3}) - \tfrac{1}{2}(\boldsymbol{y_3}, \boldsymbol{y_3})\Big] \sin 2\varphi \ .$$

Второе слагаемое. Учитывая, что матрица $\widetilde{G}$ симметрична и $R^{\intercal}\left(\frac{\pi}{2}\right) = -R\left(\frac{\pi}{2}\right)$, перепишем второе слагаемое в виде $\left(\sigma_1^0\right)^{\intercal}\Big[R\left(\frac{\pi}{2}\right)\widetilde{G} - \widetilde{G}R\left(\frac{\pi}{2}\right)\Big]X$. Тогда, с помощью непосредственных вычислений,получаем, что второе слагаемое равняется

$$\Big[-\cos 2\varphi \cdot \left(\sigma_0^1\right)^{\intercal} - \sin 2\varphi \cdot \left(\sigma_1^0\right)^{\intercal}\Big] X = -n x_{av} \cos 2\varphi - n y_{av} \sin 2\varphi \ .$$

Третье слагаемое. Как показывают непосредственные вычисления,

$$\left(\sigma_1^0\right)^{\intercal} R\left(\frac{\pi}{2}\right) = \left(\sigma_0^1\right)^{\intercal}, \qquad \widetilde{G}\sigma_1^0 = \frac{1}{2}\Big(\sigma_1^0 \cos 2\varphi - \sigma_0^1 \sin 2\varphi\Big).$$

Следовательно, третье слагаемое имеет вид

$$\left(\sigma_1^0\right)^{\intercal} R\left(\frac{\pi}{2}\right) \widetilde{G}\sigma_1^0 = \left(\sigma_0^1\right)^{\intercal} \cdot \frac{1}{2}\Big(\sigma_1^0 \cos 2\varphi - \sigma_0^1 \sin 2\varphi\Big) = -\frac{n}{2} \sin 2\varphi \ .$$

Таким образом, второе уравнение принимает вид:

$$\Big[(\boldsymbol{x_1}, \boldsymbol{y_2}) + (\boldsymbol{x_2}, \boldsymbol{y_1}) + (\boldsymbol{x_3}, \boldsymbol{y_3})\Big] \cos 2\varphi + \Big[(\boldsymbol{y_1}, \boldsymbol{y_2}) + (\boldsymbol{x_1}, \boldsymbol{x_2}) - \tfrac{1}{2}(\boldsymbol{x_3}, \boldsymbol{x_3}) - \tfrac{1}{2}(\boldsymbol{y_3}, \boldsymbol{y_3})\Big] \sin 2\varphi -$$
$$- bn\Big[x_{av} \cos 2\varphi + y_{av} \sin 2\varphi\Big] + b^2 \frac{n}{2} \sin 2\varphi = 0 \ .$$

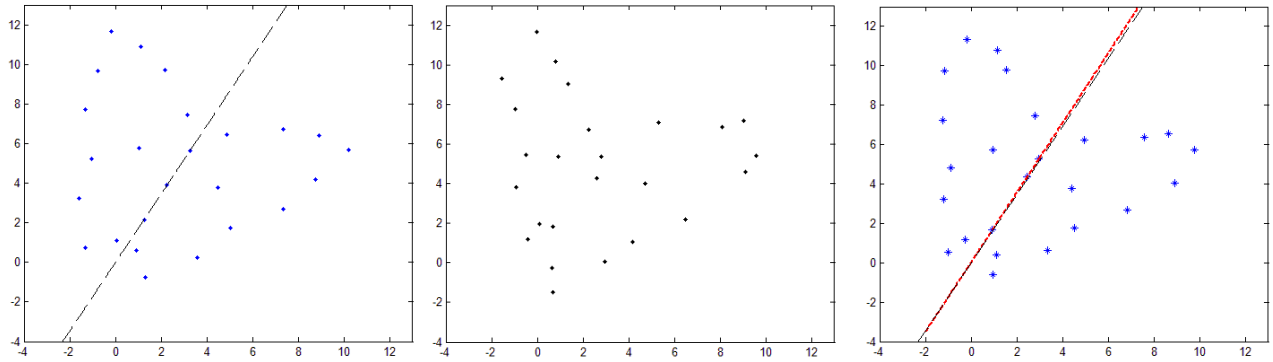


Рис. 5: Пример симметризации с нахождением оси симметрии.

Учитывая, что $b = y_{av} + x_{av}\operatorname{ctg}\varphi$, после преобразований получаем

$$\Big[(\boldsymbol{x_1},\boldsymbol{y_2}) + (\boldsymbol{x_2},\boldsymbol{y_1}) + (\boldsymbol{x_3},\boldsymbol{y_3}) - nx_{av}y_{av}\Big]\cos 2\varphi + \\ + \Big[(\boldsymbol{y_1},\boldsymbol{y_2}) - (\boldsymbol{x_1},\boldsymbol{x_2}) - \tfrac{1}{2}(\boldsymbol{x_3},\boldsymbol{x_3}) + \tfrac{1}{2}(\boldsymbol{y_3},\boldsymbol{y_3}) + \tfrac{1}{2}n\Big(x_{av}^2 - y_{av}^2\Big)\Big]\sin 2\varphi = 0\ ,$$

откуда вытекает, что

$$\operatorname{tg} 2\varphi = \frac{(\boldsymbol{x_1},\boldsymbol{y_2}) + (\boldsymbol{x_2},\boldsymbol{y_1}) + (\boldsymbol{x_3},\boldsymbol{y_3}) - nx_{av}y_{av}}{(\boldsymbol{x_1},\boldsymbol{x_2}) - (\boldsymbol{y_1},\boldsymbol{y_2}) + \tfrac{1}{2}(\boldsymbol{x_3},\boldsymbol{x_3}) - \tfrac{1}{2}(\boldsymbol{y_3},\boldsymbol{y_3}) - \tfrac{1}{2}n\Big(x_{av}^2 - y_{av}^2\Big)}\ .$$

Данному выражению можно придать более удобный вид, если перейти к центрированным координатам

$$\dot{\boldsymbol{x}}_{\boldsymbol{i}} = \boldsymbol{x_i} - x_{av}\boldsymbol{e}, \qquad \dot{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{y_i} - y_{av}\boldsymbol{e}, \qquad (i = 1, 2, 3),$$

где $\boldsymbol{e}$ — вектор соответствующей размерности, состоящий из единиц. Нетрудно видеть, что данное преобразование соответствует переносу начала системы координат в точку с координатами $\big(x_{av},\, y_{av}\big)$. Тогда после преобразований получаем

$$\operatorname{tg} 2\varphi = \frac{(\dot{\boldsymbol{x}}_{\boldsymbol{1}},\dot{\boldsymbol{y}}_{\boldsymbol{2}}) + (\dot{\boldsymbol{x}}_{\boldsymbol{2}},\dot{\boldsymbol{y}}_{\boldsymbol{1}}) + (\dot{\boldsymbol{x}}_{\boldsymbol{3}},\dot{\boldsymbol{y}}_{\boldsymbol{3}})}{(\dot{\boldsymbol{x}}_{\boldsymbol{1}},\dot{\boldsymbol{x}}_{\boldsymbol{2}}) - (\dot{\boldsymbol{y}}_{\boldsymbol{1}},\dot{\boldsymbol{y}}_{\boldsymbol{2}}) + \tfrac{1}{2}(\dot{\boldsymbol{x}}_{\boldsymbol{3}},\dot{\boldsymbol{x}}_{\boldsymbol{3}}) - \tfrac{1}{2}(\dot{\boldsymbol{y}}_{\boldsymbol{3}},\dot{\boldsymbol{y}}_{\boldsymbol{3}})}\ . \tag{13}$$

Подставляя вычисленные по формулам (13) и (12) значения параметров φ и b в первое уравнение системы, находим искомую симметризацию $AY = Z$ исходного вектора характерных точек X. Для завершения решения задачи необходимо вернуть найденный вектор в исходную (старую) систему координат.

Рисунок 5 демонстрирует результат симметризации с помощью рассмотренного выше метода. В левой части рисунка показана система точек, симметричных относительно отмеченной оси симметрии, а в средней — те же точки после зашумления, приведшего к утрате симметричности. Правая часть рисунка показывает точки после процесса симметризации, причем красным показана восстановленная ось симметрии, а черным — исходная.

Заключение

Рассмотренная выше задача симметризации характерных точек относительно осевой симметрии имеет многочисленные приложения, так как это наиболее простой и часто

встречающийся вид симметрии. Вместе с тем, в задачах, связанных с обработкой изображений, встречаются и иные виды симметрии – вращательная, диэдральная, трансляционная и другие. Особый интерес представляет симметризация в условиях аффинных искажений, которые имеют место в большинстве реальных разработок. Строгое математическое решение указанных задач может оказаться не очень простым, но зато предоставит дополнительные возможности для качественной обработки изображений.

Благодарности

Авторы благодарят А. Абраменко, И. Гречухина и С. Левашёва за помощь в проведении вычислительных экспериментов. Авторы также благодарят неизвестного рецензента за ряд замечаний, способствовавших улучшению работы.

Литература

[1] *Podolak J., Shilane P., Giesen J., Gross M., Guibas L.* Example-based 3D scan completion // Proc. Symposium on Geometry Processing, 2005. — Pp. 23–32.

[2] *Martinet A., Soler C., Holzschuch N., Sillion F.* Accurate detection of symmetries in 3D shapes. // ACM Trans. Graph, 2006. — Vol. 25, No. 2. — Pp. 439–464.

[3] *Mitra N. J., Guibas L. J., Pauly M.* Partial and approximate symmetry detection for 3D geometry // ACM Trans. Graph, 2006. — Vol. 25, No. 3. — Pp. 560–568.

[4] *Thrun S., Wegbreit B.* Shape from symmetry // Int. Conference on Computer Vision, 2005.

[5] *Simari P., Kalogerakis E., Singh K.* Folding meshes: Hierarchical mesh segmentation based on planar symmetry // Proc. Symposium on Geometry Processing, 2006.

[6] *Каркищенко А. Н., Гречухин И. А.* Статистическое распознавание лиц по геометрии характерных точек для систем транспортной безопасности // Управление большими системами. Сборник трудов. Выпуск 38. — М.: ИПУ РАН, 2012. — С. 78–90.

[7] *Каркищенко А. Н., Гречухин И. А.* Локализация характерных точек на основе естественной симметрии изображений. // Труды первой научно-технической конференции «Интеллектуальные системы управления на железнодорожном транспорте». — М.: НИИАС, 2012. — С. 261–265.

[8] *Стренг Г.* Линейная алгебра и её применения. — М.: Мир, 1980. — 454 с.

[9] *Маркус М., Минк Х.* Обзор по теории матриц и матричных неравенств. — М.: Наука, 1972. — 232 с.

Использование правил со сложной структурой для коррекции документов в формате LaTeX

Чувилин К. В.
kirill.chuvilin@gmail.com
Москва, Московский физико-технический институт (ГУ)

Рассматривается задача автоматического синтеза правил коррекции документов в формате LaTeX. Каждый документ представляется в виде синтаксического дерева. Отображения вершин деревьев черновых документов в вершины деревьев чистовых составляют обучающую выборку, по которой синтезируются правила замены. В первую очередь строятся простые правила, реализующие операции удаления, добавления или изменения одной вершины синтаксического дерева и использующие линейные последовательности вершин для выбора позиции применения. Построенные правила объединяются в группы на основе позиций применимости и оценок качества. Исследуются правила, использующие древовидные структуры вершин для выбора позиции применения. Анализируется изменение качества правил при последовательном наращивании обучающей выборки.

Ключевые слова: *автоматизация, анализ текста, лексема, машинное обучение, метрика, редактирующее расстояние, обучение с подкреплением, синтаксическое дерево, токен, LaTeX.*

The use of rules with complex structure for LaTeX documents correction

Chuvilin K. V.
Moscow Institute of Physics and Technology (SU)

The problem of automatic synthesis of LaTeX documents editing rules is investigated. Each document is represented as a parse tree. Tree nodes mappings of initial documents to edited documents form the training set, which is used to generate the rules. Simple rules that implement removal, insertion, or replacing operations of single node and use linear sequence of vertices to select a position are synthesized primarily. The constructed rules are grouped based on the positions of applicability and quality. The rules that use tree-like structure of nodes to select the position are studied. The changes in the quality of the rules during the sequental increase of training dcocuments set are analized.

Keywords: *automation, editing distance, LaTeX, lexeme, machine learning, metric, parse tree, reinforcement learning, syntax tree, text analysis, token.*

Введение

Поводом для работы послужила подготовка сборников трудов конференций ММРО и ИОИ в 2007–2012 годах. Многие конференции и издательства принимают материалы от авторов в формате LaTeX. В каждом издательстве есть определенные традиции и требования к оформлению публикуемого материала. К ним относятся оформление заголовков, списков, таблиц, библиографии, формул, чисел, и многое другое. Ошибки, связанные с несоблюдением этих правил, называются *типографическими*. Обычно авторские тексты содержат значительное количество (десятки на страницу) таких ошибок, исправление которых производится корректорами вручную. Существуют инструменты для облегчения

процесса ручной корректуры [1], но тем не менее обработка одной страницы занимает до двух часов времени. Поэтому актуальна автоматизация процесса исправления типографических ошибок для сокращения времени и объема ручной работы.

Вообще говоря, идея автоматизации коррекции текстов не нова [2], и на данный момент существуют качественные инструменты для автоматического поиска и исправления орфографических ошибок [3], использующие словари и морфологический анализ словоформ текста. Кроме того, схожая проблема возникает для интеллектуальной коррекции ошибок в запросах поиска [4], с помощью лексических и статистических признаков. Но подобные подходы не применимы для исправления типографических ошибок, рассматриваемых в данной работе, которые связаны не только с текстовым содержанием документа, но и разметкой форматирования, и зачастую для описания ошибки не достаточно локальной информации в тексте, но также требуется знание контекста, дополнительной информации о позиции в структуре документа.

С другой стороны, существует область исследований, посвященная улучшению характеристик исходного кода программ (вероятности возникновения ошибок в отдельных модулях, степени связности модулей и др.). Известны методы [5, 6], позволяющие оценивать характеристики, основываясь на анализе истории изменений репозиториев, и использовать их для поиска ошибок в коде. Они позволяют создавать рекомендательные системы [7] для улучшения качества кода программы при редактировании. Документы в формате LaTeX можно рассматривать как исходный код, который используется компилятором TeX, но в издательской практике не распространено использование репозиториев, пригодных для последующего анализа, нет единых стандартов и, кроме того, текстовое содержимое документов не может быть подвержено подобной обработке.

Таким образом, возникает необходимость нового исследования, направленного непосредственно на автоматизацию процесса исправления типографических ошибок. Предлагаемый подход заключается в следующем. Корректор работает с системой, которая использует формально описанные правила коррекции, определяет в исходном тексте возможные места исправлений и предлагает ему вариант замены. Если он согласен с заменой, ему остается только нажать на соответствующую кнопку. Если не согласен, то он делает правку самостоятельно.

Правила можно задавать вручную, непосредственно на основе практического опыта корректоров. Однако ввиду значительного числа и разнообразия ситуаций и вариантов поведения, это приведет скорее к увеличению трудозатрат, особенно на начальном этапе. Поэтому предлагается строить правила, используя корпус уже обработанных пар документов. Документы, не прошедшие корректуру, будем называть *черновиками*, прошедшие — *чистовиками*.

Задача автоматического синтеза правил преобразования документов по обучающей выборке, составленной из пар «черновик–чистовик» рассматривалась в работе [8]. Однако построенные предложенным образом правила обладали рядом недостатков, среди которых: неполное покрытие мест, соответствующих ошибкам, в документах, не входящих в обучающую выборку, и неверные предлагаемые варианты замены.

В данной работе рассматриваются два подхода к повышению качества набора правил с помощью усложнения структуры.

Выделение различий между документами

Документы формата LaTeX в своем текстовом виде представляются как последовательности (допускается вложение) *лексем* следующих типов: синтаксические скобки ({ и }),

пробел, горизонтальный отступ, вертикальный отступ, разделитель абзацев, обрыв строки, символ, цифра, буква слова, команда, тэг, обертка (тэг с замыканием), формула, верхний или нижний индекс, метка, линейное измерение, путь к файлу или папке, список, элемент списка, плавающий бокс, изображение, бинарный математический оператор, математический постоператор, математический преоператор, таблица, конец ячейки таблицы, параметры таблицы, не обрабатываемые данные.

Кроме того, такие файлы обладают естественной древовидной структурой (*синтаксическим деревом*) [9], исследуя которую, можно получить всю необходимую информацию для описания корректорской правки. Узлы этой структуры будем называть *токенами*. При построении синтаксического дерева выделяются следующие виды токенов: символ (буква, цифра, знак математической операции, кавычка, тире и т. п.), команда, параметр команды, окружение, тело окружения, пробел, разделитель абзацев, слово, число, метка, линейный размер, путь к файлу, параметры формата таблицы, не обрабатываемый фрагмент (например, текст внутри окружения `verbatim`). Каждому токену может соответствовать один из типов лексем.

Синтаксическое дерево с точностью компилятора взаимно однозначно определяет документ LaTeX. Правила замены удобно формулировать именно для деревьев.

Будем называть синтаксические деревья чистовиков *чистовыми*, а черновиков — *черновыми*.

Перед выявлением закономерностей в правках корректоров выделяются различия между черновым и чистовым деревьями для каждой пары документов из обучающей выборки. Для этого используется алгоритм Zhang-Shasha [10], который подразумевает, что к синтаксическому дереву разрешается последовательно применять следующие операции: удаление токена (все его потомки переходят родителю), вставка нового токена в произвольное место, изменение токена. Алгоритм Zhang-Shasha позволяет вычислять редактирующее расстояние между двумя деревьями и, кроме того, определять, какую операцию нужно применить к каждой вершине для реализации такого расстояния. В качестве результата его работы получаются пары (прообраз и образ) не измененных токенов, пары (прообраз и образ) измененных токенов, множество удаленных токенов, множество добавленных токенов.

Правила коррекции с простой структурой

После получения отображения черновых и чистовых деревьев строится начальный набор правил коррекции [8]. Каждое построенное правило характеризуется *шаблоном* (последовательностью соседних токенов с общим родителем), *локализатором* (токеном, к потомкам которого применяется шаблон) и *действием* (операцией, направленной на изменение синтаксического дерева).

Определение 1. *Левая (правая) шаблонная цепочка радиуса r — это последовательность соседних токенов с общим родителем, длиной не больше r. Началом цепочки считается самый правый (левый) ее токен.*

Пусть токен x чернового дерева удален или изменен на токен y. Тогда локализатор — родительский токен x, шаблон составляется из левой и правой шаблонных цепочек, наиболее близких к x и самого токена x. В таких случаях токен x будем называть *целевым токеном правила*. Действие правила заключается в удалении целевого токена или изменении его на токен y, в зависимости от типа правила.

Пусть в чистовое дерево добавлен токен y. Тогда локализатор — прообраз родительского токена y, если он существует; шаблон составляется из левой шаблонной цепочки,

начинающейся в прообразе левого соседа y, если он существует, и аналогичной правой. Действие правила заключается в добавлении токена y между левой и правой шаблонными цепочками.

Пример 1 (Правило с простой структурой). Одной из самых распространенных типографических ошибок является использование кавычек "..." вместо «...» в русском тексте. Пусть соответствующие фрагменты чернового и чистового синтаксических деревьев выглядят так, как показано на рис. 1, что соответствует преобразованию фрагмента "**слово** в <<**слово** внутри окружения `document`.

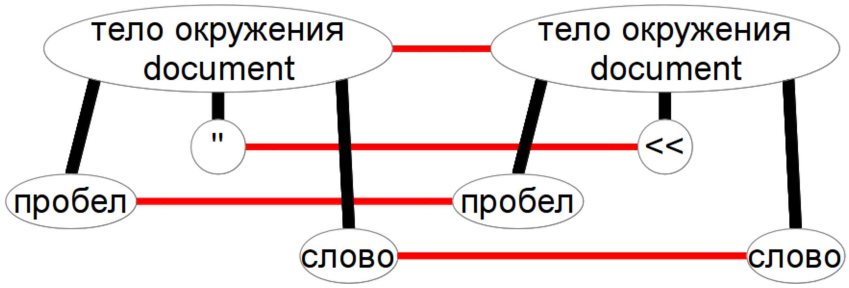


Рис. 1: Соответствующие фрагменты синтаксических деревьев

Тогда локализатором правила является токен тела окружения `document`, левая шаблонная цепочка состоит из токена пробела, правая — из токена слова, шаблон — из токена проблела, токена символа " и токена слова. Действие правила заключается в изменении токена, находящегося между токенами пробела и слова на токен символа <<.

Аналогичное правило строится для изменения правой кавычки.

Поиск соответствий правилам

Считается, что токен l дерева соответствует локализатору правила, если выполняется совпадение типов токенов и типов их лексем.

Среди потомков l ищется непрерывная последовательность, совпадающая с шаблоном по следующим правилам:

— для всех токенов шаблонных цепочек должно выполняться совпадение типов и лексем с соответствующими потомками l,
— для целевого токена должно выполняться полное совпадение с соответствующим потомком l.

Определение 2. *Позиция правила в синтаксическом дереве документа LaTeX — это совокупность токена, соответствующего локализатору правила, и набора токенов, соответствующих шаблону. Порождающая позиция правила — позиция, которая соответствует элементу отображения синтаксических деревьев, из которого было синтезировано правило. Множество позиций или позиции правила на множестве документов — совокупность всех позиций в синтаксических деревьях этих документов, удовлетворяющих правилу.*

Предварительная оценка правила

Для предварительной оценки качества каждого правила вычисляются данные по обучающей выборке [11]. Обозначим: d_t — количество позиций правила на множестве черновиков, c_t — количество позиций правила на множестве чистовиков.

Определение 3 (Предварительная точность правила). *Предварительная (на обучающей выборке) точность правила — это отношение количества позиций, которые соответствуют только черновикам, к общему числу найденных позиций:* $\frac{d_t - c_t}{d_t}$.

Выбор оптимальных шаблонов

Набор токенов образующих шаблон правила можно задавать по-разному. Из результатов экспериментов [12] можно сделать вывод, что максимальные шаблоны не всегда дают лучший результат. В данной работе оптимальный шаблон выбирается по следующим критериям:

1. предварительная точность правила не должна быть меньше 0.9,
2. выбирается наименьший размер шаблона, позволяющий построить правило с допустимой точностью,
3. выбирается правило с наибольшей точностью из всех, обладающих шаблонами выбранного размера.

Редукция набора правил

После выявления закономерностей по всем различиям между деревьями обучающей выборки оказывается, что правил избыточно, поскольку многие дублируются. Для устранения такого эффекта используется процесс редукции, который заключается в удалении некоторых правил так, чтобы общий набор выделенных закономерностей не менялся.

Определение 4. *Правило A поглощает правило B, если операции, соответствующие правилам, совпадают и множество позиций правила B на черновых документах обучающей выборки является подмножеством позиций правила A.*

Из построенного набора правил пошагово удаляются те, которые могут быть поглощены другими.

Определение 5. *Множество порождающих позиций или порождающие позиции правила A — совокупность порождающей позиции правила A и множества порождающих позиций правил, которые были поглощены правилом A.*

Оценки качества правил

Для оценки качества набора правил был проведен эксперимент, в котором использовалось 85 пар черновых и чистовых статей конференции ИОИ-8. Моделировалось адаптивное обучение набора правил. Для этого обучающее множество пар документов, используемое для построения правил, постепенно увеличивалось: 2, 3, 4, 6, 9, 13, 19, 28, 42, 63. Обозначим $S_1 \subset \ldots \subset S_{10}$ — полученные десять обучающих множеств пар документов, S_{11} — множество всех пар документов. На каждом шаге контрольное множество формировалось из пар документов, которые добавлялись к обучающему множеству на следующем шаге: $S_{i+1} \setminus S_i,\ \ i = 1, \ldots, 10$.

Вычислялись оценки качества для синтезированных правил и наборов правил [11]. Последовательности множеств строились 50 раз, данные по всем построениям были усреднены.

Обозначим: d_t — количество позиций правила на множестве черновиков, c_t — количество позиций правила на множестве чистовиков.

Определение 6. *Предварительная (на обучающей выборке) точность правила — это отношение количества позиций, которые соответствуют только черновикам, к общему числу найденных позиций:* $\frac{d_t - c_t}{d_t}$.

Определение 7. *Пусть $P(A_1), \ldots, P(A_k)$ — предварительные точности правил $A_1, \ldots, A_k$ соответственно, а их позиции правил таковы, что соответствуют изменению одного и того же токена. Весом правила A_i называется* $W(A_i) = \frac{P(A_i)}{\sum_{j=1}^{k} P(A_j)}$.

Определение 8. *Пусть* $E(A_i)$ *— число, равное* 0*, если правило* A_i *соответствует верной правке, и* 1 *в противном случае. Тогда выражение*

$$\sum_{i=1}^{k} W(A_i)E(A_i) = \frac{\sum_{i=1}^{k} P(A_i)E(A_i)}{\sum_{i=1}^{k} P(A_i)}$$

задает среднюю ошибку набора правил на выбранном токене.

Обозначим: E_t и E_c — суммы средних ошибок набора правил на всех токенах черновых деревьев обучающей и контрольной выборок соответственно, N_t и N_c — количества различных позиций всех правил набора на множествах черновиков обучающей и контрольной выборок соответственно, D_t и D_c — суммы редактирующих расстояний для всех пар черновых и чистовых деревьев обучающей и контрольной выборок соответственно.

Поскольку правила синтезируются только при добавлении, удалении или изменении токена, а сумма таких операций для двух деревьев равна редактирующему расстоянию, будут корректны следующие определения [11].

Определение 9. $\frac{N_t-E_t}{N_t}$ *— предварительная (на обучающей выборке) точность набора правил.* $\frac{N_c-E_c}{N_c}$ *— контрольная (на контрольной выборке) точность набора правил.*

Определение 10. $\frac{N_t-E_t}{D_t}$ *— предварительная (на обучающей выборке) полнота набора правил.* $\frac{N_c-E_c}{D_c}$ *— контрольная (на контрольной выборке) полнота набора правил.*

Результаты проведенных расчетов для наборов правил с простой структурой представлены на рис. 2. Кривые, соответствующие предварительным и контрольным оценкам точности и полноты набора правил, расположены довольно близко друг другу. Это означает, что синтезированные предложенным способом правила обладают неплохой обобщающей способностью.

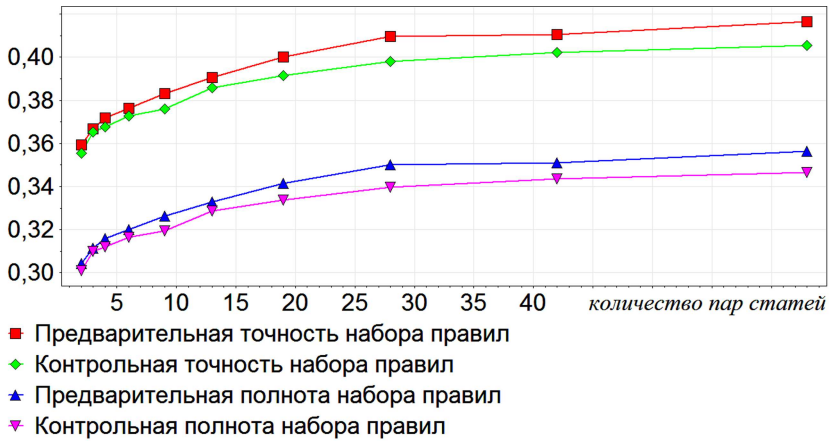


Рис. 2: Оценки точности и полноты набора правил с простой структурой

С другой стороны, и точность, и полнота наборов правил не превосходят 50%. Для точности это означает, что существуют различные правила со схожими шаблонами. Недостаток полноты можно объяснить тем, что рассмотренных типов правил недостаточно для описания действий корректора.

Групповые правила

На практике встречаются случаи, когда корректор изменяет, удаляет или добавляет более одного токена. Например, перенос одного токена на другую позицию представляет

собой совокупность удаления и добавления токена. Для увеличения спектра обрабатываемых правок корректора мы будем использовать группировку правил.

Пусть для двух правил существуют позиции такие, что:

— токены, соответствующие локализаторам, совпадают;
— наборы токенов, соответствующих шаблонам, имеют общие элементы.

Тогда построим новое *групповое правило*, локализатор которого совпадает с локализатором рассматриваемых правил, а шаблон образуется объединением их шаблонов. Построенное правило добавляется в набор, если его предварительная точность выше, чем предварительная точность каждого из рассматриваемых правил.

Пример 2 (Групповое правило). Стандарты оформления знаков тире могут различаться для издательств. Это приводит, например, к необходимости изменения `~---` на `’’---`. Первый фрагмент состоит из токена неразрывного пробела (~) и символа тире (---), второй — из токена символа неотделимого тире (’’---). Каждое правило с простой структурой может изменить только один токен, поэтому такая замена ими не реализуется. С другой стороны, совокупность удаления токена неразрывного пробела, стоящего перед токеном тире, и замены токена тире на токен неотрывного тире дает нужную замену.

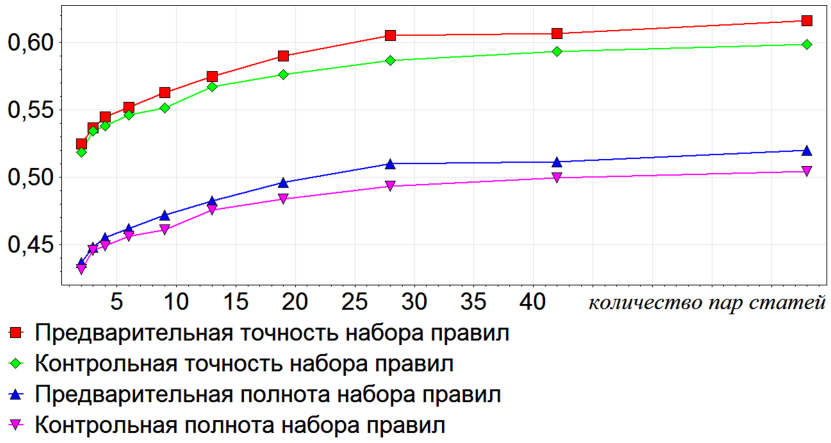


Рис. 3: Оценки точности и полноты набора правил с учетом группировки

На рис. 3 показаны оценки качества набора правил с учетом групповых правил, построенные в соответствии с экспериментом, описанным выше. Можно видеть, что такой подход позволил получить точность заметно больше половины, но полнота все еще находится на уровне 50%.

Правила с древовидными шаблонами

Еще один способ повышения точности и полноты набора правил заключается в усложнении структуры шаблона. Шаблон правила с простой структурой позволяет использовать только соседние токены для определения позиции, что, вообще говоря, не означает использование всего текста, соответствующего этим токенам, поскольку не учитываются структура и содержимое поддеревьев, корни которых образуют шаблон.

Шаблон *древовидного правила* будем строить из двух *шаблонных деревьев: левого и правого*. В этом случае *длина шаблона* — количество токенов в этих деревьях.

Синтез таких правил, выбор оптимальных шаблонов и редукция происходят так же, как и правил с простой структурой, а поиск мест применимости осуществляется следующим образом. Считается, что токен l дерева соответствует локализатору правила, если

выполняется совпадение типов токенов и типов их лексем. Шаблонные деревья и все их поддеревья проверяются на применимость с помощью проверки на каждом уровне, начиная с потомков токена l, условий:

— совпадение самых правых (для левых поддеревьев) или левых (для правых поддеревьев) токенов-потомков с токенами шаблона,
— применимость соответствующего поддерева для каждого токена-потомка.

Пример 3 (Правило с древовидным шаблоном). Еще одной из самых распространенных типографических ошибок является включение знака препинания в тело формулы. Пусть соответствующие фрагменты чернового и чистового синтаксических деревьев выглядят так, как показано на рис. 4, что соответствует преобразованию фрагмента `$ ,$` в `$ $,` внутри окружения `document`.

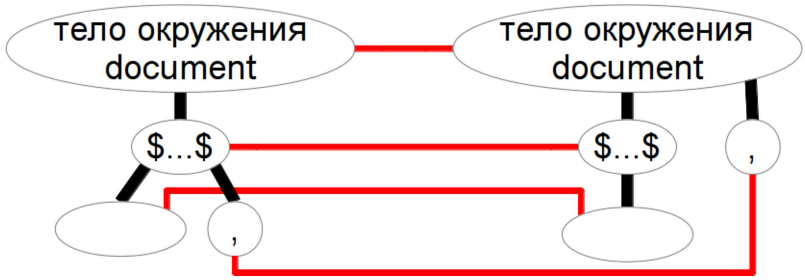


Рис. 4: Соответствующие фрагменты синтаксических деревьев

Такую замену можно описать как совокупность добавления токена запятой после токена формулы, содержащей в конце запятую, и удаления токена запятой из токена формулы. Но для определения наличия запятой внутри формулы требуется обратиться к токенам внутри нее, поэтому шаблона правила с простой структурой будет не достаточно.

Поэтому правило добавления запятой описывается древовидным шаблоном, левое поддерево которого состоит из из токена формулы, содержащего токен запятой, а правое поддерево пустое. Действие заключается в добавлении токена запятой после левого поддерева.

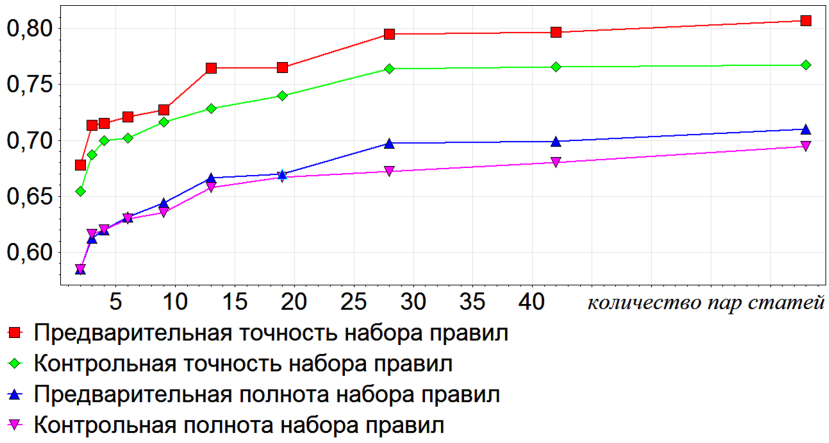


Рис. 5: Оценки точности и полноты набора правил с учетом древовидных правил

На рис. 5 показаны оценки качества набора правил с учетом синтеза древовидных правил, построенные в соответствии с экспериментом, описанным выше. Можно видеть,

что подобный подход позволил значительно увеличить точность синтезированного набора правил.

Заключение

Исследование направлено на улучшение результатов, полученных в работах [8, 9, 11, 12]. Были предложены два новых подхода: группировка правил и древовидные шаблоны. Каждый из подходов позволил заметно улучшить точность и полноту синтезируемого набора правил. Важным является преодоление уровня 50% для точности.

Литература

[1] *André J., Richy H.* Paper-less editing and proofreading of electronic documents. — 1999. — `http://www.irisa.fr/imadoc/articles/1999/heidelberg.pdf`.

[2] *Большаков И. А.* Проблемы автоматической коррекции текстов на флективных языках // *Итоги науки и техн. Сер. Теор. вероятн. Мат. стат. Теор. кибернет.* 1988. Т. 28. С. 111–139.

[3] `http://extensions.services.openoffice.org/project/lightproof` — Lightproof grammar checker development framework — 2013.

[4] *Панина М. Ф., Байтин А. В., Галинская И. Е.* Автоматическое исправление опечаток в поисковых запросах без учета контекста // *Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог»*. 2013. Вып. 12. Т. 1. С. 556–567.

[5] *Williams C., Hollingsworth J.* Automatic Mining of Source Code Repositories to Improve Bug Finding Techniques // *IEEE Transactions on Software Engineering table of contents archive*. 2005. Vol. 31, no. 6. P. 466–480.

[6] *Князев Е. Г.* Методы обнаружения закономерностей эволюции программного кода // *Труды XIV Всероссийской научно-методической конференции «Телематика-2007»*. СПбГУ ИТМО. 2007. Т. 2. С. 435–436.

[7] *Madou F., Agüero M., Esperón G., López De Luise D.* Software for improving source code quality // *World Academy of Science, Engineering and Technology*. 2011. Vol. 59. P. 1259–1265.

[8] *Чувилин К. В.* Синтез правил коррекции документов в формате LaTeX с помощью сопоставления синтаксических деревьев // *Доклады 15-й Всероссийской конференции «Математические методы распознавания образов» ММРО-15*. 2012. С. 597–600.

[9] *Чувилин К. В.* Использование синтаксических деревьев для автоматизации коррекции документов в формате LaTeX // *Компьютерные исследования и моделирование*. 2012. Т. 4, no 4. С. 871–883.

[10] *Zhang K., Shasha D.* Simple fast algorithms for the editing distance between trees and related problems // *SIAM Journal of Computing*. 1989. No. 18. P. 1245–1262.

[11] *Чувилин К. В.* Адаптивное обучение правил коррекции документов в формате LaTeX // *Доклады 9-й Международной конференции «Интеллектуализация обработки информации» ИОИ*. 2012. С. 652–655.

[12] *Чувилин К. В.* Автоматический синтез правил коррекции документов в формате LaTeX и их улучшение на основе статистической оценки качества // *Труды II Всероссийской научной конференции молодых ученых с международным участием «Теория и практика системного анализа» ТПСА*. 2012. С. 17–25.

Применение Машины Релевантных Объектов в задачах восстановления числовых зависимостей

Разин Н. А.[1]***, Черноусова Е. О.***[2]***, Красоткина О. В.***[3]***, Моттль В. В.***[4]
nrmanutd@gmail.com[1], elchernousova@inbox.ru[2], o.v.krasotkina@yandex.ru[3], vmottl@yandex.ru[4]
МФТИ[1,2]; ТулГу[3,4]; ВЦ РАН[3,4]

В работе рассматривается задача беспризнакового распознавания образов в предположении, что объекты попарно сравниваются при помощи произвольной действительной функции. Такой подход является гораздо более общим, чем традиционный метод потенциальных функций (кернелов), требующий положительной полуопределенности матрицы функции сравнения объектов. Последнее требование в большинстве случаев является чрезмерным, причем обучение еще более осложняется, если существует несколько различных способов сравнительного представления объектов. В таких случаях экспериментатор вынужден решать задачу исключения как избыточных базисных объектов для сравнительного представления объектов обучающей совокупности, так и способов сравнения. В терминах общего пространства попарного сравнительного представления объектов предлагаемая постановка становится математически аналогичной классической задаче отбора признаков. Получившийся выпуклый критерий обучения аналогичен методу релевантных векторов Типпинга, но является существенно более общим, поскольку содержит структурный параметр, контролирующий селективность отбора.

Elastic-Net Relevance Vector Machines for selective multimodal regression estimation

Razin N. A.[1]***, Chernousova E. O.***[2]***, Krasotkina O. V.***[3]***, Mottl V. V.***[4]
MIPT[1,2]; Tula State University[3,4]; Computing Centre of the Russian Academy of Sciences[3,4]

We address the problem of regression estimation under the assumption that pair-wise comparison of objects is arbitrarily scored by real numbers. Such a linear embedding is much more general than the traditional kernel-based approach, which demands positive semi-definiteness of the matrix of object comparisons. This demand is frequently prohibitive and is further complicated if there exist a large number of comparison functions, i.e., multiple modalities of object representation. In these cases, the experimenter typically also has the problem of eliminating redundant modalities and objects. In the context of the general pair-wise comparison space this problem becomes mathematically analogous to that of wrapper-based feature selection. The resulting convex training criterion based on the principle of Elastic Net regression estimation is analogous to Tipping's Regression Relevance Vector Machine, but essentially generalizes it via the presence of a structural parameter controlling the selectivity level.

Введение

Одна из современных задач информатики — это восстановление зависимостей по эмпирическим данным. Пусть задано множество объектов реального мира $\omega \in \Omega$, каждому из которых поставлено в соответствие значение ненаблюдаемой переменной $y \in \mathbb{Y}$. Предположим, что наблюдателю известны значения функции $y(\omega) : \Omega \to \mathbb{Y}$ только на объектах обучающей совокупности.

$$\Omega^* \Rightarrow \left\{\omega_j, y(\omega_j), j = 1, ..., N\right\} \subset \Omega. \tag{1}$$

Необходимо продолжить эту функцию на все множество объектов $\hat{y}(\omega) : \Omega \to \mathbb{Y}$ таким образом, чтобы было возможно оценить скрытую характеристику объектов $\omega \in \Omega \setminus \Omega^*$ [1]. В частности, если скрытая переменная принимает значение из множества действительных чисел $\hat{y}(\omega) : \Omega \to \mathbb{Y} = \mathbb{R}$, то такая задача известна как задача восстановления числовой зависимости (задача восстановления регрессии).

В простейшем случае предположим, что каждый объект представлен в компьютере вектором действительных чисел $\mathbf{x}(\omega) = \big(x_i(\omega), i \in \mathbb{I}\big) \in \mathbb{R}^n, \mathbb{I} = \{1, ..., n\}$, а числовая зависимость рассматривается как линейное решающее правило:

$$\hat{y}(\omega) = \hat{y}\big(\mathbf{x}(\omega)\big) = \sum_{i \in \mathbb{I}} a_i x_i(\omega) + b, \quad a_i, b \in \mathbb{R}. \tag{2}$$

То, что решающее правило (2) линейное, не является ограничивающим фактором, т.к. всегда можно выбрать такой способ описания объектов и задания их признаков $x_i(\omega)$, чтобы они погрузились ровно в такое линейное пространство, которое необходимо.

В качестве альтернативы к признаковому и кернельному подходам Дьюин и его соавторы предложили в [2] беспризнаковый подход, в котором объекты представлены произвольной функцией попарного сравнения $S(\omega', \omega'') : \Omega \times \Omega \to \mathbb{R}$. Идея заключалась в использовании значения этой функции между конкретным объектом $\omega \in \Omega$ и всеми остальными объектами обучающей совокупности $\{\omega_j, j = 1, ..., N\}$ как вектора вторичных признаков $\mathbf{x}(\omega) = \big(x_j(\omega) = S(\omega_j, \omega), j = 1, ..., N\big)$ и применении стандартных методов восстановления числовой зависимости (2), основанных на признаковом описании объекта в $\mathbb{R}^N$ с $\mathbb{J} = \{j = 1, ..., N\}$ вместо $\mathbb{I} = \{i = 1, ..., n\}$.

Позже Бишоп и Типпинг предложили «машину» (метод) Релевантных Векторов (Relevance Vector Machine – RVM) [3], где основной идеей был отбор только небольшого количества наиболее информативных релевантных объектов из обучающей совокупности. Авторы называли их релевантными векторами, поскольку $S(\omega', \omega'')$ рассматривалась как кернел, погружающий объекты в линейное пространство. В исходном методе RVM селективность релевантных векторов намеренно была принята очень высокой, чтобы придать им аналогию с опорными векторами в модели SVM. К тому же, метод RVM основан на байесовском принципе отбора, поэтому приводит к невыпуклой задаче обучения.

В данной работе рассматривается более общая ситуация, типичная для приложений, когда заданы несколько модальностей попарного представления объектов $S_k(\omega', \omega'')$ в виде различных функций парного сравнения $k \in \mathbb{K} = \{1, ..., m\}$, то количество вторичных признаков каждого объекта $n = mN$ превышает объем обучающей совокупности N:

$$\mathbf{x}(\omega) = \big(x_i(\omega) = x_{kj}(\omega) = S_k(\omega_j, \omega),\ k \in \mathbb{K},\ j \in \mathbb{J},\ i \in \mathbb{I} = \mathbb{K} \times \mathbb{J}\big). \tag{3}$$

Таким образом, чтобы избежать эффекта переобучения, необходимо выполнить отбор наиболее информативных вторичных признаков $(kj) \in \hat{\mathbb{I}} \subset \mathbb{I}$ и исключить остальные $(kj) \notin \hat{\mathbb{I}}$.

Мы рассмотрим класс задач обучения, которые, являясь выпуклыми в отличие от классической машины RVM [3], позволяют отбирать не только релевантные объекты $\hat{\mathbb{J}} \subset \mathbb{J}$, но и релевантные модальности их попарного представления $\hat{\mathbb{K}} \subset \mathbb{K}$, а также контролировать степень отбора с помощью специального параметра селективности $\mu \geqslant 0$.

Пусть дана обучающая совокупность $\big\{(\tilde{\mathbf{x}}_j, \tilde{y}_j), j = 1, ..., N\big\}$. Задача восстановления регрессии $\tilde{y}_j \in \mathbb{R}$ с избыточным количеством признаков $\mathbb{I}$ (вторичных признаков в случае задачи восстановления регрессии через RVM $\mathbb{I} = \mathbb{K} \times \mathbb{J}$) может быть записана как задача восстановления гребневой регрессии, т.е. взвешенной суммы квадратов ошибок отклонения $\sum_{j=1}^{N} \big[\tilde{y}_j - \big(\sum_{i \in \mathbb{I}} \tilde{a}_i \tilde{x}_{ij} + \tilde{b}\big)\big]^2 \to \min(\tilde{a}_i, i \in \mathbb{I}, \tilde{b})$ и штрафа на сумму квадратов

коэффициентов $\beta\sum_{i\in\mathbb{I}}\tilde{a}_i^2 \to \min(\tilde{a}_i, i\in\mathbb{I})$, где β — коэффициент, взвешивающий эти два слагаемых между собой (trade-off coefficient). Очевидно, что подобный подход имеет смысл только если исходная обучающая совокупность $\big\{(\tilde{\mathbf{x}}_j, \tilde{y}_j), j=1,...,N\big\}$ центрирована и нормирована:

$$\big\{(\mathbf{x}_j, y_j), j=1,...,N\big\},\ \mathbf{x}_j=\big(x_{ij}, i\in I\big)\in\mathbb{R}^n,\ y_j\in\mathbb{R}, \tag{4}$$

$$\frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j=\mathbf{0},\ \frac{1}{N}\sum_{j=1}^{N}y_j=0,\ \frac{1}{N}\sum_{j=1}^{N}x_{ij}^2=1,\ i\in\mathbb{I}. \tag{5}$$

В противном случае штраф на сумму квадратов β_i должен быть уникальным для каждого признака $i\in\mathbb{I}$, чтобы скомпенсировать разницу в их масштабах. Если $\big\{(\tilde{\mathbf{x}}_j, \tilde{y}_j), j=1,...,N\big\}$ исходная обучающая совокупность, то нормализация может быть выполнена следующим образом:

$$x_{ij}=\frac{\tilde{x}_{ij}-\bar{x}_i}{\tilde{c}_i},\ y_j=\tilde{y}_j-\bar{y},\ \bar{x}_i=\frac{1}{N}\sum_{j=1}^{N}\tilde{x}_{ij},\ \bar{y}=\frac{1}{N}\sum_{j=1}^{N}\tilde{y}_j,\ \tilde{c}_i=\sqrt{\frac{1}{N}\sum_{j=1}^{N}(\tilde{x}_{ij}-\bar{x}_i)^2}. \tag{6}$$

Обратно, когда модель регрессии $\hat{y}(\mathbf{x})=\hat{\mathbf{a}}^T\mathbf{x}$ построена по нормализованной обучающей совокупности, она должна быть преобразована таким образом, чтобы корректно работать с произвольным новым объектом $\omega\in\Omega$, подаваемым на вход и описанным оригинальными признаками: $\tilde{\mathbf{x}}(\omega)=\big(\tilde{x}_i(\omega), i\in\mathbb{I}\big)$:

$$\hat{\tilde{y}}(\tilde{\mathbf{x}})=\sum_{i=1}^{m}\tilde{\hat{a}}_i\tilde{x}_i+\tilde{b},\ \tilde{\hat{a}}_i=\frac{\hat{a}_i}{\tilde{c}_i},\ \tilde{b}=\bar{y}-\sum_{i=1}^{m}\frac{\hat{a}_i}{\tilde{c}_i}\bar{x}_i. \tag{7}$$

Таким образом, предполагая, что обучающая совокупность нормирована и центрирована (4)-(5), задача восстановления гребневой регрессии записывается в следующей форме:

$$F_{RR}(\mathbf{a}|\beta,\hat{\mathbb{I}})=\beta\sum_{i\in\hat{\mathbb{I}}}a_i^2+\sum_{j=1}^{N}\Big(y_j-\sum_{i\in\hat{\mathbb{I}}}a_ix_{ij}\Big)^2\to\min(a_i, i\in\hat{\mathbb{I}}). \tag{8}$$

Критерий гребневой регрессии содержит два структурных параметра – коэффициент β и подмножество активных признаков $\hat{\mathbb{I}}$. Выбор первого из них не является большой проблемой, т.к. достаточно взять малое значение коэффициента $\beta>0$ с той лишь целью, чтобы гарантировать строгую выпуклость критерия в случае коллинеарности векторов признаков из $\hat{\mathbb{I}}$. Но что касается выбора $\hat{\mathbb{I}}$, то это очень сложная задача.

Если задан конкретный набор признаков, то, воспользовавшись следствием из формулы Шермана-Вудбери-Моррисона [5], можно применить эффективный алгоритм [4] для вычисления ошибки leave-one-out критерия обучения гребневой регрессии (8). Тем не менее, это невозможно сделать для всех 2^n наборов признаков $\hat{\mathbb{I}}\subseteq\mathbb{I}$. В нашем случае, в соответствии с (3), должно быть найдено лучшее подмножество признаков $i=(kj)\in\mathbb{I}=\mathbb{K}\times\mathbb{J}$, количество которых $n=mN$ — произведение чила модальностей $m=|\mathbb{K}|$ и числа объектов в обучающей совокупности $N=|\mathbb{J}|$.

В данной работе разработана Машина Релевантных Объектов (Relevance Object Machine) с возможностью отбора модальностей для задачи восстановления регрессии. В основу предлагаемого метода мы положили критерий обучения Elastic Net, представленный Zou и Hastie в [6] как результат комбинирования регуляризующего слагаемого гребневой регрессии $\beta\sum_{i\in\mathbb{I}}a_i^2\to\min$ и Лассо Тибширани: $\mu\sum_{i\in\mathbb{I}}|a_i|\to\min$ [7]:

$$F_{EN}(\mathbf{a}|\beta,\mu)=\beta\sum_{i\in\mathbb{I}}a_i^2+\mu\sum_{i\in\mathbb{I}}|a_i|+\sum_{j=1}^{N}\Big(y_j-\sum_{i\in\mathbb{I}}a_ix_{ij}\Big)^2\to\min(a_i,i\in\mathbb{I}). \tag{9}$$

В отличие от гребневой регрессии (8), суммирование здесь выполняется по всем возможным признакам из $i\in\mathbb{I}$, т.е. вторичным признакам $(kj)\in\mathbb{I}=\mathbb{K}\times\mathbb{J}$ (3), и, следовательно, искомый набор признаков $\hat{\mathbb{I}}\subset\mathbb{I}$ не будет найден при помощи стандартного Elastic Net.

Наличие регуляризующего слагаемого Lasso наделяет данный критерий возможностью строго обнулять избыточные признаки и, тем самым, находить набор информативных признаков $\hat{\mathbb{I}}_{\beta,\mu}=\big\{i:\hat{a}_{i,\beta,\mu}\neq 0\big\}\subset\mathbb{I}$, не выполняя полного перебора. Степень отбора признаков регулируется параметром μ.

Когда коэффициент гребневой регрессии положителен, т.е. $\beta>0$, критерий обучения Elastic Net (9) является строго выпуклым, однако более не является квадратичным, в отличие от гребневой регрессии (8). Тем не менее, минимизировать его легко.

Одним из наиболее удобных преимуществ Elastic Net является простота, с которой можно реализовать алгоритм регуляризации [6]. Такой алгоритм позволяет в один проход найти последовательность наборов признаков, охватывающих все возможные значения параметра селективности, от тех, при которых все признаки исключены из решающего правила, и до тех, при которых все включены. Тем не менее, выбор наиболее подходящего уровня селективности остается задачей внешних механизмов, например, кросс-валидации.

В данной работе предлагается встроить алгоритм регуляризации в алгоритм Машины Релевантных Векторов Регрессии для поиска последовательности наборов вторичных признаков $\hat{\mathbb{I}}_l=\hat{\mathbb{K}}_l\times\hat{\mathbb{J}}_l=\big\{i=(kj),k\in\hat{\mathbb{K}}_l,j\in\hat{\mathbb{J}}_l\big\}$, $l=|\hat{\mathbb{J}}_l\times\hat{\mathbb{K}}_l|$, начиная от единственного признака $l=1$, состоящего из одного объекта и одной модальности, $x_1(\omega)=x_{k_1j_1}(\omega)=S_{k_1}(\omega_{j_1},\omega)$, $\hat{\mathbb{K}}_1=\{k_1\}$, $\hat{\mathbb{J}}_1=\{j_1\}$, и заканчивая полным Декартовым произведением $\big\{x_i(\omega)=x_{kj}(\omega)=S_k(\omega_j,\omega),(kj)\in\mathbb{I}=\mathbb{K}\times\mathbb{J}\big\}$, $l=n=mN=|\mathbb{K}\times\mathbb{J}|$.

Как только найдены все подмножества-кандидаты вторичных признаков $\hat{\mathbb{I}}_l\subset\mathbb{I}$, $l=1,2,3,...,n$, остается лишь выбрать лучшее из них с точки зрения обобщающей способности. Больше не нужен регуляризационный член Лассо $\mu\sum_{i\in\mathbb{I}}|a_i|\to\min$, и мы возвращаемся к обычной модели гребневой регрессии (8) с целью использования ее квадратичной формы для того, чтобы вычислить ее ошибку leave-one-out, применив эффективный беспереборный алгоритм, описанный в [4].

Статья начинается разделом 1, в котором математически тщательно описывается главное свойство критерия Elastic Net в точке минимума (9) – разбивать множество признаков на два подмножества: активные и неактивные. Активные, в свою очередь, разбиваются на положительные и отрицательные коэффициенты регрессии Elastic Net. Этот факт становится особенно очевидным в сравнении со стандартными прямыми попытками обоснования в [6], если выписать двойственную задачу для задачи оптимизации Elastic Net

В разделе 2 предложен общий итеративный алгоритм для решения двойственной задачи при фиксированных параметрах регуляризации и поиска подмножества активных коэффициентов регрессии.

Далее, в разделе 3, излагается алгоритм регуляризации, а именно, алгоритм, решающий задачу Elastic Net в один проход для всех возможных значений параметра селективности.

Раздел 4 посвящен быстрому способу вычисления ошибки leave-one-out, представленному в [4], но адаптированному к двойственной задаче исходного критерия обучения.

Наконец, в разделе 5, приводятся результаты экспериментов на тестовых и реальных данных.

В статье пропущены элементарные доказательства некоторых математических предложений.

Двойственная задача для критерия Elastic Net и разбиение множества признаков

Обобщение критерия обучения Elastic Net (9) превращает его в многомодальную Машину Релевантных Объектов, которая отличается от исходной более сложным смыслом вторичных признаков $i = (k, l)$ и их количеством $|\mathbb{I}| = mN$. Для того, чтобы явно увидеть механизм отбора признаков, удобно записать критерий Elastic Net в его наглядной форме – выпуклой оптимизационной задачи с условиями-равенствами:

$$\begin{cases} \beta \sum\limits_{i\in\mathbb{I}} a_i^2 + \mu \sum\limits_{i\in\mathbb{I}} |a_i| + \sum\limits_{j=1}^{N} \delta_j^2 \to \min\big(a_i \in \mathbb{R}, i \in \mathbb{I}; \delta_j \in \mathbb{R}, j = 1, ..., N\big), \\ \delta_j = y_j - \sum\limits_{i\in\mathbb{I}} a_i x_{ij},\ j = 1, ..., N. \end{cases} \tag{10}$$

Утверждение 1. *Решение выпуклой задачи Elastic Net с ограничениями* (10) *выражается через решение двойственной выпуклой задачи без ограничений, т.е. через коэффициенты* $\boldsymbol{\delta} = (\delta_1, ..., \delta_N) \in \mathbb{R}^N$*, которые играют роль множителей Лагранжа:*

$$\begin{gathered} W(\boldsymbol{\delta}|\beta,\mu) = \frac{1}{\beta}\sum_{i\in\mathbb{I}}\Big\{\min\Big[\frac{\mu}{2} + \bar{\mathbf{x}}_i^T\boldsymbol{\delta},\ 0,\ \frac{\mu}{2} - \bar{\mathbf{x}}_i^T\boldsymbol{\delta}\Big]^2\Big\} + (\boldsymbol{\delta}-\mathbf{y})^T(\boldsymbol{\delta}-\mathbf{y}) = \\ \frac{1}{\beta}\sum_{i\in\mathbb{I}}\Big\{\min\Big[\frac{\mu}{2} + \sum_{j=1}^{N}\delta_j x_{ij},\ 0,\ \frac{\mu}{2} - \sum_{j=1}^{N}\delta_j x_{ij}\Big]^2\Big\} + \textstyle\sum_{j=1}^{N}(\delta_j - y_j)^2 \to \min, \end{gathered} \tag{11}$$

где $\bar{\mathbf{x}}_i = (x_{i,1}, ..., x_{i,N}) \in \mathbb{R}^N$ *— векторы соответствующих вторичных признаков в обучающей выборке* $i \in \mathbb{I} = \mathbb{K} \times \mathbb{J}$ (3).

Решение $\hat{\boldsymbol{\delta}}_{\beta\mu} = (\hat{\delta}_{1,\beta\mu}, ..., \hat{\delta}_{N,\beta\mu})$ *двойственной задачи* (11)*, в свою очередь, определяет разбиение множества признаков* $\hat{P}_{\beta\mu} = (\hat{\mathbb{I}}^-_{\beta\mu}, \hat{\mathbb{I}}^0_{\beta\mu}, \hat{\mathbb{I}}^+_{\beta\mu})$ *на активные* $\hat{\mathbb{I}}_{\beta\mu} = \hat{\mathbb{I}}^-_{\beta\mu} \bigcup \hat{\mathbb{I}}^+_{\beta\mu}$ *и неактивные* $\hat{\mathbb{I}}^0_{\beta\mu} = \big\{i \in \mathbb{I} : \hat{a}_{i,\beta\mu} = 0\big\}$*:*

$$\begin{gathered} \hat{\mathbb{I}}^-_{\beta\mu} = \Big\{i \in \mathbb{I} : \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta\mu} < -\frac{\mu}{2}, \hat{a}_{i,\beta\mu} < 0\Big\},\ \hat{\mathbb{I}}^+_{\beta\mu} = \Big\{i \in \mathbb{I} : \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta\mu} > \frac{\mu}{2}, \hat{a}_{i,\beta\mu} > 0\Big\}, \\ \hat{\mathbb{I}}^0_{\beta\mu} = \Big\{i \in \mathbb{I} : -\frac{\mu}{2} \leqslant \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta\mu} \leqslant \frac{\mu}{2}, \hat{a}_{i,\beta\mu} = 0\Big\}. \end{gathered} \tag{12}$$

Полученные активные коэффициенты регрессии являются линейными функциями от остатков, найденных из (11)

$$\hat{a}_{i,\beta\mu} = \frac{1}{\beta}\Big(\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta\mu} + \frac{\mu}{2}\Big) < 0 \text{ если } i \in \hat{\mathbb{I}}^-_{\beta\mu}, \quad \hat{a}_{i,\beta\mu} = \frac{1}{\beta}\Big(\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta\mu} - \frac{\mu}{2}\Big) > 0 \text{ если } i \in \hat{\mathbb{I}}^+_{\beta\mu}, \tag{13}$$

тогда как оставшиеся коэффициенты равны нулю $\hat{a}_{i,\beta\mu} = 0$ *если* $i \in \hat{\mathbb{I}}^0_{\beta\mu}$.

Таким образом, решение задачи Elastic Net (9)-(10) полностью определяется решением двойственной задачи (11), чья целевая функций строго выпукла, как сумма двух функций, одна из которых выпуклая функция, а другая квадратичная. Следовательно, решение $\hat{\boldsymbol{\delta}}_{\beta\mu} = (\hat{\delta}_{1,\beta\mu}, ..., \hat{\delta}_{N,\beta\mu})$ всегда единственно.

Наглядно видно из (12) что, если коэффициент гребневой регрессии $\beta > 0$ зафиксирован, то штраф на сумму модулей $\mu > 0$ действует как параметр селективности, контролирующий количество признаков, попавших в выбранное множество активных признаков, или, что эквивалентно, регулирует количество отличных от нуля коэффициентов регрессии $\hat{\mathbf{a}}_{\beta\mu} = \big(\hat{a}_{i,\beta\mu}, i \in \mathbb{I}\big)$.

Когда $\mu = 0$, задача Elastic Net становится стандартной задачей гребневой регрессии. Коэффициенты гребневой регрессии (8) вычисляются через оптимальное решение двойственной задачи (11) в соответствии с (13):

$$W\big(\boldsymbol{\delta}|\beta, \mu = 0\big) = \frac{1}{\beta}\boldsymbol{\delta}^T\Big(\sum_{i\in\mathbb{I}} \bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\Big)\boldsymbol{\delta} + (\boldsymbol{\delta} - \mathbf{y})^T(\boldsymbol{\delta} - \mathbf{y}) \to \min(\boldsymbol{\delta}), \tag{14}$$

$$\hat{\boldsymbol{\delta}}_{\beta 0} = \beta\Big(\sum_{i\in\mathbb{I}} \bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T + \beta\mathbf{I}\Big)^{-1}\mathbf{y}, \;\; \hat{a}_{i,\beta 0} = \frac{1}{\beta}\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta 0}, \;\; i \in \mathbb{I}, \;\; \hat{\mathbb{I}}^0_{\beta 0} = \varnothing. \tag{15}$$

Небольшое увеличение параметра селективности μ не изменит исходно пустое множество неактивных признаков, в то время как вектор остатков $\hat{\boldsymbol{\delta}}_{\beta 0}$ удовлетворяет неравенствам $\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta 0} \leqslant -(\mu/2)$ или $\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta 0} \geqslant (\mu/2)$ для всех $i \in \mathbb{I}$.

Напротив, когда μ слишком велико, все признаки становятся неактивными. В самом деле, второе слагаемое в целевой функции двойственной задачи (11) является квадратичной функцией $(\boldsymbol{\delta}-\mathbf{y})^T(\boldsymbol{\delta}-\mathbf{y})$. Если ее точка минимума $\boldsymbol{\delta} = \mathbf{y}$ удовлетворяет неравенствам $-(\mu/2) < \bar{\mathbf{x}}_i^T\mathbf{y} < (\mu/2)$ для всех $i \in \mathbb{I}$, то первое слагаемое становится равным нулю, и вектор $\mathbf{y}$ является решением двойственной задачи.

Таким образом, в соответствии с (12), для того, чтобы перебрать все значения селективности, начиная с тех, при которых активных признаков нет вообще, $\big(\hat{\mathbb{I}}^0_{\beta\mu} = \varnothing, \; \hat{\mathbb{I}}_{\beta\mu} = \mathbb{I}\,\big)$ и заканчивая теми значениями, когда все признаки являются активными, $\big(\hat{\mathbb{I}}^0_{\beta\mu} = \mathbb{I}, \; \hat{\mathbb{I}}_{\beta\mu} = \varnothing\big)$, достаточно менять селективность в интервале

$$2\min_{i\in\mathbb{I}}|\bar{\mathbf{x}}_i^T\hat{\boldsymbol{\delta}}_{\beta 0}| = \mu_{\min} \leqslant \mu \leqslant \mu_{\max} = 2\max_{i\in\mathbb{I}}|\bar{\mathbf{x}}_i^T\mathbf{y}|. \tag{16}$$

Итерационный алгоритм решения задачи Elastic Net для фиксированных параметров регуляризации

Пусть параметры регуляризации $\beta > 0$ и $\mu > 0$ в критерии обучения (9)-(10) зафиксированы. Для того, чтобы описать итерационный алгоритм решения выпуклой двойственной задачи оптимизации (11), мы введем понятие произвольного разбиения P набора признаков $i \in \mathbb{I}$ на три подмножества:

$$P = \big\{\mathbb{I}^-_P, \mathbb{I}^0_P, \mathbb{I}^+_P\big\}, \quad \mathbb{I}^-_P \bigcup \mathbb{I}^0_P \bigcup \mathbb{I}^+_P = \mathbb{I}. \tag{17}$$

Пусть, далее, $P_{\boldsymbol{\delta}} = \big\{\mathbb{I}^-_{\boldsymbol{\delta}}, \mathbb{I}^0_{\boldsymbol{\delta}}, \mathbb{I}^+_{\boldsymbol{\delta}}\big\}$ обозначает разбиение, связанное с конкретным вектором остатков $\boldsymbol{\delta} = (\delta_1, ..., \delta_N) \in \mathbb{R}^N$:

$$\mathbb{I}^-_{\boldsymbol{\delta}} = \Big\{i \in \mathbb{I} : \bar{\mathbf{x}}_i^T\boldsymbol{\delta} < -\frac{\mu}{2}\Big\}, \; \mathbb{I}^0_{\boldsymbol{\delta}} = \Big\{i \in \mathbb{I} : \frac{\mu}{2} \leqslant \bar{\mathbf{x}}_i^T\boldsymbol{\delta} \leqslant \frac{\mu}{2}\Big\}, \; \mathbb{I}^+_{\boldsymbol{\delta}} = \Big\{i \in \mathbb{I} : \bar{\mathbf{x}}_i^T\boldsymbol{\delta} > \frac{\mu}{2}\Big\}. \tag{18}$$

В частности, разбиение в точке минимума критерия Elastic Net (12) может быть переписано в этих терминах как:

$$\hat{P}_{\beta\mu} = P_{\hat{\boldsymbol{\delta}}_{\beta\mu}},\ \hat{\mathbb{I}}^-_{\beta\mu} = \mathbb{I}^-_{\hat{\boldsymbol{\delta}}_{\beta\mu}},\ \hat{\mathbb{I}}^0_{\beta\mu} = \mathbb{I}^0_{\hat{\boldsymbol{\delta}}_{\beta\mu}},\ \hat{\mathbb{I}}^+_{\beta\mu} = \mathbb{I}^+_{\hat{\boldsymbol{\delta}}_{\beta\mu}}.$$

Наконец, мы зафиксируем разбиение P в двойственной задаче оптимизации (11), и рассмотрим квадратичную задачу условной оптимизации и ее очевидное решение:

$$\begin{aligned} W(\boldsymbol{\delta}|P) &= \frac{1}{\beta}\Big\{\sum_{i\in\mathbb{I}^-_P}\Big(\bar{\mathbf{x}}_i^T\boldsymbol{\delta} + \frac{\mu}{2}\Big)^2 + \sum_{i\in\mathbb{I}^-_P}\Big(\bar{\mathbf{x}}_i^T\boldsymbol{\delta} - \frac{\mu}{2}\Big)^2\Big\} + (\boldsymbol{\delta}-\mathbf{y})^T(\boldsymbol{\delta}-\mathbf{y}) \to \min(\boldsymbol{\delta}), \\ \hat{\boldsymbol{\delta}}_P &= \Big(\sum_{i\in\mathbb{I}_P}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T + \beta\mathbf{I}\Big)^{-1}\Big[\mathbf{y} + \frac{1}{2}\Big(\sum_{i\in\mathbb{I}^+_P}\bar{\mathbf{x}}_i - \sum_{i\in\mathbb{I}^-_P}\bar{\mathbf{x}}_i\Big)\mu\Big], \quad \mathbb{I}_P = \mathbb{I}^-_P\bigcup\mathbb{I}^+_P. \end{aligned} \tag{19}$$

Матрица $\sum\limits_{i\in\mathbb{I}_P}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T + \beta\mathbf{I}$, через которую выражается решение $\hat{\boldsymbol{\delta}}_P$ в (19), всегда обратима.

Утверждение 2. *Пусть P обозначет произвольное разбиение множества признаков, и $\hat{\boldsymbol{\delta}}_P$ является решением соответствующей задачи условной оптимизации* (19). *Пусть, кроме этого, $P_{\hat{\boldsymbol{\delta}}_P}$ обозначает разбиение, полченное из $\hat{\boldsymbol{\delta}}_P$* (18). *Совпадение этих двух разбиений $P_{\hat{\boldsymbol{\delta}}_P} = P$ – необходимое и достаточное условие того, чтобы $\hat{\boldsymbol{\delta}}_P$ являлось решением двойственной задачи* (11).

Утверждение 2 описывает итерационный алгоритм решения двойственной задачи (11) и, следовательно, задачи Elastic Net (9)-(10), согласно Утверждению 1.

Алгоритм. *Пусть $\hat{\boldsymbol{\delta}}_k$ является некоторым приближением искомого решения двойственной задачи и $P_k = P_{\hat{\boldsymbol{\delta}}_k}$* (18) *является разбиением множества признаков, соответствующим данному решению. Решение условной задачи оптимизации* (19) *с $P = P_k$ кладется равным следующему приближению $\hat{\boldsymbol{\delta}}_{k+1}$. Алгоритм останавливается, когда разбиения $P_{k+1} = P_{\hat{\boldsymbol{\delta}}_{k+1}}$ и P_k совпадают.*

Опыт показывает, что целесообразно начать с $\hat{\boldsymbol{\delta}}_0 = \mathbf{y}$, то есть с тривиального разбиения $P = \big\{\mathbb{I}^-_P = \varnothing, \mathbb{I}^0_P = \mathbb{I}, \mathbb{I}^+_P = \varnothing\big\}$ при котором все признаки неактивны. В этом случае алгоритм, описанный выше, постепенно расширяет множество активных признаков $\hat{\mathbb{I}}_k = \hat{\mathbb{I}}^-_k\bigcup\hat{\mathbb{I}}^+_k$, но при этом расширение множества практически никогда не является жадным. С таким исходным приближением алгоритм сходится обычно за 15-30 итераций.

Количество итераций часто сокращается до 1-2 только если исходный вектор $\hat{\boldsymbol{\delta}}_0$ близок к решению. Алгоритм регуляризации использует это очень удачное свойство.

Алгоритм регуляризации

Идея алгоритма регуляризации впервые была сформулирована для критерия Лассо ($\beta = 0$) в [7] и затем расширена на случай Elastic Net ($\beta > 0$) в [6]. Мотивацией послужило предположение, что поиск зависимости вектора коэффициентов регрессии от параметра селективности $\hat{\mathbf{a}}_{\beta\mu} = \hat{\mathbf{a}}_\beta(\mu)$ в один проход сразу для всех значений (16) должен быть вычислительно гораздо эффективнее, чем независимое вычисление искомого вектора коэффициентов регрессии для нескольких отдельных значений селективности.

Пусть коэффициент $\beta > 0$ взят достаточно малым, чтобы гарантировать строгую выпуклость целевой функции Elastic Net (9). В основе теоретического обоснования алгоритма регуляризации для всех значений парметра селективности $\mu > 0$ лежит математический факт, что зависимость $\hat{\mathbf{a}}_\beta(\mu)$, определенная обучающим критерием Elastic Net — непрерывная, кусочно-линейная функция, имеющая конечное количество точек бифуркации. Будем рассматривать эти точки как уменьшающуюся последовательность $\big(\mu_{max} = \mu_0 > \mu_1 >, ..., > \mu_M = 0\big)$. Можно доказать, что количество нетривиальных точек M удовлетворяет неравенству $n \leqslant M \leqslant 2^n$, где $n = |\mathbb{I}|$ — количество признаков.

Это значит, что в терминах теоретического подхода, освещенного в разделах 1 и 2, для каждой обучающей совокупности (4), убывающая последовательность точек бифуркации $\mu_{max} \to \mu_k \to 0$ (16) задает соответствующую дискретную последовательность разбиений множества признаков P_k, начиная с разбиения, в котором все признаки отсутствуют $\hat{\mathbb{I}}_0 = \hat{\mathbb{I}}_0^- \bigcup \hat{\mathbb{I}}_0^+ = \varnothing$, и заканчивая разбиением, включающим в себя все признаки $\hat{\mathbb{I}}_M = \mathbb{I}$. Между двумя соседними точками бифуркации разбиение остается неизменным $\mu_{k-1} > \mu > \mu_k$, но следует отметить, что, вообще говоря, размер множества активных признаков $|\hat{\mathbb{I}}_k|$ не является монотонно возрастающей функцией от значения параметра селективности μ.

Теоретически всего точек бифуркации может быть 2^n. Однако такое число чрезмерно велико для применения предлагаемого алгоритма Машины Релевантных Объектов, базирующегося на принципе отбора признаков. В нашем случае количество вторичных признаков $n = |\mathbb{I}| = mN$ превышает размер обучающей совокупности $N = |\mathbb{J}|$ в такое же количество раз, сколько задано функций парного сравнения объектов $m = |\mathbb{K}|$. Поэтому мы будем использовать «урезанный» алгоритм регуляризации.

Пусть μ_{max} обозначает максимальное значение параметра селективности, вычисленное по обучающей совокупности (4) в соответствии с (16), и M — количество значений селективности, которое мы хотим попробовать. Предлагаемая последовательность $(\mu_0 > \mu_1 >, ..., > \mu_M)$ вычисляется как

$$\mu_k = \Big(1 - \frac{k}{M}\Big)\mu_{max}, \quad k = 0, 1, ..., M, \quad \text{при этом} \quad \mu_0 = \mu_{max}, \; \mu_M = 0. \tag{20}$$

Окончательным результатом алгоритма регуляризации является последовательность подмножеств активных признаков $\hat{\mathbb{I}}_k$, $k = 0, 1, ..., M$. Мощность этих подмножеств $|\hat{\mathbb{I}}_k|$ в общем случае возрастает от 0 до полного количества признаков $n = mN$, но ее зависимость от параметра селективности необязательно монотонная.

Нет необходимости вычислять коэффициенты регрессии Elastic Net, потому что мы используем эту модель только лишь для отбора признаков. Вместо этого, мы вычисляем коэффициенты гребневой регрессии (13) для каждого из отобранных подмножеств активных признаков в соответствии с (15):

$$\hat{\boldsymbol{\delta}}_{\hat{\mathbb{I}}_k} = \big(\hat{\delta}_{1,\hat{\mathbb{I}}_k}, ..., \hat{\delta}_{N,\hat{\mathbb{I}}_k}\big) = \Big(\sum_{i \in \hat{\mathbb{I}}_k} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T + \beta \mathbf{I}\Big)^{-1} \mathbf{y}, \quad \hat{a}_{i,\hat{\mathbb{I}}_k} = \begin{cases} (1/\beta)\bar{\mathbf{x}}_i^T \hat{\boldsymbol{\delta}}_{\hat{\mathbb{I}}_k}, & i \in \hat{\mathbb{I}}_k, \\ 0, & i \notin \hat{\mathbb{I}}_k. \end{cases} \tag{21}$$

Бесперебoрная процедура верификации leave-one-out для подмножеств активных признаков

На каждом k-ом шаге алгоритма регуляризации полученное среднее значение суммы квадратов остатков гребневой регрессии (8) для заданного подмножества признаков $\hat{\mathbb{I}}_k \subseteq \mathbb{I}$ дается выражениями:

$$Q_{\hat{\mathbb{I}}_k} = \frac{1}{N} \sum_{j=1}^{N} (\hat{\delta}_{j,\hat{\mathbb{I}}_k})^2, \tag{22}$$

где вектор остатков $\hat{\boldsymbol{\delta}}_{\hat{\mathbb{I}}_k} = \big(\hat{\delta}_{1,\hat{\mathbb{I}}_k}, ..., \hat{\delta}_{N,\hat{\mathbb{I}}_k}\big)$ вычисляется согласно (21). Отметим, что остатки являются непосредственным решением двойственной задачи (11), а вычисление коэффициентов регрессии $\hat{a}_{i,\hat{\mathbb{I}}_k}$ не является обязательным для того, чтобы посчитать $Q_{\hat{\mathbb{I}}_k}$.

Ошибка leave-one-out

$$Q_{\hat{\mathbb{I}}_k}^{LOO} = \frac{1}{N} \sum_{j=1}^{N} (\hat{\delta}_{j,\hat{\mathbb{I}}_k}^{(j)})^2, \tag{23}$$

является средним арифметическим квадратов элементов вектора ошибок, каждый из которых вычисляется по всей обучающей совокупности, исключая соответствующий j-ый элемент, в отличие от (21):

$$\begin{aligned}
&\hat{\boldsymbol{\delta}}^{(j)}_{\hat{\mathbb{I}}_k} = \big(\hat{\delta}^{(j)}_{1,\hat{\mathbb{I}}_k}, ..., \hat{\delta}^{(j)}_{N,\hat{\mathbb{I}}_k}\big) = \Big(\sum_{i\in\hat{\mathbb{I}}_k} \tilde{\bar{\mathbf{x}}}^{(j)}_i (\tilde{\bar{\mathbf{x}}}^{(j)}_i)^T + \beta\tilde{\mathbf{I}}\Big)^{-1} \tilde{\mathbf{y}}^{(j)},\\
&\tilde{\bar{\mathbf{x}}}^{(j)}_i = \big(x_{il}, l \neq j\big) \in \mathbb{R}^{N-1},\ \tilde{\mathbf{y}}^{(j)} = \big(y_l, l \neq j\big) \in \mathbb{R}^{N-1},\ \ \tilde{\mathbf{I}}\big[(N-1)\times(N-1)\big].
\end{aligned} \tag{24}$$

Как показано в [4], квадратичная форма критерия гребневой регрессии (8) позволяет избежать множественных обращений матриц при вычислении ошибки leave-one-out. В терминах напрямую посчитанных остатков в (23)-(24) при решении двойственной задачи (11), ошибка leave-one-out определяется выражением

$$Q^{LOO}_{\hat{\mathbb{I}}_k} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\hat{\delta}_{j,\hat{\mathbb{I}}_k}}{1-q_{j,\hat{\mathbb{I}}_k}}\right)^2,\ \ q_{j,\hat{\mathbb{I}}_k} = \left[\Big(\sum_{i\in\hat{\mathbb{I}}_k}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\Big)\Big(\sum_{i\in\hat{\mathbb{I}}_k}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T + \beta\mathbf{I}\Big)^{-1}\right]_{jj}, \tag{25}$$

где $[...]_{jj}$ означает jый диагональный элемент квадратной матрицы.

Можно заметить, что обратная матрица $\Big(\sum_{i\in\hat{\mathbb{I}}_k}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T + \beta\mathbf{I}\Big)^{-1}$ вычисляется лишь однажды на каждом шаге алгоритма регуляризации при оценке остатков гребневой регрессии по всей обучающей совокупности (21).

Результаты экспериментов

Эксперименты на искусственных данных

Продемонстрируем использование предложенной Машины Релевантных Объектов с отбором модальностей на двухмерных искусственных данных. Данные устроенны таким образом, что по ним явно можно понять истинное множество релевантных объектов и релевантных модальностей. Набор данных состоит из объектов $\omega \in \Omega$, представленных двумя наблюдаемыми признаками $\mathbf{z}(\omega) = \big(z_1(\omega), z_2(\omega)\big) \in \mathbb{R}^2$ и одним скрытым признаком $y(\omega) \in \mathbb{R}$.

Генератор случайных данных основан на предзаданной функции $\mathbb{R}^2 \to \mathbb{R}$, определенной следующим образом:

$$\begin{aligned}
&y^*(\mathbf{z}) = K(\mathbf{z}^*_1, \mathbf{z}) + K(\mathbf{z}^*_2, \mathbf{z}) - K(\mathbf{z}^*_3, \mathbf{z}) - K(\mathbf{z}^*_4, \mathbf{z}),\\
&K(\mathbf{z}^*, \mathbf{z}) = \exp\big\{-5.5||\mathbf{z}^* - \mathbf{z}||^2\big\},\\
&\mathbf{z}^*_1 = (-0.26,\ 0.69),\ \mathbf{z}^*_2 = (0.47,\ 0.76),\ \mathbf{z}^*_3 = (-0.7,\ 0.32),\ \mathbf{z}^*_4 = (0.25,\ 0.4).
\end{aligned} \tag{26}$$

Каждый случайный объект представлен парой $(\mathbf{z}, y)$, где $\mathbf{z}$ – случайный вектор, равномерно распределенный в прямоугольнике $\big(-1 \leqslant z_1 \leqslant 1,\ -0.2 \leqslant z_2 \leqslant 1.2\big)$ и y — нормально распределенная случайная величина с условным математическим ожиданием $E(y|\mathbf{z}) = y^*(\mathbf{z})$ и дисперсией $Var(y) = 0.1$.

Набор данных содержит $|\Omega| = 1250$ объектов, которые случайно разбиты на $N = 150$ объектов для обучения и $N_{test} = 1100$ объектов для контроля. На Рис. 1 показана заданная скрытая функция $y^*(\mathbf{z})$, которую нужно оценить, и обучающая совокупность $\big\{(\mathbf{z}_j, y_j), j = = 1, ..., N\big\}$. Контурные линии заданной скрытой функции хорошо демонстрируют два "холма"в верхней части координатной плоскости и две "впадины"в нижней.

Далее мы применили к обучающей совокупности предложенную в работе Машину Релевантных Объектов с отобором модальностей для фиксированного значения параметра $\beta = 0.1$ и для $m = 4$ функций парного сравнения:

$$\begin{aligned} S_1(\omega', \omega'') &= S_1(\mathbf{z}', \mathbf{z}'') = \exp\{-5.5[(z'_1 - z''_1)^2 + (z'_2 - z''_2)^2]\}, \\ S_2(\omega', \omega'') &= S_2(\mathbf{z}', \mathbf{z}'') = |z'_1 - z''_1|, \\ S_3(\omega', \omega'') &= S_3(\mathbf{z}', \mathbf{z}'') = |z'_2 - z''_2|, \\ S_4(\omega', \omega'') &= S_4(\mathbf{z}', \mathbf{z}'') = |(z'_2 - z''_2) - (z'_1 - z''_1)|. \end{aligned} \tag{27}$$

Особенность скрытой заданной функции (26) указывает на то, что алгоритм должен отобрать только первую функцию $S_1(\mathbf{z}', \mathbf{z}'')$ как единственную адекватную функцию искомой зависимости, а в качестве релевантных объектов отобрать объекты обучающей совокупности, расположенные преимущественно в окрестности "холмов"и "впадин".

Таким образом, мы имеем полный набор $\mathbb{I} = \{i = (kj)\}$ из $n = mN = 400$ вторичных признаков (3), сформированный из $m = 4$ модальностей и $N = 100$ объектов обучения. Для подобной экспериментальной структуры максимальное значение параметра селективности (16) равняется $\mu_{max} \approx 11$.

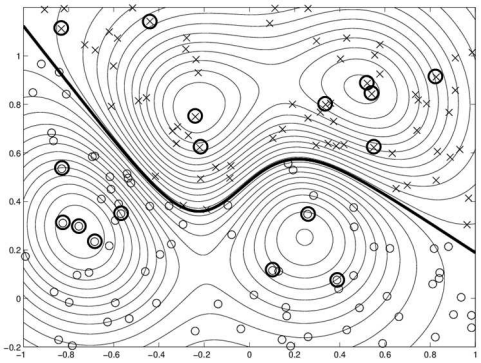


Рис. 1: Контурные линии функции $y^*(\mathbf{z})$ и обучающая совокупность $\{(\mathbf{z}_j, y_j), j = 1, ..., N\}$, $N = 150$. Жирная линия обозначает множество точек, для которых $y^*(\mathbf{z}) = 0$, маркеры $\times$ и $\odot$ обозначают объекты обучающей совокупности $\mathbf{z}_j$ и, соответственно, положительные и отрицательные значения наблюдаемой целевой переменной y_j. Релевантные объекты подсвечены.

Применение алгоритма регуляризации привело к тому, что лучшее значение селективности равно $\mu = 1.21$, давая минимальное срденее значение ошибки leave-one-out $Q^{LOO}_{\hat{\mathbb{I}}} = 0.089$, вычисленное согласно (25). Средний квадрат ошибки на контроле оказался немного выше: $Q^{test}_{\hat{\mathbb{I}}} = 0.125$. Обе оценки хорошо согласуются со значением дисперсии $Var(y) = 0.1$, для которой генерировались искуственные данные.

Соответствующее подмножество релевантных вторичных признаков $\hat{\mathbb{I}} = \{(kj)\}$ содержит 18 элементов. Средин них 17 элементов были сформированы разными релевантными объектами $\hat{\mathbb{J}} = \{j\} \subset \mathbb{J}$, связанными с одной и той же функцией парного сравнения $S_1(\mathbf{z}', \mathbf{z}'')$, как и ожидалось. И только один релевантный объект предпочел другую функцию парного сравнения $S_4(\mathbf{z}', \mathbf{z}'')$ (28). Релевантные объекты расположились в окрестностях "холмов"и "впадин"(Рис. 1), как и ожидалось.

Эксперименты на реальных данных

Эффективность Машины Релевантных Объектов мы проверили на задаче восстановления числовой зависимости, которая заключается в моделировании намагниченно-

сти в сверхпроводниках [8]. Данные для этой задачи предоставлены Национальным Технологическим Институтом Стандартов (National Institute of Standards and Technology) `http://www.nist.gov/index.html`.

Набор данных состоит из объектов $\omega \in \Omega$, представленных одним наблюдаемым признаком $\mathbf{z}(\omega) = z_1(\omega) \in \mathbb{R}$ и одним скрытым признаком $y(\omega) \in \mathbb{R}$.

Скрытый признак — это измеренная намагниченность сверхпроводника, наблюдаемый признак — это логарифм от времени, прошедшего с момента старта эксперимента до момента его завершения. Истинная зависимость между скрытым и наблюдаемым признаком описывается $y = -2000(50 + z_1)^{-10/9}$.

Набор данных содержит $|\Omega| = 154$ объекта, которые случайно разбиты на $N = 75$ объектов для обучения и $N_{test} = 79$ объекта для контроля. На Рис. 2 показана заданная скрытая функция $y^*(\mathbf{z})$, которую нужно оценить, и обучающая совокупность $\{(\mathbf{z}_j, y_j), j = = 1, ..., N\}$.

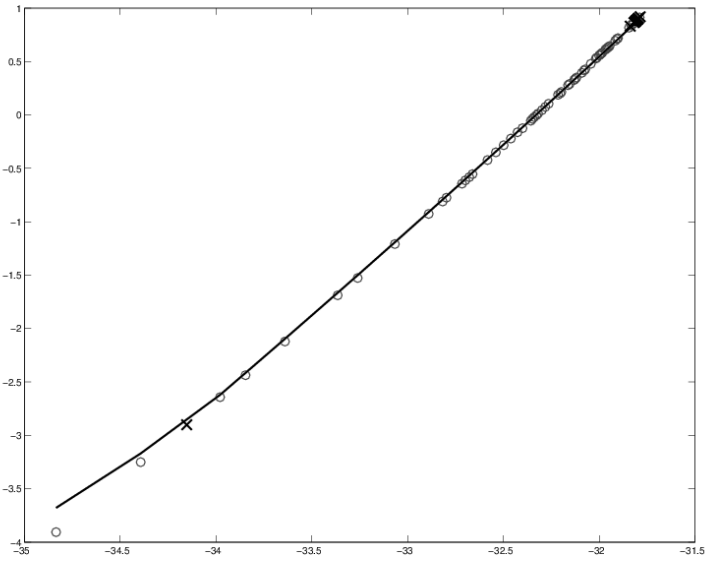


Рис. 2: Обучающая совокупность $N = 75$ и восстановленная по ней зависимость $y^*(\mathbf{z})$. Маркеры $\odot$ и $\times$ обозначают объекты обучающей совокупности $\mathbf{z}_j$ и, соответственно, релевантные объекты.

Далее мы применили к обучающей совокупности предложенную в работе Машину Релевантных Объектов с отбором модальностей для фиксированного значения параметра $\beta = 15$ и для $m = 4$ функций парного сравнения:

$$\begin{aligned} S_1(\omega', \omega'') &= S_1(\mathbf{z}', \mathbf{z}'') = (1 + |z_1' - z_1''|)^{-10/9}, \\ S_2(\omega', \omega'') &= S_2(\mathbf{z}', \mathbf{z}'') = \exp(-1.5(z_1' - z'')^2), \\ S_3(\omega', \omega'') &= \exp(-1.5 * |z_1' - z_1''|)/(1 + (z_1' + z_1'')^2), \\ S_4(\omega', \omega'') &= S_4(\mathbf{z}', \mathbf{z}'') = \exp(-1.5|z_1' - z_1''|) \end{aligned} \tag{28}$$

Особенность скрытой заданной функции указывает на то, что алгоритм должен отобрать только первую функцию $S_1(\mathbf{z}', \mathbf{z}'')$ как единственную адекватную функцию искомой зависимости, а в качестве релевантных объектов отобрать небольшое количество объектов обучающей совокупности.

Таким образом, мы имеем полный набор $\mathbb{I} = \{i = (kj)\}$ из $n = mN = 300$ вторичных признаков (3), сформированный из $m = 4$ модальностей и $N = 75$ объектов обучения. Для подобной экспериментальной структуры максимальное значение параметра селективности (16) равняется $\mu_{max} \approx 94.5$.

Применение алгоритма регуляризации привело к тому, что лучшее значение селективности равно $\mu = 186.4$, давая минимальное срденее значение ошибки leave-one-out $Q^{LOO}_{\hat{\mathbb{I}}} = 0.000695$, вычисленное согласно (25). Средний квадрат ошибки на контроле оказался немного выше: $Q^{test}_{\hat{\mathbb{I}}} = 0.000805$. Обе оценки хорошо согласуются со значением квадрата отклонения истинной зависимости от наблюдаемых значений скрытой переменной $Var(y) = 0.00052$.

Соответствующее подмножество релевантных вторичных признаков $\hat{\mathbb{I}} = \{(kj)\}$ содержит 11 элементов. Все эти элементы были сформированы разными релевантными объектами $\hat{\mathbb{J}} = \{j\} \subset \mathbb{J}$, связанными с одной и той же функцией парного сравнения $S_1(\mathbf{z}', \mathbf{z}'')$, как и ожидалось.

Выводы

В данной работе предложена методология решения задач восстановления регрессии на основе множества произвольных действительнозначных функций парного сравнения, не обязанных удовлетворять трудновыполнимым на практике условиям положительной полуопределенности. Предложенная разработка является обобщением относительного дискриминантного анализа (Relational Discriminant Analysis) Роберта Дьюина, с одной стороны, и Машины Релевантных Векторов Типпинга, с другой стороны.

Критерий обучения отличается от озвученных ранее, тем, что он явно включает в себя структурный параметр, контролирующий степень отобора признаков, описывающих объекты, каждый из которых связан с соответствующим объектом в обучающей совокупности и с конкретной функцией парного сравнения. Кроме того, многомодальная Машина Релевантных Объектов с селективностью является выпуклой, следовательно, гарантированно имеет единственное решение.

Явное наличие параметра селективности делает критерий удобным для задач, в которых присутствуют несколько функций парного сравнения, т.е. биоинформатику и многомодальную биометрию.

Литература

[1] *V. Vapnik. Estimation of Dependencies Based on Empirical Data* Springer, 1982.

[2] *R. Duin, E. Pekalska, D. de Ridder* Relational discriminant analysis *Pattern Recognition Letters*, 1999, Vol. 20, pp. 1175-1181.

[3] *C. Bishop, M. Tipping* Variational Relevance Vector Machines *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 46-53. Morgan Kaufmann, 2000.

[4] *G.C. Cawley, N.L.C. Talbot* Fast exact leave-one-out cross-validation of sparse least-squares support vector machines *Neural Networks*, 2004, Vol. 17, pp. 1467-1475.

[5] *M.S. Bartlett* An inverse matrix adjustment arising in discriminant analysis *Annals of Mathematical Statistics*, 1951, Vol. 22(1), pp. 107–111.

[6] *H. Zou and T. Hastie* Regularization and variable selection via the elastic net *Journal of the Royal Statistical Society*, 67: 301–320, 2005.

[7] *R. Tibshirani* Regression shrinkage and selection via the lasso *Journal of the Royal Statistical Society*, 58(1): 267–288, 1996.

[8] *L. Bennett, L. Swartzendruber, H. Brown* Superconductivity Magnetization Modeling NIST 1994