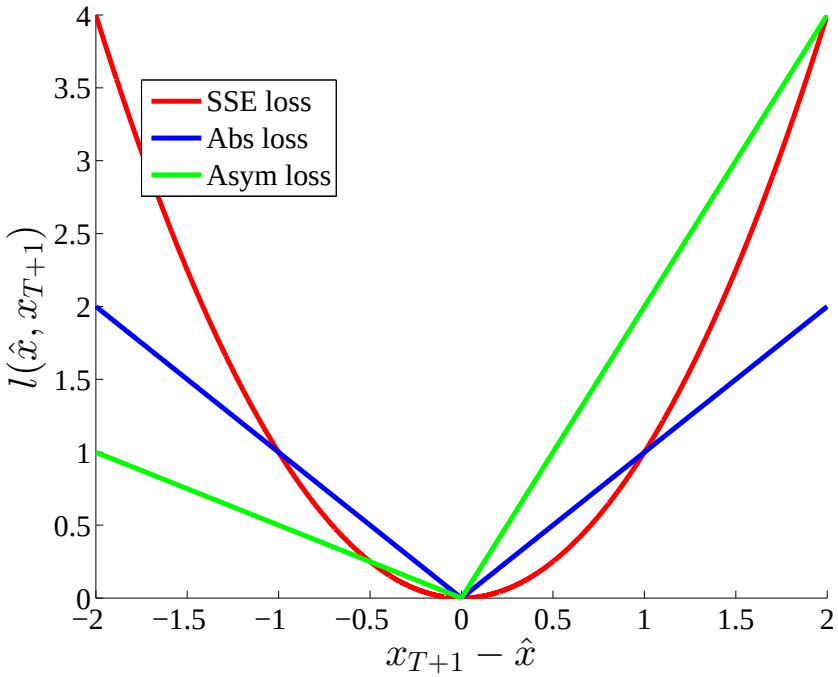
ISSN 2223-3792

Машинное обучение и анализ данных

2015 год **Том 1, номер 14**



Машинное обучение и анализ данных

Журнал «Машинное обучение и анализ данных» публикует новые теоретические и обзорные статьи с результатами научных исследований в области искусственного интеллекта, теоретических основ информатики и 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. Журнал зарегистрирован в системе Crossref, doi http://dx.doi.org/10.21469/22233792.

- Новостной сайт http://jmlda.org/
- Электронная система подачи статей http://jmlda.org/papers/
- Правила подготовки статей http://jmlda.org/papers/doc/authors-guide.pdf

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

Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Москва, 2015

Journal of Machine Learning and Data Analysis

The journal Machine Learning and Data Analysis publishes original research papers and reviews of the developments in the field of artificial intelligence, theoretical computer science and its applications. The journal aims to promote the theory of machine learning and data mining and methods of conducting computational experiments. Papers are accepted in English and Russian.

The journal is included in the Russian science citation index RSCI. Information about citation to articles can be found at the Russian science citation index website. ISSN 2223-3792. Mass media registration certificate ЭЛ № ФС 77-55486. The Crossref journal doi is `http://dx.doi.org/10.21469/22233792`.

- Journal news and archive http://jmlda.org/
- Open journal system for papers submission http://jmlda.org/papers/
- Style guide for authors http://jmlda.org/papers/doc/authors-guide.pdf

The scope of the journal:
— classification, clustering, regression analysis;
— multidimensional statistical analysis;
— Bayesian methods for regression and classification;
— model selection and complexity;
— deep learning;
— Statistical Learning Theory;
— time series forecasting techniques;
— methods of signal processing and speech recognition;
— optimization methods for solving machine learning and data mining problems;
— methods of big data analysis;
— data visualization techniques;
— methods of image processing and recognition;
— text analysis, text mining and information retrieval;
— applied data analysis problems.

Dorodnicyn Computing Centre FRC CSC RAS
Moscow Institute of Physics and Technology
Department of Control and Applied Mathematics
Division “Intelligent Systems”

Moscow, 2015

Содержание

Contents

Прогнозирование нестационарных временных рядов при несимметричных функциях потерь*

В. Ю. Черных[1]***, М. М. Стенина***[1,2]
vladimir.chernykh@phystech.edu, mmedvednikova@gmail.com
[1]Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9
[2]Высшая школа экономики, Россия, г. Москва, ул. Мясницкая, 20

Рассматривается задача прогнозирования временны́х рядов при несимметричных функциях потерь. Предлагается двухэтапный алгоритм прогнозирования ARIMA + Hist. На первом этапе используется авторегрессионное интегрированное скользящее среднее ARIMA с сезонной компонентой в случае необходимости. Параметры модели подбираются согласно методологии Бокса–Дженкинса. На втором этапе проводится анализ регрессионных остатков и находится оптимальная добавка к прогнозу, полученному на первом шаге, минимизирующая математическое ожидание потерь. Для оценки ожидаемых потерь используется свертка функции потерь с гистограммой регрессионных остатков. Работа предлагаемого двухэтапного алгоритма иллюстрируется на временны́х рядах с различными элементами нестационарности (тренд, сезонность) и для различных симметричных и несимметричных функций потерь. Демонстрируется, что качество прогнозов двухэтапного алгоритма превосходит качество прогнозов модели ARIMA в случае несимметричных функций потерь.

Ключевые слова: *прогнозирование; временные ряды; нестационарность; ARIMA; свертка с функцией потерь; несимметричная функция потерь*

DOI: 10.21469/22233792.1.14.01

1 Введение

Рассматривается задача прогнозирования нестационарных временны́х рядов в случае несимметричных функций потерь. Предлагается двухэтапный алгоритм прогнозирования ARIMA + Hist, на первом этапе которого отслеживаются свойства временно́го ряда, обусловливающие его нестационарность, такие как тренд и сезонность. На втором этапе предлагается находить поправку, обеспечивающую оптимальность прогноза в случае несимметричной функции потерь.

Свойства прогнозов временны́х рядов при использовании несимметричных функций потерь были исследованы в работе [1], авторы которой отмечают смещенность оптимальных прогнозов при несимметричных потерях и делают вывод о необходимости разработки специальных методов прогнозирования временны́х рядов в условиях несимметричности функции потерь.

Один из используемых методов прогнозирования нестационарных временны́х рядов, авторегрессионное интегрированное скользящее среднее ARIMA [2], позволяет с хорошим качеством прогнозировать временны́е ряды с трендом, а также при небольшой модификации и ряды с сезонной компонентой. Однако настройка параметров этого алгоритма осуществляется путем минимизации квадратичной функции потерь, но функция потерь, по которой производится оценка качества прогноза, может существенно отличаться от

*Работа выполнена при финансовой поддержке РФФИ, проект № 14-07-31046.

квадратичной. Это приводит к тому, что оптимальный прогноз для модели ARIMA является несмещенным, а регрессионные остатки должны удовлетворять условиям, описанным далее. Ввиду вышесказанного, модель ARIMA не подходит для решения задачи прогнозирования в случае несимметричной функции потерь, что отмечается в [1, 3].

Авторами работ [4, 5] были предложены модификации модели ARIMA, позволяющие учесть несимметричность функции потерь при настройке параметров алгоритма. Однако обе предложенные модификации сложны в реализации, не позволяют использовать пакеты для прогнозирования временны́х рядов, в которых есть стандартные реализации ARIMA, и требуют для каждой функции потерь создания и обучения индивидуальной модели, что неприемлемо в промышленных задачах. Еще одним методом, предложенным для работы с несимметричными функциями потерь, является квантильная регрессия [6]. Она позволяет находить оптимальный смещенный прогноз для несимметричных функций потерь кусочно-линейного вида, но не дает возможности работать с функциями потерь других видов, а также применима только для стационарных временны́х рядов.

Предлагаемый алгоритм ARIMA + Hist использует результат из [7] о том, что при несимметричной функции потерь оптимальный прогноз смещен, причем его смещение зависит только от функции потерь и дисперсии временно́го ряда. Также используется идея из статьи [8], где автор для построения прогноза использовал авторегрессионную модель с минимизацией квадратичной функции потерь для получения несмещенного прогноза и анализ регрессионных остатков для оценки оптимального смещения прогноза.

Алгоритм ARIMA + Hist строит прогноз в два этапа. На первом этапе используется модель ARIMA с сезонной компонентой в случае необходимости, параметры которой подбираются при помощи анализа временно́го ряда по методологии Бокса–Дженкинса [2]. На этом этапе получается несмещенный прогноз. На втором этапе производится анализ регрессионных остатков модели ARIMA с целью оценки оптимального смещения прогноза для минимизации математического ожидания потерь. Оптимальное смещение находится при помощи алгоритма Hist. Финальный прогноз получается путем прибавления к несмещенному прогнозу, полученному с помощью ARIMA, найденной при помощи алгоритма Hist добавки.

Алгоритм Hist является обобщением алгоритма квантильной регрессии [6]. Он находит приближенное решение задачи минимизации математического ожидания потерь и используется только для прогнозирования стационарных временны́х рядов. Такая задача минимизации рассматривалась в работах [9, 10], где математическое ожидание потерь было представлено как свертка функции потерь с функцией плотности распределения значений временно́го ряда. На втором этапе алгоритма ARIMA + Hist в качестве временно́го ряда выступают регрессионные остатки, однако их плотность распределения неизвестна. В качестве оценки плотности используется гистограмма значений регрессионных остатков, как предложено в [11]. В алгоритме Hist используется ряд упрощений задачи минимизации свертки функции потерь с оценкой плотности распределения регрессионных остатков, которые приводят к задаче приближенного нахождения минимума путем перебора конечного числа значений, из которых выбирается то, которое обеспечивает наименьшее значение свертки.

Основное преимущество ARIMA + Hist состоит в том, что не накладывается ограничений на класс функций потерь, которые можно использовать в задаче прогнозирования.

Алгоритм тестируется на наборе временны́х рядов, обладающих различными элементами нестационарности. Качество полученных прогнозов сравнивается с качеством прогнозов модели ARIMA при использовании различных функций потерь. Демонстрируется,

что чем более несимметричная будет функция потерь, тем более существенный выигрыш в качестве можно будет получить при помощи двухэтапного алгоритма ARIMA + Hist по сравнению с ARIMA.

2 Задача прогнозирования временных рядов

Данные представляют собой временной ряд $\mathbf{x} = \{(x_i)_{i=1}^T \,|\, x_i \in \mathbb{R}\}$. Также задается горизонт прогнозирования h. Ставится задача прогнозирования этого временно́го ряда, т. е. нахождения регрессионной модели

$$f : (\mathbf{w}, \mathbf{x}, h) \mapsto \hat{\mathbf{x}},$$

где $\mathbf{w}$ — вектор параметров; $\hat{\mathbf{x}}$ — вектор прогнозов длины h. В данной работе прогнозирование производится с горизонтом $h = 1$, поэтому вектор прогнозов $\hat{\mathbf{x}}$ является скаляром и обозначается далее как $\hat{x}_{T+1}$.

2.1 Прогнозирование стационарных временных рядов

Временной ряд $\mathbf{x}$ называется *стационарным*, если для любых v многомерное распределение $x_t, \ldots, x_{t+v}$ не зависит от t, т. е. его свойства не зависят от времени. Из определения немедленно следует, что все значения ряда $x_1, \ldots, x_T$ генерируются из одного распределения $\rho(u)$, которое не меняется во времени. Пусть задана функция потерь $\mathscr{L}(\hat{x}, x)$ и требуется получить прогноз $\hat{x}_{T+1}$ следующего значения x_{T+1} временно́го ряда, минимизируя ожидаемые потери. Предполагается, что следующее значение временно́го ряда генерируется из того же распределения, что и все предыдущие. При этом задача прогнозирования запишется как

$$\hat{x}_{T+1} = \arg\min_{c \in \mathbb{R}} \mathsf{E}\, \mathscr{L}(c, x_{T+1}).$$

Если предположить, что плотность распределения $\rho(u)$, из которого генерируются значения временно́го ряда, известна, математическое ожидание потерь запишется как

$$L(c) = \mathsf{E}\, \mathscr{L}(c, x_{T+1}) = \int_{-\infty}^{+\infty} \mathscr{L}(c, u)\, \rho(u)\, du.$$

В таком случае задача прогнозирования формулируется как

$$\hat{x}_{T+1} = \arg\min_{c \in \mathbb{R}} \int_{-\infty}^{+\infty} \mathscr{L}(c, u)\, \rho(u)\, du \equiv \arg\min_{c \in \mathbb{R}} L(c). \tag{1}$$

2.2 Прогнозирование нестационарных временных рядов

В случае, когда временной ряд не является стационарным, необходимо оценить и исключить из временно́го ряда нестационарные особенности, прежде чем минимизировать ожидаемые потери в задаче (1). Таким образом, прогноз $\hat{x}_{T+1}$ нестационарного временно́го ряда будет складываться из двух частей: прогноз нестационарной компоненты $\hat{x}^{\mathrm{ns}}_{T+1}$ и прогноз стационарной компоненты $\hat{x}^{\mathrm{s}}_{T+1}$:

$$\hat{x}_{T+1} = \hat{x}^{\mathrm{ns}}_{T+1} + \hat{x}^{\mathrm{s}}_{T+1}.$$

Алгоритм прогнозирования нестационарной компоненты временного ряда должен быть таким, чтобы регрессионные остатки при прогнозе доступной для обучения истории $\mathbf{x}$

$$\mathbf{r} = \{(r_i)_{i=1}^T \,|\, r_i = x_i - \hat{x}_i^{\text{ns}}\}$$

были стационарным временны́м рядом, значения которого сгенерированы из одного распределения с плотностью $\gamma(u)$.

В качестве алгоритма прогнозирования нестационарной части ряда предлагается использовать ARIMA. Для оптимизации параметров этот алгоритм использует квадратичную функцию потерь $\mathscr{L}_{\text{sq}}(\hat{x}, x) = (\hat{x} - x)^2$, по которой строится функционал потерь:

$$\mathcal{Q}(f^{\text{ns}}, \mathbf{x}) = \frac{1}{T}\sum_{i=1}^{T} \mathscr{L}_{\text{sq}}(f^{\text{ns}}(\mathbf{w}, \mathbf{x}_i, 1), x_{i+1}); \quad \mathbf{x}_i = \{x_1 \cdots x_i\}. \tag{2}$$

Тогда решение задачи минимизации (2) дает вектор параметров искомой регрессионной модели:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^n} \mathcal{Q}(f^{\text{ns}}, \mathbf{x}).$$

При этом прогноз вычисляется следующим образом:

$$\hat{x}_{T+1}^{\text{ns}} = f^{\text{ns}}(\mathbf{w}^*, \mathbf{x}, 1).$$

После получения прогноза нестационарной компоненты временно́го ряда $\hat{x}_{T+1}^{\text{ns}}$ прогноз стационарной компоненты $\hat{x}_{T+1}^{\text{s}}$ может быть получен при помощи оценки плотности распределения $\gamma(u)$ регрессионных остатков $\mathbf{r}$ и решения для этой плотности задачи минимизации ожидаемых потерь (1).

3 Прогнозирование нестационарной компоненты. ARIMA. Методология Бокса–Дженкинса

В данном разделе описывается модель авторегрессионного интегрированного скользящего среднего ARIMA и методология Бокса–Дженкинса прогнозирования временны́х рядов. Принято записывать модель в виде ARIMA(p, d, q), где $p, d, q \in \mathbb{Z}_+$ — структурные параметры, характеризующие порядок для соответствующих частей модели — авторегрессионной, интегрированной и скользящего среднего. ARIMA с подходящими параметрами для каждого временно́го ряда предлагается использовать для получения прогноза нестационарной компоненты $\hat{x}_{T+1}^{\text{ns}}$. Анализ того, насколько хорошо выбранная модель аппроксимирует временной ряд, по методологии Бокса–Дженкинса, включает проверку регрессионных остатков на несмещенность, стационарность и неавтокоррелированность. Модель считается подходящей для аппроксимации временно́го ряда, если все эти свойства выполняются для ряда регрессионных остатков, как это описано в [12]. Таким образом, при выборе подходящей модели ARIMA для прогнозирования нестационарной компоненты временно́го ряда получается стационарный ряд регрессионных остатков, который можно использовать для построения прогноза $\hat{x}_{T+1}^{\text{s}}$ стационарной компоненты временно́го ряда.

Стационарный временной ряд со средним значением μ описывается моделью ARMA(p, q), если выполняется

$$x_t = \alpha + \varepsilon_t + \sum_{i=1}^{q} \psi_i \varepsilon_{t-i} + \sum_{i=1}^{p} \theta_i x_{t-i}; \quad \alpha = \mu\left(1 - \sum_{i=1}^{p} \theta_i\right).$$

где $\theta_1, \ldots, \theta_p$, $\psi_1, \ldots, \psi_q$ — константы; ε_t — гауссов белый шум с нулевым средним и постоянной дисперсией. Вводя оператор сдвига L, действующий по правилу $Lx_i = x_{i-1}$, можно записать модель ARMA(p, q) в следующем виде:

$$\theta(L)x_t = \alpha + \psi(L)\varepsilon_t; \quad \theta(L) = 1 - \sum_{i=1}^{p} \theta_i L^i; \quad \psi(L) = 1 + \sum_{i=1}^{q} \psi_i L^i. \tag{3}$$

Временной ряд описывается моделью ARIMA(p, d, q), если ряд его разностей

$$\nabla^d x_t = (1 - L)^d x_t$$

описывается моделью (3), при этом модель ARIMA(p, d, q) записывается как

$$\theta(L)\nabla^d x_t = \alpha + \psi(L)\varepsilon_t.$$

Временной ряд, обладающий мультипликативной сезонностью с периодом S, описывается моделью ARIMA$(p, d, q) \times (P, D, Q)_S$, если

$$\theta_p(L)\Theta_P(L^S)\nabla^d\nabla_S^D x_t = \alpha + \psi_q(L)\Psi_Q(L^S)\varepsilon_t\,.$$

3.1 Методология Бокса–Дженкинса анализа временных рядов

Методология Бокса–Дженкинса используется для оценки параметров модели ARIMA. Согласно этой методологии, порядок дифференцирования временно́го ряда d выбирается так, чтобы ряд разностей порядка d был стационарным. Параметры p и q выбирают при помощи анализа автокорреляционной и частичной автокорреляционной функций.

Определение 1. *Автокорреляционная функция* ACF_τ *с лагом автокорреляции* τ *для временно́го ряда* $\mathbf{x}$ *вычисляется по формуле:*

$$\mathrm{ACF}_\tau = \frac{\sum_{i=1}^{T-\tau}(x_i - \bar{x})(x_{i+\tau} - \bar{x})}{\sum_{i=1}^{T}(x_i - \bar{x})^2}; \quad \bar{x} = \frac{1}{T}\sum_{i=1}^{T} x_i\,.$$

Определение 2. *Частичная автокорреляционная функция* PACF_τ *с лагом автокорреляции* τ *для стационарного временно́го ряда* $\mathbf{x}$ *вычисляется по формуле:*

$$\mathrm{PACF}_\tau = \begin{cases} \mathsf{E}\left[x_{t+1}x_t\right], & \tau = 1; \\ \mathsf{E}\left[(x_{t+\tau} - x_{t+\tau}^{\tau-1})(x_t - x_t^{\tau-1})\right], & \tau \geqslant 2; \end{cases}$$

$$x_t^{\tau-1} = \beta_1 x_{t+1} + \beta_2 x_{t+2} + \cdots + \beta_{\tau-1} x_{t+\tau-1};$$

$$x_{t+\tau}^{\tau-1} = \beta_1 x_{t+\tau-1} + \beta_2 x_{t+\tau-2} + \cdots + \beta_{\tau-1} x_{t+1},$$

где $\beta_1, \ldots, \beta_{\tau-1}$ *— коэффициенты линейной регрессии.*

Выбор параметров p и q осуществляется из следующих соображений:

(1) в модели ARIMA$(p, d, 0)$ автокорреляционная функция экспоненциально затухает или имеет синусоидальный вид, а частичная автокорреляционная функция значимо отличается от нуля при лагах, не больших p;

(2) в модели ARIMA$(0, d, q)$ частичная автокорреляционная функция экспоненциально затухает или имеет синусоидальный вид, а автокорреляционная функция значимо отличается от нуля при лагах, не больших q.

Множество структурных параметров сезонной компоненты ARIMA назначается с помощью анализа автокорреляционной и частичной автокорреляционной функций. При наличии сезонной компоненты у временно́го ряда на графиках этих функций будут наблюдаться характерные максимумы в лагах, соответствующих периоду S сезонной компоненты.

Необходимые коэффициенты многочленов $\theta(L)$ и $\psi(L)$ оптимизируются при использовании квадратичной функции потерь $\mathscr{L}_{\text{sq}}(\hat{x}, x)$ и функционала потерь (2), основанного на ней.

После оптимизации параметров модели проводится анализ остатков. Регрессионные остатки проверяются на

(1) несмещенность $\mathsf{E}\, r_t = 0$;
(2) стационарность $\forall t \hookrightarrow r_t \sim \gamma(u)$;
(3) неавтокоррелированность $\mathsf{E}\,[r_t r_{t+k}] = 0, k \neq 0$,

Если регрессионные остатки обученной модели обладают всеми этими свойствами, то модель признается подходящей для аппроксимации анализируемого временно́го ряда.

4 Прогнозирование стационарной компоненты. Hist. Влияние функции потерь

После обучения модели ARIMA с выбранными параметрами даваемый ею прогноз $\hat{x}^{\text{ns}}_{T+1}$ учитывает характерные особенности временно́го ряда $\mathbf{x}$, но не функции потерь $\mathscr{L}(\hat{x}, x)$. Пусть ряд регрессионных остатков $\mathbf{r}$ описывается неизвестной плотностью распределения $\gamma(u)$. Предлагается построить добавочный прогноз $\hat{x}^{\text{s}}_{T+1}$ для стационарного ряда из регрессионных остатков, который минимизирует математическое ожидание потерь (1). Такая добавка к несмещенному прогнозу $\hat{x}^{\text{ns}}_{T+1}$ позволит учесть особенности несимметричной функции потерь.

В случае квадратичной $\mathscr{L}_{\text{sq}}(\hat{x}, x) = (\hat{x} - x)^2$ или абсолютной $\mathscr{L}_{\text{abs}}(\hat{x}, x) = |\hat{x} - x|$ функций потерь добавочный прогноз x^{s}_{T+1} можно найти аналитически, не зная при этом конкретного вида распределения $\gamma(u)$: для квадратичной функции потерь $x^{\text{s}}_{T+1} = \mathsf{E}\, r_t$; для абсолютной $x^{\text{s}}_{T+1} = \operatorname{med} \gamma(u)$, что можно получить, продифференцировав $L(c)$. В случае же более общего вида функции потерь задача не поддается аналитическому решению без знания конкретного распределения $\gamma(u)$, а в практических задачах оно, как правило, неизвестно.

Алгоритм Hist предлагает следующий путь для решения этой проблемы. Он состоит из двух приближений функции $L(c)$.

4.1 Оценка плотности $\gamma(u)$ гистограммой

Плотность вероятности $\gamma(u)$ приближается гистограммой значений ряда, т. е. кусочно-постоянной функцией $\hat{\gamma}(u)$. Обозначим $u_{\min} = \min\limits_t r_t$; $u_{\max} = \max\limits_t r_t$. Они существуют, так как рассматриваются только конечные множества $\mathbf{r}$. Произведем разбиение области $[u_{\min}; u_{\max}]$ на n отрезков $[u_i; u_{i+1}]$ равной длины, где

$$u_i = u_{\min} + ia; \;\; a = \frac{u_{\max} - u_{\min}}{n}.$$

На этих отрезках положим значения функции $\hat{\gamma}(u)$ постоянными и равными y_i на отрезке $[u_{i-1}; u_i]$, где y_i пропорционально количеству точек ряда $\mathbf{r}$, значения которых $r_t \in [u_{i-1}; u_i]$.

Точное значение y_i определяется из условия нормировки функции $\hat{\gamma}(u)$:

$$\int\limits_{u_{\min}}^{u_{\max}} \hat{\gamma}(u)\,du = \sum_{i=1}^{n} y_i(u_{i-1} - u_i) = a\sum_{i=1}^{n} y_i = 1.$$

Тогда $\hat{\gamma}(u)$ есть оценка плотности распределения. При использовании этого приближения функция математического ожидания потерь $L(c)$ оценивается как

$$L_{\text{hist}}(c) = \int\limits_{u_{\min}}^{u_{\max}} \mathscr{L}(c,u)\hat{\gamma}(u)\,du = \sum_{i=1}^{n} y_i \int\limits_{u_{i-1}}^{u_i} \mathscr{L}(c,u)\,du. \tag{4}$$

4.2 Приближение интеграла

Интеграл от функции потерь, присутствующий в (4), приближается по методу прямоугольников со средней точкой:

$$\int\limits_{u_{i-1}}^{u_i} \mathscr{L}(c,u)\,du \approx a\,\mathscr{L}\left(c, \frac{u_{i-1}+u_i}{2}\right).$$

После этого приближенная функция математического ожидания потерь примет окончательный вид:

$$L_{\text{conv}}(c) = a\sum_{i=1}^{n} y_i\,\mathscr{L}\left(c, \frac{u_{i-1}+u_i}{2}\right). \tag{5}$$

Точность приближений растет с ростом числа отрезков n; в первом случае (4) это связано с уточнением приближения гистограммой исходной плотности распределения, во втором (5) — с уточнением оценки интеграла.

Работа алгоритма заключается в поиске c^*, на котором достигается минимум $L_{\text{conv}}(c)$, и взятии значения c^* в качестве прогноза. В силу возможной сложности функции потерь $\mathscr{L}(\hat{x},x)$ минимум ищется среди значений функции в ограниченном наборе точек

$$G = \left\{\frac{u_0+u_1}{2}, \cdots, \frac{u_{n-1}+u_n}{2}\right\},$$

состоящем из середин отрезков разбиений:

$$\hat{x}^{\text{s}}_{T+1} = \arg\min_{c\in G} L_{\text{conv}}(c). \tag{6}$$

4.3 Алгоритм Hist

Вход: стационарный ряд регрессионных остатков $\mathbf{r}$, функция потерь $\mathscr{L}(\hat{x}, x)$, число столбцов гистограммы n;
Выход: прогноз $\hat{x}^s_{T+1}$, минимизирующий математическое ожидание потерь;

1: вычислить ширину столбцов гистограммы $a = (\max \mathbf{r} - \min \mathbf{r})/n$ и координаты концов отрезков постоянства $u_0, u_1, \ldots, u_n$ для функции $\hat{\gamma}(u)$;
2: построить гистограмму, найти функцию $\hat{\gamma}(u)$, отнормировав гистограмму, получить значения функции на отрезках постоянства $y_1, \ldots, y_n$;
3: найти значения свертки $\sum\limits_{i=1}^{n} y_i \, \mathscr{L}\left(c, (u_i + u_{i-1})/2\right)$
для всех $c \in \{(u_0 + u_1)/2, \ldots, (u_{n-1} + u_n)/2\}$;
4: выбрать c^*, дающее минимальное значение свертки;
5: $\hat{x}^s_{T+1} = c^*$.

Основной параметр алгоритма, который можно варьировать, — число столбцов гистограммы n. При малых n оценка плотности распределения $\hat{\gamma}(u)$ получается огрубленной, при больших n — более детальной. В следующем разделе будут приведены результаты исследования свойств алгоритма и на регрессионных остатках различных временны́х рядов.

Алгоритм Hist минимизирует математическое ожидание потерь прогнозирования при любом распределении регрессионных остатков $\mathbf{r}$ и произвольной функции потерь $\mathscr{L}(\hat{x}, x)$. Если регрессионные остатки имеют нулевое среднее, то смещение $\hat{x}^s_{T+1}$ будет обусловлено лишь несимметричностью функции потерь. Однако использование алгоритма ARIMA + Hist при двухэтапном прогнозе может уменьшить средние потери и в том случае, если подобранная для прогнозирования нестационарной компоненты модель дает смещенные прогнозы. Это может происходить при смене характера тренда. Например, увеличивается темп роста величины или величина сначала убывала, а потом начинает возрастать. Если изменение однократное и смещение прогнозов ARIMA, обученной на первом характере тренда, постоянное, то при поиске решения задачи минимизации ожидаемых потерь (6) это смещение будет скомпенсировано, что приведет к повышению качества даже в случае симметричной функции потерь. Примером такого ряда может служить ряд [13], во второй половине которого тренд более сильный. В разд. 7 будет показано, что для этого временно́го ряда использование алгоритма Hist действительно дает ощутимый выигрыш в качестве по сравнению с ARIMA даже для квадратичной функции потерь.

Отметим также, что прогноз Hist является константой в том смысле, что не зависит от горизонта прогнозирования, и для любого будущего момента времени t ответ будет одним и тем же, т. е. Hist строит регрессионную модель нулевого порядка.

5 Двухэтапное прогнозирование. Алгоритм ARIMA + Hist

Как было показано в (2), ARIMA настраивается таким образом, чтобы минимизировать регрессионные остатки для квадратичной функции потерь $\mathscr{L}_{\text{sq}}(\hat{x}, x)$. Если же функция потерь $\mathscr{L}(\hat{x}, x)$, по которой производится оценка качества прогноза, не является квадратичной, то и регрессионные остатки в общем случае минимальными не будут.

Далее запускается алгоритм Hist с действительной функцией ошибок на ряде $\mathbf{r}$, т. е. минимизируем остатки в смысле новой функцией потерь $\mathscr{L}(\hat{x}, x)$. Отметим, что первый этап действительно необходим и нельзя давать на вход Hist просто продифференцирован-

ный ряд, так как этот алгоритм не учитывает некоторые особенности, которые не уходят при дифференцировании, например сезонность.

Итоговой прогноз суммируется из двух $\hat{x}_{T+1} = \hat{x}^{\text{ns}}_{T+1} + \hat{x}^{\text{s}}_{T+1}$. По сути же Hist добавляет одинаковый на всем горизонте сдвиг вверх или вниз от исходного прогноза ARIMA в зависимости от конкретного вида функции потерь.

5.1 Алгоритм ARIMA + Hist

Вход: временной ряд $\mathbf{x}$, функция потерь $\mathscr{L}(\hat{x}, x)$;
Выход: прогноз $\hat{x}_{T+1}$;

1: подобрать подходящую для временно́го ряда модель ARIMA по методологии Бокса–Дженкинса;
2: вычислить прогноз нестационарной компоненты $\hat{x}^{\text{ns}}_{T+1}$ с помощью выбранной модели ARIMA;
3: вычислить регрессионные остатки $\mathbf{r}$ для выбранной модели ARIMA;
4: задать число столбцов n в гистограмме для алгоритма Hist;
5: вычислить прогноз стационарной компоненты $\hat{x}^{s}_{T+1}$ с помощью алгоритма Hist;
6: $\hat{x}_{T+1} = \hat{x}^{\text{ns}}_{T+1} + \hat{x}^{\text{s}}_{T+1}$;

6 Исследование свойств алгоритма Hist

В данном разделе описываются и исследуются свойства алгоритма Hist, основанного на свертке гистограммы с функцией потерь.

6.1 Используемые временные ряды

Для исследования свойств алгоритма Hist и последующего вычислительного эксперимента были использованы временны́е ряды [13–17], изображенные на рис. 1. Рассматриваемые ряды отличаются друг от друга длиной истории, наличием или отсутствием сезонности и тренда, диапазоном значений. После первого этапа алгоритма ARIMA + Hist получаются стационарные ряды регрессионных остатков. Именно на этих рядах и проводится исследование свойств Hist. Первый этап алгоритма и сравнение качества прогнозов описаны в следующем разделе.

6.2 Функции потерь

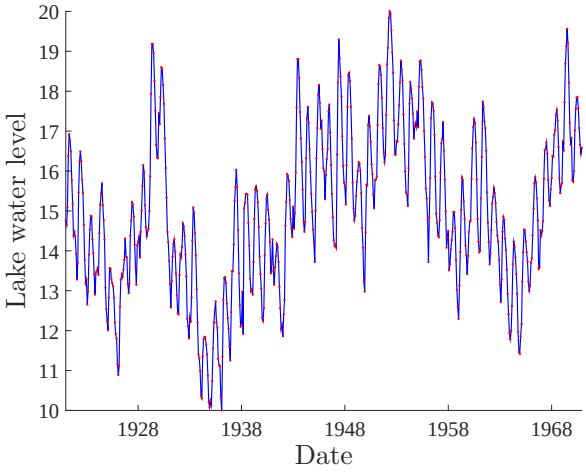
Эксперименты проводились для трех различных функций потерь:

$$\mathscr{L}_{\text{sq}}(\hat{x}, x) = (\hat{x} - x)^2\,; \tag{7}$$

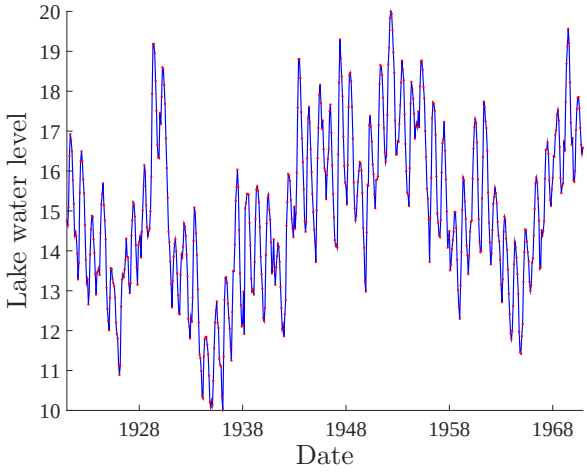
$$\mathscr{L}_{\text{abs}}(\hat{x}, x) = |\hat{x} - x|\,; \tag{8}$$

$$\mathscr{L}_{\text{asym}}(\hat{x}, x) = \begin{cases} \dfrac{1}{2}|\hat{x} - x|, & x \leqslant \hat{x}\,; \\ 2|\hat{x} - x|, & x > \hat{x}\,. \end{cases} \tag{9}$$

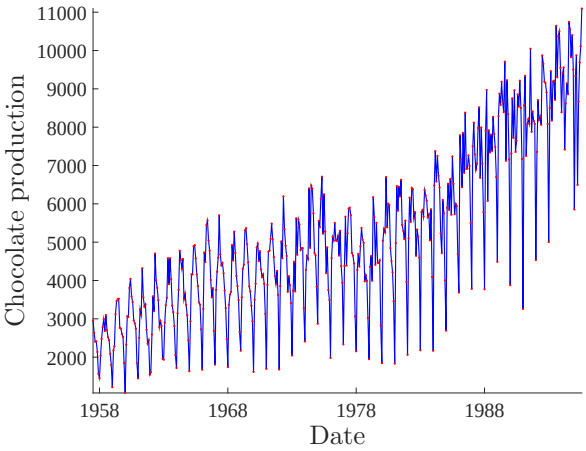
Графики квадратичной, абсолютной и ассимметричной функций потерь изображены на рис. 2. Все три функции выпуклые, достигают минимума при совпадении прогноза и действительного значения временно́го ряда. Первые две функции симметричные, последняя — несимметричная кусочно-линейная функция.



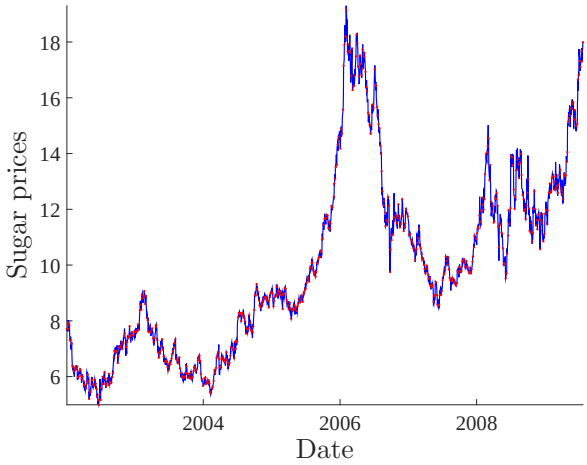
(*а*) Fraser River at hope



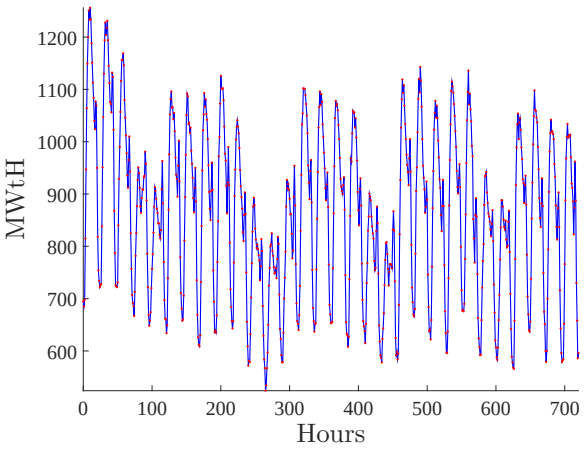
(*б*) Monthly Lake Erie Levels



(*в*) Monthly production of chocolate confectionery in Australia



(*г*) Sugar price



(*д*) Electricity consumption July 2005

Рис. 1 Временны́е ряды

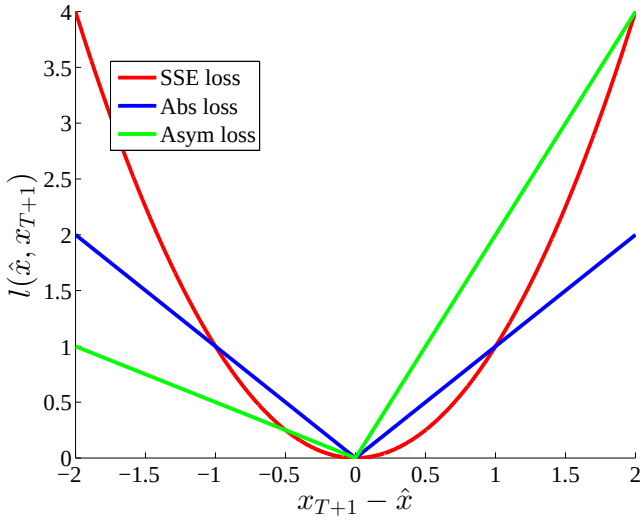


Рис. 2 Функции потерь

6.3 Свойства прогноза алгоритма Hist

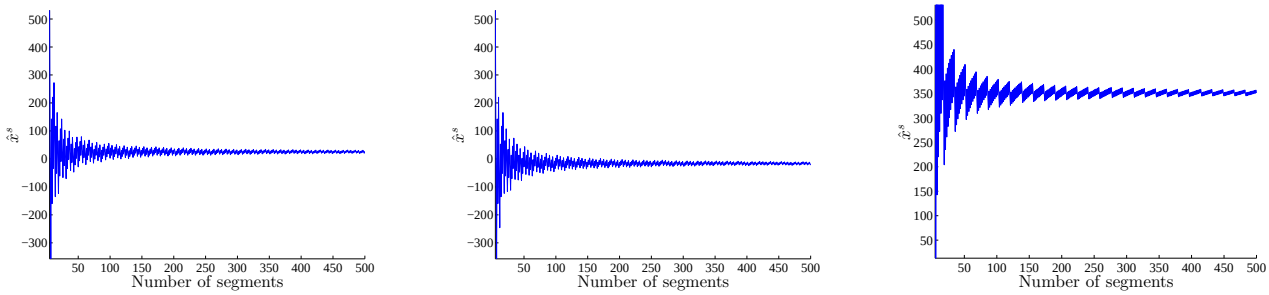
Чтобы определить, как зависит поведение алгоритма Hist от функции потерь и количества столбцов в гистограмме, для каждого ряда регрессионных остатков были построены графики зависимости прогноза алгоритма Hist от количества столбцов в гистограмме для каждой функции потерь (7)–(9). Графики изображены на рис. 3–7. На каждом графике по оси абсцисс отложено количество столбцов гистограммы, по оси ординат — прогноз, полученный алгоритмом Hist при использовании заданной функции потерь и гистограммы с заданным числом столбцов.

На рис. 3–7 видно, что для всех временны́х рядов и любой функции потерь с увеличением числа столбцов гистограммы полученные прогнозы стабилизируются вокруг предельного значения. Для симметричных функций потерь (7) и (8) предельное значение для прогнозов близко к нулю, что означает, что для симметричных функций потерь алгоритм Hist не дает существенной поправки к прогнозу нестационарной компоненты, полученному с помощью модели ARIMA. В то же время для несимметричной функции потерь (9) предельное значение прогнозов существенно больше нуля. Это значит, что суммарный прогноз будет значительно превышать прогноз нестационарной компоненты, поскольку рассматриваемая функция потерь (9) штрафует недопрогноз гораздо сильнее, чем перепрогноз.

Стабилизация прогнозов алгоритма Hist с увеличением количества столбцов в гистограмме связана с увеличением точности оценки плотности распределения регрессионных остатков $\gamma(u)$, о которой говорилось в разд. 4. Однако для конечных временны́х рядов добиться сходимости прогнозов к предельному значению с любой наперед заданной точностью невозможно из-за конечного количества доступных данных для оценки плотности распределения $\gamma(u)$.

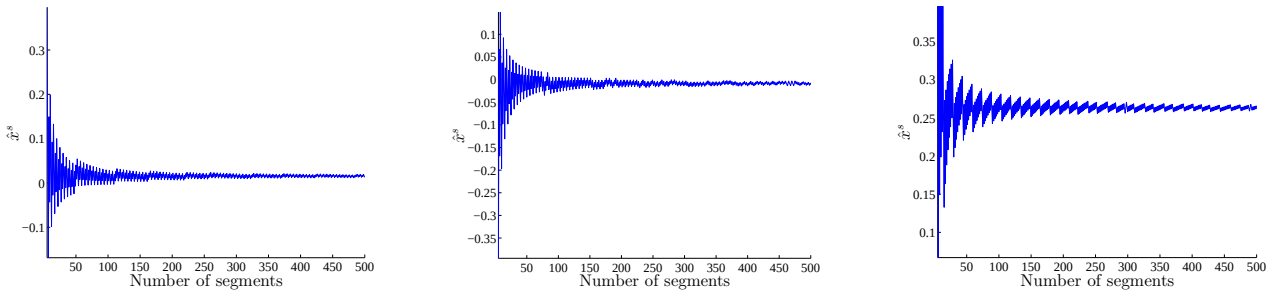
7 Вычислительный эксперимент

Целью проведенного вычислительного эксперимента является сравнение средних потерь прогнозирования различных временны́х рядов для различных функций потерь при использовании модели ARIMA и предложенного двухэтапного алгоритма ARIMA + Hist. Рассмотрены пять различных временны́х рядов [13–17], изображенных на рис. 1, и три



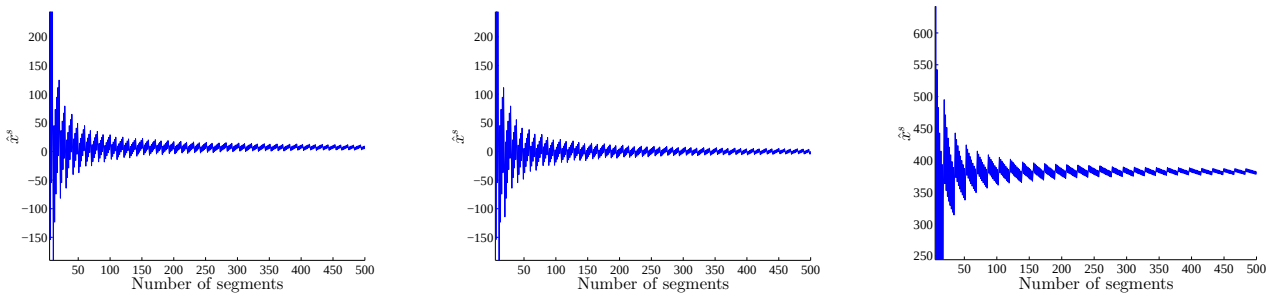
(*а*) Квадратичная функция потерь (*б*) Абсолютная функция потерь (*в*) Асимметричная функция потерь

Рис. 3 Прогнозы алгоритма Hist для регрессионных остатков ряда Fraser River at hope



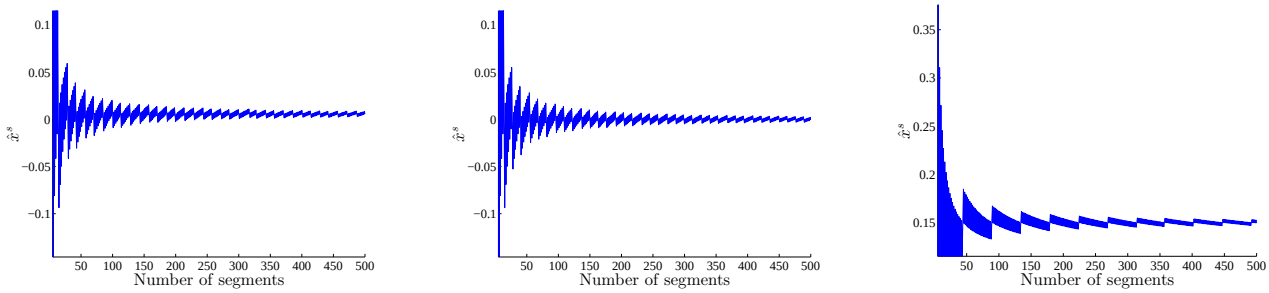
(*а*) Квадратичная функция потерь (*б*) Абсолютная функция потерь (*в*) Асимметричная функция потерь

Рис. 4 Прогнозы алгоритма Hist для регрессионных остатков ряда Monthly Lake Erie Levels



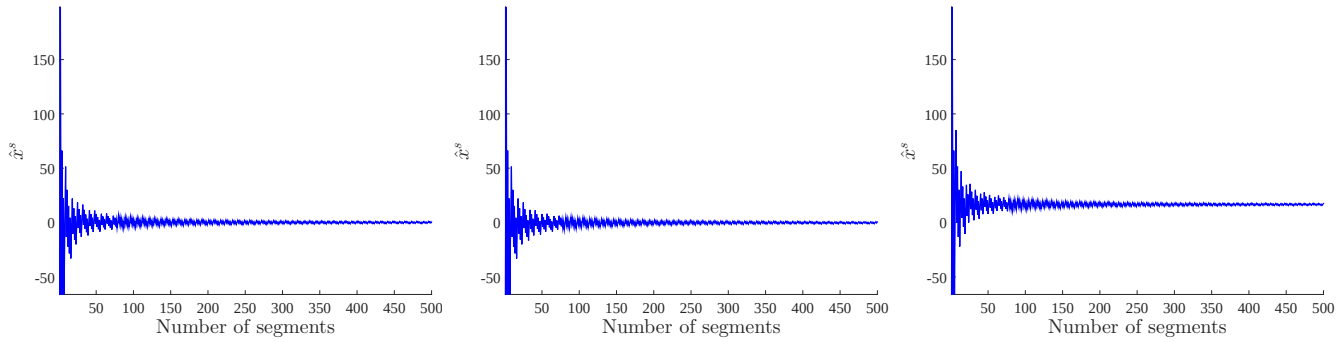
(*а*) Квадратичная функция потерь (*б*) Абсолютная функция потерь (*в*) Асимметричная функция потерь

Рис. 5 Прогнозы алгоритма Hist для регрессионных остатков ряда Monthly production of chocolate confectionery in Australia



(*а*) Квадратичная функция потерь (*б*) Абсолютная функция потерь (*в*) Асимметричная функция потерь

Рис. 6 Прогнозы алгоритма Hist для регрессионных остатков ряда Sugar price



(*а*) Квадратичная функция потерь

(*б*) Абсолютная функция потерь

(*в*) Асимметричная функция потерь

Рис. 7 Прогнозы алгоритма Hist для регрессионных остатков ряда Electricity consumption

функции потерь (7)–(9), одна из которых несимметричная. Получено экспериментальное подтверждение того, что при несимметричных потерях использование двухэтапного прогнозирования позволяет уменьшить средние потери.

Для каждого временного ряда, изображенного на рис. 1, подбиралась модель ARIMA по методологии Бокса–Дженкинса [2]. Выбранные модели для каждого временно́го ряда показаны в табл. 1.

Для прогнозирования стационарной компоненты используются регрессионные остатки алгоритма ARIMA.

Для сравнения качества прогнозов модели ARIMA и связки алгоритмов ARIMA + Hist 20% последних точек каждого временно́го ряда использовались как контрольные. Для каждой контрольной точки по доступной истории временно́го ряда (все точки от первой до предшествующей рассматриваемой контрольной) обучалась выбранная для временно́го ряда модель ARIMA, затем для обученной модели вычислялся ряд регрессионных остатков. По ряду регрессионных остатков обучался алгоритм Hist с заданной функцией потерь и заданным количеством столбцов в гистограмме. Прогноз для контрольной точки складывался из прогноза ARIMA и Hist. Эксперимент был проведен для функций потерь (7)–(9) и вариантов алгоритма Hist с 20, 50, 300 и 500 столбцами в гистограмме. Средние потери для каждой функции потерь приведены для всех вариантов алгоритма в табл. 2.

Как видно из табл. 2, при использовании асимметричной функции потерь двухэтапный алгоритм прогнозирования ARIMA + Hist позволяет получать среднюю ошибку прогноза существенно ниже, чем прогнозирование с помощью модели ARIMA. В большинстве случаев для симметричных функций потерь использование двухэтапного алгоритма прогнозирования не приводит к значительным изменениям по сравнению с прогнозом модели ARIMA. Исключение составляют только средние потери прогнозирования для временно́го ряда Monthly production of chocolate confectionery in Australia. Для этого временно́го ряда

Таблица 1 Выбранные модели ARIMA для прогноза нестационарной компоненты

Временной ряд	Модель ARIMA
Fraser River at hope	$\text{ARIMA}(1,0,0) \times (1,0,1)_{12}$
Monthly Lake Erie Levels	$\text{ARIMA}(2,0,0) \times (1,0,1)_{12}$
Chocolate production in Australia	$\text{ARIMA}(1,1,1) \times (1,0,1)_{12}$
Sugar price	$\text{ARIMA}(1,0,0)$
Electricity consumption	$\text{ARIMA}(2,1,2) \times (1,0,1)_{24} \times (1,0,1)_{168}$

Таблица 2 Средние потери прогнозирования

Ряд	Алгоритм	**Квадратичная**	**Абсолютная**	**Асимметричная**
River	No Hist	52400	498	616
	Hist(20)	53500	495	523
	Hist(50)	**52200**	496	516
	Hist(300)	52500	493	516
	Hist(500)	52400	**492**	**515**
Lake	No Hist	0,172	0,313	0,410
	Hist(20)	0,182	0,316	0,315
	Hist(50)	0,171	**0,313**	0,312
	Hist(300)	0,171	0,314	0,311
	Hist(500)	**0,171**	0,314	**0,311**
Chocolate	No Hist	71500000	8350	4180
	Hist(20)	66000	612	579
	Hist(50)	65800	609	575
	Hist(300)	65300	610	575
	Hist(500)	**65100**	**609**	**575**
Sugar	No Hist	0,127	0,265	0,340
	Hist(20)	0,128	0,267	**0,260**
	Hist(50)	0,127	0,266	0,267
	Hist(300)	0,127	0,265	0,266
	Hist(500)	**0,127**	**0,265**	0,266
Electricity	No Hist	500	16,9	19,9
	Hist(20)	717	19,1	**12,7**
	Hist(50)	589	**16,5**	13,4
	Hist(300)	**498**	17,1	13,0
	Hist(500)	502	16,9	13,0

использование двухэтапного алгоритма прогнозирования ARIMA + Hist привело к существенному уменьшению потерь для всех функций потерь. Это связано с тем, что, как видно на рис. 1, *в*, этот временной ряд во второй половине истории имеет более высокий темп роста, чем в первой половине. Обученная преимущественно по первой половине истории модель ARIMA дает прогнозы в контрольных точках, сильно смещенные в одну сторону относительно реальных значений. При использовании двухэтапного алгоритма прогнозирования ARIMA + Hist на втором шаге с помощью алгоритма Hist удается оценить это смещение и сделать более точный прогноз.

8 Заключение

Предложен двухэтапный алгоритм прогнозирования нестационарных временны́х рядов ARIMA + Hist, минимизирующий ожидаемые потери. Он не накладывает на вид функции потерь ограничений — она может быть симметричной или несимметричной, дифференцируемой или нет. На первом этапе строится прогноз нестационарной компоненты временно́го ряда путем выбора подходящей модели ARIMA. На втором этапе оценивается плотность распределения регрессионных остатков выбранной модели ARIMA и строится прогноз стационарной компоненты путем минимизации математического ожи-

дания потерь, которые заданы несимметричной функцией потерь. Финальный прогноз вычисляется как сумма прогнозов нестационарной и стационарной компоненты временно́го ряда. С помощью вычислительного эксперимента показано, что двухэтапное прогнозирование в случае несимметричной функции потерь позволяет уменьшить средние потери по сравнению с одноэтапным прогнозированием ARIMA. Также на практике средние потери можно уменьшить с помощью двухэтапного прогнозирования в случае симметричной функции потерь и смещенных прогнозов нестационарной компоненты временно́го ряда.

Литература

[1] Patton A. J., Timmermann A. Properties of optimal forecasts under asymmetric loss and nonlinearity // J. Econometrics, 2007. Vol. 140. No. 2. P. 884–918. doi: http://dx.doi.org/10.1016/j.jeconom.2006.07.018

[2] Box G. E. P., Jenkins G. M., Reinsel G. C. Time series analysis: Forecasting and control. — 3rd ed. — Englewood Cliffs, NJ: Prentice Hall, 1994.

[3] Berk R. Asymmetric loss functions for forecasting in criminal justice settings // J. Quantitative Criminology, 2011. Vol. 27. No. 1. P. 107–123. doi: http://dx.doi.org/10.1007/s10940-010-9098-2

[4] Cipra, T. Asymmetric recursive methods for time series // Appl. Math., 1994. Vol. 39. No. 3. P. 203–214.

[5] Koenker R., Xiao Z. Quantile autoregression // J. Am. Stat. Association, 2006. Vol. 101. No. 475. P. 980–990. doi: http://dx.doi.org/10.1198/016214506000000672

[6] Koenker R. Quantile regression. — Cambridge University Press, 2005. doi: http://dx.doi.org/10.1017/CBO9780511754098

[7] Christoffersen P. F, Diebold F. X. Optimal prediction under asymmetric loss // Econometric Theory, 1997. Vol. 13. No. 06. P. 808–817. doi: http://dx.doi.org/10.1017/S0266466600006277

[8] Granger C. W. J. Prediction with a generalized cost of error function // OR, 1969. Vol. 20. No. 2. P. 199–207. doi: http://dx.doi.org/10.2307/3008559

[9] Christoffersen P. F, Diebold F. X. Further results on forecasting and model selection under asymmetric loss // J. Appl. Econometrics, 1996. Vol. 11. No. 5. P. 561–571. doi: http://doi.org/bs5mh8

[10] Diebold F. X., Gunther T., Tay A. Evaluating density forecasts // Int. Econ. Rev., 1998. Vol. 39. P. 863–883. doi: http://dx.doi.org/10.2307/2527342

[11] Biau G., Bleakley k., Györfi L., Ottucsák G. Nonparametric sequential prediction of time series // J. Nonparametric Stat., 2010. Vol. 22. No. 3. P. 297–317.

[12] Hyndman R. J., Athanasopoulos G. Forecasting: Principles and practice. — OTexts, 2006. https://www.otexts.org/book/fpp.

[13] Monthly production of chocolate confectionery in Australia. https://datamarket.com/data/set/22rl/monthly-production-of-chocolate-confectionery-in-australia-tonnes-july-1957-aug-1995#!ds=22rl&display=line.

[14] Fraser River at hope. https://datamarket.com/data/set/22nm/fraser-river-at-hope-1913-1990#!ds=22nm&display=line.

[15] Monthly Lake Erie Levels. https://datamarket.com/data/set/22pw/monthly-lake-erie-levels-1921-1970#!ds=22pw&display=line.

[16] Sugar price. https://mlalgorithms.svn.sourceforge.net/svnroot/mlalgorithms/TSForecasting/TimeSeries/Sources/tsSugarPrice.csv.

[17] Electricity consumption. https://mlalgorithms.svn.sourceforge.net/svnroot/mlalgorithms/TSForecasting/TimeSeries/Sources/tsEnergyConsumption.csv.

Поступила в редакцию 30.08.15

Forecasting nonstationary time series under asymmetric loss*

***V. Y. Chernykh*[1] *and M. M. Stenina*[1,2]**

vladimir.chernykh@phystech.edu, mmedvednikova@gmail.com

[1]Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Russia

[2]Higher School of Economics, 22 Myasnitskaya st., Moscow, Russia

The problem of forecasting time series under asymmetric loss functions is considered. A new two-step forecasting algorithm ARIMA + Hist is presented. At the first step, autoregression integrated moving average algorithm ARIMA with seasonal components is used. The parameters of the model are selected according to Box–Jenkins methodology. At the second step, the residuals are analyzed and optimal addition to the forecast of the first step which minimize the expected value of losses is found. Expected loss is estimated by convolution of loss function with the histogram of regression residuals. The performance of the algorithm is demonstrated on time series with different types of nonstationarity (i. e., trend or seasonality) and for different symmetric and asymmetric loss functions. The results obtained during this experiment show that the quality of the forecast of two-step ARIMA+Hist exceed the quality of usual ARIMA in case of asymmetric loss functions.

Keywords: *forecasting; time series; nonstationary; ARIMA; convolution with loss function; asymmetric loss*

DOI: 10.21469/22233792.1.14.01

References

[1] Patton, A. J., and A. Timmermann. 2007. Properties of optimal forecasts under asymmetric loss and nonlinearity. *J. Econometrics* 140(2):884–918. doi: http://dx.doi.org/10.1016/j.jeconom.2006.07.018

[2] Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. *Time series analysis: Forecasting and control.* 3rd ed. Englewood Cliffs: NJ: Prentice Hall.

[3] Berk, R. 2011. Asymmetric loss functions for forecasting in criminal justice settings. *J. Quantitative Criminology* 27(1):107–123. doi: http://dx.doi.org/10.1007/s10940-010-9098-2

[4] Cipra, T. 1994. Asymmetric recursive methods for time series. *Appl. Math.* 39(3):203–214.

[5] Koenker, R., and Z. Xiao. 2006. Quantile autoregression. *J. Am. Stat. Association* 101(475):980–990. doi: http://dx.doi.org/10.1198/016214506000000672

[6] Koenker, R. 2005. *Quantile regression.* Cambridge University Press. doi: http://dx.doi.org/10.1017/CBO9780511754098

*This work was done under financial support of the Russian Foundation for Basic Research (grant 14-07-31046)

[7] Christoffersen, P. F, and F. X. Diebold. 1997. Optimal prediction under asymmetric loss. *Econometric Theory* 13(6):808–817. doi: http://dx.doi.org/10.1017/S0266466600006277

[8] Granger, C. W. J. 1969. Prediction with a generalized cost of error function. *OR* 20(2):199–207. doi: http://dx.doi.org/10.2307/3008559

[9] Christoffersen, P. F, and F. X. Diebold. 1996. Further results on forecasting and model selection under asymmetric loss. *J. Appl. Econometrics* 11(5):561–571. doi: http://doi.org/bs5mh8

[10] Diebold, F. X., T. Gunther, and A. Tay. 1998. Evaluating density forecasts. *Int. Econ. Rev.* 39:863–883. doi: http://dx.doi.org/10.2307/2527342

[11] Biau, G., K. Bleakley, L. Györfi, and G. Ottucsák. 2010. Nonparametric sequential prediction of time series. *J. Nonparametric Stat.* 22(3):297–317.

[12] Hyndman, R. J., and G. Athanasopoulos. 2006. *Forecasting: Principles and practice*. OTexts. Available at: https://www.otexts.org/book/fpp (accessed December 29, 2015).

[13] Monthly production of chocolate confectionery in Australia. Available at: https://datamarket.com/data/set/22rl/monthly-production-of-chocolate-confectionery-in-australia-tonnes-july-1957-aug-1995#!ds=22rl&display=line (accessed December 29, 2015).

[14] Fraser River at hope. Available at: https://datamarket.com/data/set/22nm/fraser-river-at-hope-1913-1990#!ds=22nm&display=line (accessed December 29, 2015).

[15] Monthly Lake Erie Levels. Available at: https://datamarket.com/data/set/22pw/monthly-lake-erie-levels-1921-1970#!ds=22pw&display=line (accessed December 29, 2015).

[16] Sugar price. Available at: https://mlalgorithms.svn.sourceforge.net/svnroot/mlalgorithms/TSForecasting/TimeSeries/Sources/tsSugarPrice.csv (accessed December 29, 2015).

[17] Electricity consumption. Available at: https://mlalgorithms.svn.sourceforge.net/svnroot/mlalgorithms/TSForecasting/TimeSeries/Sources/tsEnergyConsumption.csv (accessed December 29, 2015).

Received August 30, 2015

Using generalized precedents for big data sample compression at learning*

***V. V. Ryazanov*[1], *A. P. Vinogradov*[1], *and Yu. P. Laptin*[2]**

vngrccas@mail.ru

[1]Dorodnicyn Computing Centre of the Russian Academy of Sciences, 40 Vavilova st., Moscow, Russia

[2]Glushkov Institute of Cybernetics of the Ukrainian National Academy of Sciences, 40 Glushkova ave., Kiev, Ukraine

The role of intrinsic and introduced data structures at constructing efficient recognition algorithms is analyzed. The concept of generalized precedent as representation of stable local regularity in data and based on its use methods of reduction of the dimension of tasks has been investigated. Two new approaches to the problem based on positional data representation and on cluster means for elementary logical regularities are proposed. The results of computational experiment with data compression in parametric spaces for several practical tasks are presented.

Keywords: *generalized precedent; logical regularity; positional representation; bit slice; hypercube; correct decision rule*

DOI: 10.21469/22233792.1.14.02

1 Introduction

Methods of solving problems of recognition and data analysis use various structures in data. At a choice of appropriate structure, two main objectives are pursued:

a) identification natural clusters of density in the feature space in which the vectors of realizations are condensed; and
b) optimization of the computational expenses necessary for creation of the decision rule and subsequent calculations.

Both purposes are closely related with each other and at their realization compete for computational resources. For this reason, the overwhelming share of principles of structurization can be refereed to both directions simultaneously, and the choice of a concrete method in many respects is determined by assignment of priorities for (a) and (b). Now, the huge number of approaches, algorithms, and methods, more or less successful, are applied to achievement of both purposes. Note some survey publications on this subject [1, 2] where the most actual and perspective decisions are outlined. In them, both conceptual and technical aspects of the choice of the compromise are concerned.

In this paper, the close relationship which exists between the concepts 'precedent' and 'cluster' is investigated. The question how the mobility of the border between admissible realization of the concepts 'precedent' and 'cluster' in the computational environment can be used at the search of a compromise for (a) and (b) has been studied. Let the sum

$$F(x) = \sum_i \mu_i e^{-0.5(\boldsymbol{x_i}-\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\sigma}(\boldsymbol{x_i}-\boldsymbol{x})} \tag{1}$$

be parametrical approximation of empirical distribution by uniform normal mix with constant covariance matrix σ^{-1}. The component $\mathcal{N}(x_i, \sigma^{-1})$ represents compact spatial cluster C_i with

*This work was done under financial support of the Russian Foundation for Basic Research (grants 15-01-05776-a and 14-01-90413 Ukr_a).

the center x_i which is unambiguously described by the couple (x_i, μ_i). The natural treatment (1) implies that each cluster of C_i is filled by vectors corresponding to casual deviations from the parameters of the central object x_i. The recognized object x_0 can also be considered as a single realization of distribution of probable localizations of the true center which also form cluster C_0 with the center x_0 and with the same form of distribution $\mu_i e^{-0.5(\boldsymbol{x_i}-\boldsymbol{x})^{\top}\boldsymbol{\sigma}(\boldsymbol{x_i}-\boldsymbol{x})}$ where coordinates of the center x_0 and variable x interchange positions according to the Bayes's law. Thereby, internally inherent structure of the sample gains simple representation; however, this simplicity is reached at the price of creation of representation (1) as a solution of hard multiparametric inverse problem, and also with difficulties of reference of the cluster C_0 to one of the classes, each of which is represented by several clusters of type C_i. Certainly, the example is exaggerated, but it correctly reflects relationship between two concepts.

Opposite example in which injected structure of data appears, one can find in IP (Internet protocol) technologies where rigid hierarchy of clusters forcedly introduced into the R^2 plane in the form of quadtree provides high computational efficiency at training and recognition, but the hierarchy is thus invariable, and in orthodox approaches, it is not adjusted in any way to internal structure of the training sample [3]. The coordinates of clusters of quadtree are unambiguously fixed, and substantial information is coded by only the density of filling of clusters at different levels.

Further in this work, recognition problems will be considered in which the balance between the accuracy of representation and computational efficiency can be reached via structural reduction within the pair 'precedent–cluster.'

2 Generalized precedents: Feature space replacement

Application of models of type (1) assumes the use of Euclidean norm

$$\|x\| = \left(\sum_i x_i^2\right)^{1/2}$$

for estimation and comparison vectors in R^N. The norm binds together the values of different parameters, in particular, qualitatively incomparable ones. That is often convenient, but can cause questions at substantial interpretation of results.

On the contrary, for hierarchy of clusters in a quad- or oktree, the scales in different dimensions does not interact. In case of IP, it is the main drawback of quadtree-type models which limits their use [4]. Really, in case of images or scenes, it is usually assumed that spatial directions possess equal properties. At the same time, the models of this type are noninvariant to rotations in R^2 and R^3 and, therefore, the results achieved with their use are difficult to reproduce after rotation of the basis.

In abstract feature space, the assumption of 'equality axes in rights' takes place rarely. Moreover, invariance of a model to independent scaling of the main dimensions (in general, to independent nonlinear changes of scale on axes) becomes an important advantage. One of the successful approaches based on the use of this invariance is the approach with logical regularities [5–7]. In this approach, the clusters are hyperparallelepipeds in R^N, each cluster is described by conjunction of the following kind:

$$L^i = \&_n R_n^i, R_n^i = (A_n^i < x_n < B_n^i), \;\; n = 1, \ldots, N\,, \tag{2}$$

and substantively interpreted as a recurring joint manifestation of feature values $x_1, x_2, \ldots, x_N$ of the vector x at intervals $A_n^i < x_n < B_n^i, n = 1, \ldots, N$. The principle of proximity to each other

precedents of the same phenomenon here is embodied in the requirement of filling the interior of a certain type of cluster by the objects of the same class. The shape of clusters becomes of particular importance, and multiple joint appearance of feature values at the selected intervals in this approach is seen as an independent phenomenon called *elementary logical regularity*.

In all approaches mentioned above, just limited number of parameters is used to describe the spatial arrangement of the cluster and its filling. In case of quadtree, each cluster is encoded by one integer and one real parameter (q_i, μ_i); for the normal mixture (1), it is a pair of kind (x_i, μ_i); in case of logical regularities, it is a set of $2N$ border markers on axes A_n^i and B_n^i, $n = 1, \ldots, N$, and also, the weight of regularity μ_i.

Recently, V. V. Ryazanov has proposed the idea of reduction of dimension of the problem through the use of substantial clusters such as hyperparallelepiped or component $\mathcal{N}(x_i, \sigma^{-1})$ with significant aprioristic weight as new training objects. Each combined object is regarded as geometric manifestation of some separate regularity in initial data and is called *generalized precedent*. Such generalized precedents are just proposed to use in training. Generalized precedents are described by geometric parameters of corresponding clusters and dimensions of the new feature spaces in the above examples are $2, N + 1$, and $2N + 1$, respectively. Thus, dimension of the space of generalized precedents may change as the upward and downward, but big training sample receives more compact representation as the result.

3 Examples of usage of generalized precedents for sample reduction

3.1 Positional representation

In case of positional data representation, structural elements belong also to the special family of logical regularities of the 1st type (2), when real numbers are truncated to real ones, and the intervals used $(A_n^i < x_n < B_n^i), n = 1, \ldots, N$, are equal in length. Thus, hyperparallelepipeds become hypercubes of restricted variety of kinds.

Positional notation is the development of quadtree model in dimensions greater than 2. The main advantage is that the structuring of positional hierarchy is already automatically injected into any numerical data when registering them, and it is immediately ready for use. It was also noted above that in models of this type, independent scaling of the main axes is naturally implemented, and this fact makes prospects of using the proposed approach in a variety of recognition problems, including the ones with incomparable numerical features.

Let finite sets X_k are preset in R^N and represent classes k, $k = 1, \ldots, K$, of the training sample X.

Positional representation [8] of data in R^N is defined by a bit grid $D^N \subseteq R^N$ where $|D| = 2^d$ for some integer d.

The parameter d is not fixed in advance. As it will be shown, its value is determined by the results of the analysis of the mutual arrangement of classes in the training sample.

Each grid point $x_1, x_2, \ldots, x_N$, $n = 1, \ldots, N$, corresponds to effectively performed transformation on bit slices in D^N, when the mth bit in binary representation $x_n \in D$ of the nth coordinate of x becomes $p(n)$-bit of binary representation of the mth digit of 2^N-ary number that represents vector x as whole. Here, it is supposed $0 < m \leqslant d$, and function $p(n)$ defines a permutation on $1, \ldots, N, p \in S_N$. The result is a linearly ordered scale S of length 2^{dN}, representing one-to-one all the points of the grid in the form of a curve that fills the space D^N densely. For chosen grid D^N, an exact solution of the problem of recognition with K classes results in K-valued function f defined on the scale S. As known, m-digit in 2^N-ary positional representation corresponds to n-dimensional cube of volume $2^{N(m-1)}$. It is called m-point. For each m, the entire set of m-points is called m-slice. Thus, one has

Lemma 1. *There are just one d-point, 2^N ones of $(d-1)$-points, and 2^{dN} ones of 1-points on the scale S.*

Each of m-points, $0 < m \leqslant d$, can be regarded as separate cluster in D^N. If it is nonempty and filled with data of certain class only, one has got generalized precedent.

Further, for every $k, k = 1, \ldots, K$, and every $m, 0 < m \leqslant d$, let us look for the set of all of m-points, which are generalized precedents, i. e., elementary logical regularities of class k. The larger uniform regions in the domain of function f (corresponding to generalized precedents as elder m-points), the better the decision rule. In the description of positional generalized precedent, the filled volume is represented latently by parameter m (i. e., by the level in the hierarchy) and actual new feature space is formed of pair (p_i, m_i).

Here, let describe the scheme of algorithm **A** that realizes this search on hierarchy of m-points of the grid D^N from top to bottom.

The search is carried out for all classes k, $k = 1, \ldots, K$, simultaneously. Data of the training sample $X = \bigcup X_k \subseteq R^N$ are transformed into 2^N-ary indices of the grid D^N.

All objects of the sample are processed in turn. Each next object x marks all m-points, $m > 1$, of the own branch in hierarchy D^N with the index k. Notice that for $m > 1$, there are no more than $\sum_{m=2}^{d} 2^{N(d-m)}$ different m-points. For dimensions $N > 3$, this number is negligible in comparison with the total number of 1-points of the grid D^N, and this fact provides the mechanism of compression of the sample.

Upon termination of search in each marked point of hierarchy D^N, the final attributing is carried out: if some $(m+1)$-point was marked with indexes of various classes (i. e., is not the generalized precedent), and all m-points subordinated to it are the generalized precedents, then the entire last are included in the decision rule. Further specification and attributing of subordinated $(m-1)$-points are not required.

As for all classes $k, k = 1, \ldots, K$, the analysis began with the same d-point as the top of hierarchy, one has

Lemma 2. *Algorithm **A** finds all generalized precedents of specified kind in the training sample $X = \bigcup X_k \subseteq R^N$.*

Since the number of m-points is final, any m-point that hashes classes at actual choice of the parameter d, will be further resolved by next iteration of algorithm **A** under this m-point regarded as new top and, thus, one has got

Lemma 3. *Iterative process on the basis of algorithm **A** provides creation of exact decision rule that is correct on the training sample $X = \bigcup X_k \subseteq R^N$.*

Thus, one has to decide what is better in this or that case: big d or many iterations of **A**.

Since data of training sample $X = \bigcup X_k \subseteq R^N$ are analyzed consecutively, further retraining of any recognition algorithm constructed on this way will demand investigation of objects not more than inside one generalized precedent for each new object.

When sets of generalized precedents for all classes are built, one can combine within each class some collected m-points as hypercubes in larger hyperparallelepipeds according to criteria of contiguity [7]. So, one more way of building space of generalized precedents taking the form of elementary logical regularities of the 1st kind can be realized.

3.2 Cluster means as generalized precedents

Generally, large number of various ways of creation logical regularities of the 1st kind is developed now, and the choice of one of them is determined by properties of data and the character

of a problem of recognition or forecasting [5]. As the second example of use of generalized precedents, here, a new method of compression of data is considered which consists in transformation of feature space R^N to the space (c^i, μ^i) with dimension $N+1$. Class means c^i in clusters of regularities $L^i = \&_n R^i_n, R^i_n = (A^i_n < x_n < B^i_n), n = 1, \ldots, N$, are used as generalized precedents in this method. The space $(A^i_n, B^i_n, \mu^i), n = 1, \ldots, N$, itself is used thus at the intermediate stage.

Notice that in hierarchy of m-points of D^N, very rigid criterion of selection of generalized precedents was applied. Existence of the only object of alien class as a part of any m-point (hypercube of large volume, when m is close to d) excludes the last from among the generalized precedents and strongly reduces thereby potential efficiency of compression of the sample. For this reason, in the majority of methods of creation of logical regularities, softer selection criteria are used when existence of certain share of objects of others classes as a part of this or that hyperparallelepiped (corresponding to elementary logical regularity) is allowed. Thus, flexibility of the model of logical regularities in general is reached and possibility of creation of simple decision rule with small set of elements of the sort $L^i = \&_n R^i_n, R^i_n = (A^i_n < x_n < B^i_n)$ is provided where each of them represents essential part of the training sample.

The proposed method of compression uses the specified opportunity fully, but realizes also a way of disposal of the difficulties related with the existence of alien objects in the cluster of regularity $L^i = \&_n R^i_n, R^i_n = (A^i_n < x_n < B^i_n)$. Let $x^L_t, t = 1, \ldots, T^L$, be a set of objects of the kth class as a part of cluster of elementary logical regularity L. Construct in the space R^{N+1} a new sample that is made of vectors of averages $c^L = \sum_t^{T^L} x^L_t$, and their shares T^L in each regularity L. Thereby, the space of the generalized precedents (c^i, T^L) is set, each point of which corresponds to nonuniformly filled cluster of initial space R^N where the objects of class k dominate. The role of cluster geometry thus partially loses its value, important is only that the share T^L of objects of the kth class within the cluster is big.

4 Reconstruction of the decision rule in initial feature space

Reconstruction of the decision rule in initial feature space consists in the return replacement of sets of the essential generalized precedents with clusters of the chosen for them geometric forms. Replacement is carried out directly and does not cause difficulties. In Fig. 1, it is shown how it takes place in case of hypercubes of positional data representation. For cluster means, this transition is even more direct since in this case, the space of generalized precedents differs from initial feature space only in additional equipment of weight coordinate T^L.

5 Computational experiment

Computational experiment in the framework of this training sample compression model was made for several types of generalized precedents on real tasks. The best accuracy was achieved by approach on the basis of cluster means. Here, the generalized precedents are used for representation of the training sample in the form of sets of new precedents that match as the source precedents and classes and the results of analysis of the initial training sample. As additional information for each class $K_\lambda, \lambda = 1, 2, \ldots, l$, there are used multiple logical regularities of classes $P_\lambda = \{P_t(\mathbf{x})\}$, i. e., predicates of the form

$$P^{\Omega_1, \mathbf{c}^1, \Omega_2, \mathbf{c}^2}(\mathbf{x}) = \bigwedge_{j \in \Omega_1} (c^1_j \leqslant x_j) \bigwedge_{j \in \Omega_2} (c^1_j \geqslant x_j),$$

$$\Omega_1, \Omega_2 \subseteq 1, 2, \ldots, n, \mathbf{c}^1, \mathbf{c}^2 \in R^n,$$

where

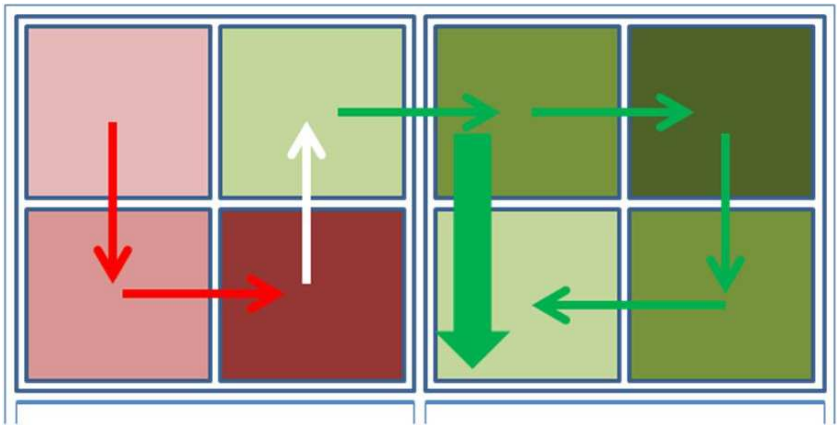


Figure 1 Fragment of quadtree with the scale S of Peano-type. Two classes (red and green) are separated in S by white arrows. Color intensity depicts the density of filling. Left big square represents an m-point that hashes classes. Right square is filled with objects of the green class only detected in all subordinated $(m-1)$-points and so, this m-point represents generalized precedent as large uniform region included in the decision rule (big green arrow)

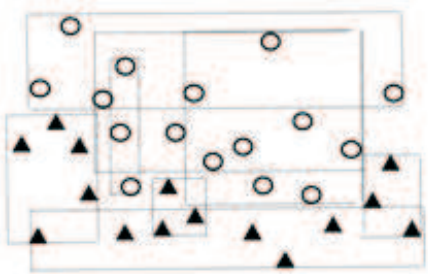


Figure 2 Intervals of two-dimensional regularities of two classes are the marked boxes

1) $\exists \mathbf{x}_t \in K_2^0 | P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_t) = 1$;
2) $\forall \mathbf{x}_t \neg \in K_2^0 | P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_t) = 0$; and
3) $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_t)$ represents a local optimum of the standard criterion of predicates' quality.

Here, through $\tilde{K}_\lambda$, the training sample objects from class λ are designated. Two schemes of definition of generalized predicates are used.

In the first scheme, sets of objects that satisfy the predicates of P_λ correspond to the set $\tilde{K}_\lambda$. Figure 2 shows a model example. An analog of the "nearest neighbor" algorithm was used.

Object $\mathbf{x}$ is assigned to the class, the regularity of which is considered the closest, the "distance" to the patterns is calculated by the formula:

$$d_\alpha(\mathbf{x}) = \frac{\sum_{\mathbf{x}_t : P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_t)=1} \rho(\mathbf{x},\mathbf{x}_t)}{|\{\mathbf{x}_t : P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_t) = 1\}|}$$

where ρ is the Euclidean metric in R^n.

Comparison was carried out on the data of the credit scoring (2 classes, 15 features, 348 test objects) [8]. The accuracy of the standard and the modified method of "nearest neighbor" was on the test data, respectively, 75.6% and 77.5% of correct answers.

In the second scheme, generalized precedent is considered as the set of values of all logical regularities of the object, disjunction of their negations, the set of values of all logical regularities of another class, and disjunction of their negations (classification with 3 classes and more used

Figure 3 Visualization of the original training sample (a) and of the sample of generalized precedents in the parametric space where classes become linearly separable (b)

scheme "one against all"). Thus, each object corresponds to a vector of numbers $\{0, 1\}$, and the generalized precedent is simply a description of the object in the new feature space. Figure 3 shows the visualization of the original training sample and the sample in new parametric space on the task of recognition of breast cancer [9]. The objects of different classes are presented in a plane gray and black circles. Generalized precedents of the training sample are linearly separable.

The version of support vector machine implemented in [5] was used as the main classification method. The results of the comparison of methods of recognition of test data on various tasks are presented in Table 1.

In general, the achieved positive results testify to prospects of the approach and to need of further development of this direction of researches.

6 Concluding remarks

The use of some inherent and injected structures in data has been considered. The opportunities arising from the use of generalized precedents for creation of detailed decision rule have been analyzed. It was shown that in case of positional data representation, the feature space R^N can be reduced to two-dimensional space where training data become represented by compact clusters. Reduced representation realizes the one-dimensional scan of R^N, which is loaded with weights of generalized precedents. A scheme for an iterative process is proposed that yields to construct exact solutions which are correct on the training data. A new method of training

Table 1 Results of comparison of recognition methods on various tasks

Task	Classes	Dimension	Objects	Reference objects	Accuracy on reference objects	Accuracy on generalized precedents
"Breast"	2	9	344	355	94.6 (0.8)	96.1
"Credit"	2	15	342	348	80.5 (4.3)	64.5
"Image"	7	16	210	2100	68.8 (27.7)	92.0 (0.6)

data compression has been developed and investigated based on the use of cluster means for elementary logical regularities and on its use as generalized precedents in transformed $(N+1)$-dimensional feature space. Computational experiment was made for several types of generalized precedents on real tasks. Good results approve the new opportunities and open prospects of the use of generalized precedents in recognition tasks with big data samples.

References

[1] De Berg, M., M. van Kreveld, M. O. Overmars, and O. Schwarzkopf. 2000. *Computational geometry: Algorithms and applications*. 2nd ed. Springer. 291–306. doi: `http://dx.doi.org/10.1007/978-3-662-04245-8`

[2] Berman, J. 2013. *Principles of big data*. Elsevier. 1–14.

[3] Samet, H., and R. Webber. 1985. Storing a collection of polygons using quadtrees. *ACM Trans. Graph.* 4(3):182–222. doi: `http://dx.doi.org/10.1145/282957.282966`

[4] Eberhardt, H., V. Klumpp, and U. D. Hanebeck. 2010. Density trees for efficient nonlinear state estimation. *13th Conference (International) on Information Fusion Proceedings*. Edinburgh. 1–8. doi: `http://dx.doi.org/10.1109/ICIF.2010.5712086`

[5] Zhuravlev, Yu. I., V. V. Ryazanov, and O. V. Senko. 2006. *Raspoznavanie. Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya*. Moscow: FAZIS. 168 p. (In Russian.)

[6] Ryazanov, V. V. 2007. Logicheskie zakonomernosti v zadachakh raspoznavaniya (parametricheskiy podkhod). Zhurnal vychislitelnoy matematiki i matematicheskoy fiziki 47(10):1793–1809. (In Russian.)

[7] Vinogradov, A., and Yu. Laptin. 2010. Usage of positional representation in tasks of revealing logical regularities. *VISIGRAPP-2010, Workshop IMTA-3 Proceedings*. Angers. 100–104.

[8] Aleksandrov, V. V., and N. D. Gorskiy. 1983. *Algoritmy i programmy strukturnogo metoda obrabotki dannykh*. Leningrad: Nauka. 208 p. (In Russian.)

[9] Mangasarian, O. L., and W. H. Wolberg. 1990. Cancer diagnosis via linear programming. *SIAM News* 23(5):1–18.

Received June 15, 2015

Использование обобщенных прецедентов для сжатия больших выборок при обучении*

***В. В. Рязанов*[1], *А. П. Виноградов*[1], *Ю. П. Лаптин*[2]**

vngrccas@mail.ru

[1]Вычислительный центр РАН им. А.А.Дородницына, Россия, г. Москва, ул. Вавилова, 40

[2]Институт кибернетики им. В. М. Глушкова Национальной академии наук Украины, Украина, г. Киев, пр. Ак. Глушкова, 40

Анализируется роль внутренне присущих и привнесенных структур данных при построении эффективных алгоритмов распознавания. Исследуется понятие обобщенного прецедента как способа представления устойчивой локальной закономерности в данных и методы снижения размерности задач на основе его использования. Предложены два новых подхода к проблеме, основанные на позиционном представлении и на средних по кластерам элементарных логических закономерностей. Представлены результаты вычислительного эксперимента по сжатию данных в параметрических пространствах для нескольких практических задач.

Ключевые слова: *обобщенный прецедент; логическая закономерность; позиционное представление; битовый слой; гиперкуб; корректное решающее правило*

DOI: 10.21469/22233792.1.14.02

Литература

[1] *De Berg M., van Kreveld M., Overmars M. O. Schwarzkopf O.* Computational geometry: Algorithms and applications. — 2nd ed. — Springer, 2000. P. 291–306. doi: http://dx.doi.org/10.1007/978-3-662-04245-8

[2] *Berman J.* Principles of big data. — Elsevier, 2003. P. 1–14.

[3] *Samet H., Webber R.* Storing a collection of polygons using quadtrees // ACM Trans. Graph., 1985. Vol. 4. Iss. 3. P. 182–222. doi: http://dx.doi.org/10.1145/282957.282966

[4] *Eberhardt H., Klumpp V., Hanebeck U. D.* Density trees for efficient nonlinear state estimation // 13th Conference (International) on Information Fusion Proceedings. Edinburgh, 2010. doi: http://dx.doi.org/10.1109/ICIF.2010.5712086

[5] *Журавлев Ю. И., Рязанов В. В., Сенко О. В.* Распознавание. Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. 168 с.

[6] *Рязанов В. В.* Логические закономерности в задачах распознавания (параметрический подход) // Ж. вычислительной математики и математической физики, 2007. Т. 47. № 10. С. 1793–1809.

[7] *Vinogradov A., Laptin Yu.* Usage of positional representation in tasks of revealing logical regularities // VISIGRAPP-2010, Workshop IMTA-3 Proceedings. Angers, 2010. P. 100–104.

[8] *Александров В. В., Горский Н. Д.* Алгоритмы и программы структурного метода обработки данных. — Л.: Наука, 1983. 208 с.

[9] *Mangasarian O. L., Wolberg W. H.* Cancer diagnosis via linear programming // SIAM News, 1990. Vol. 23. No. 5. P. 1–18.

Поступила в редакцию 15.06.15

*Работа выполнена при финансовой поддержке РФФИ (проекты №№15-01-05776-а и 14-01-90413 Укр_а)

Определение видимой области радужки классификатором локальных текстурных признаков*

И. А. Соломатин, И. А. Матвеев
ivan.solomatin@phystech.edu
Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9
Вычислительный центр РАН им. А.А.Дородницына, Россия, г. Москва, ул. Вавилова, 40

Распознавание человека по изображению радужной оболочки — актуальная задача в биометрических системах. Помимо выделения радужки как кольцевой области для повышения точности распознавания определяют области затенения (блики, веки, ресницы и т. д.). Задача выделения затенений радужки может быть поставлена как классификация пикселей кольцевой области на два класса: «радужка» и «затенение». В кольцевой области определяется сектор с минимальной дисперсией яркости, который, как правило, не содержит затенений (в данной работе этот сектор не вычисляется, а рассматривается как часть входных данных алгоритма). Далее строится классификатор на основе многомерного гауссиана, который обучается на выборке, задаваемой по пикселям этого сектора. Параметры классификатора были оптимизированы с помощью генетического алгоритма. Проблема шума и ошибок с классификацией некоторых участков изображения решается с помощью применения морфологической постобработки. Был проведен вычислительный эксперимент, и получено распределение функционала качества алгоритма.

Ключевые слова: *обработка изображений; распознавание образов; биометрическая идентификация по радужке*

DOI: 10.21469/22233792.1.14.03

1 Введение

Необходимый этап биометрической идентификации по радужной оболочке — сегментация радужки, т. е. выделение области изображения, содержащей только радужку. В сегментацию помимо выделения окружностей, аппроксимирующих внешнюю и внутреннюю границы радужной оболочки, входит выделение областей затенения. Последняя задача достаточно важна для распознавания человека по радужке, так как ее решение задает конечную эффективную область для распознавания. В англоязычной литературе эта задача называется EES (Eyelids, Eyelashes, Shadows) localization. Существуют различные подходы к ее решению. В [1] осуществляется выделение области век с использованием интегрально-дифференциальных операторов, граница века аппроксимируется некоторой кривой. В [2] для локализации век используется преобразование Хафа. В [3] граница века аппроксимируется параболой. В [4] для обнаружения затенений используется мера качества радужки, вводится вероятностная мера качества радужки, основанная на смеси многомерных гауссианов, которая сравнивается с мерой, основанной на преобразовании Фурье. Для распознавания ресниц в [5] применяется статистическая оценка. В [6] описан метод распознавания век, использующий горизонтальный медианный фильтр и распознавание границ. В [7] для распознавания ресниц используется оператор Собеля, затем ресницы удаляются с изображения с помощью медианного фильтра. Существуют решения, выделяющие все затенения сразу (без разделения на веки, ресницы, блики и т. п.). Одно из таких решений описано в [8], в нем применяется классификатор, основанный на смесях

*Работа выполнена при финансовой поддержке РФФИ, проект № 15-01-05552.

гауссианов, разделяющий точки на два класса: «радужка» и «затенение». Для обучения классификатора используются изображения, размеченные вручную.

Недостатками метода [8] являются необходимость в ручной разметке и использование одного и того же классификатора для всех изображений. Единый классификатор для всех изображений является сомнительным решением вследствие того, что изображения разных радужных оболочек могут иметь значительные вариации текстурных признаков (разная структура, разное освещение и т. д.), что при обучении на большой выборке порождает очень широкий класс «радужка», обладающий малой разделяющей способностью, а при обучении на малой выборке — узкий класс, не пригодный для многих изображений.

В данной работе предлагается использовать классификатор на основе многомерного гауссиана, обучающийся на незатененном секторе S. Использование сектора S для обучения позволяет избежать использования обучающей выборки, размеченной вручную, т. е. построить полностью автоматический метод. Таким образом, содержание работы — реализация алгоритма поиска областей затенения с использованием сектора S вместо ручной разметки. В качестве сектора S используется сектор с минимальной дисперсией яркости с фиксированным углом раствора. Важно заметить, что в данной работе прежде всего ставился вопрос о принципиальной работоспособности метода распознавания затенений на основе классификатора, обучающегося на элементе того же изображения. Вопрос максимальной оптимизации пока не ставился. В дальнейшем, в частности, будет подробнее изучен вопрос нахождения сектора S и вопрос выбора признаков, по которым строится классификатор.

2 Постановка задачи

На вход подается черно-белое изображение I, которое является прямоугольной матрицей $W \times H$ из целых беззнаковых однобайтовых чисел $I(i,j) \in [0;255]$, задающих яркость каждого пикселя изображения. Также на вход подаются координаты центров и радиусы двух окружностей, аппроксимирующих внешнюю и внутреннюю границу радужки (вычисленные, например, методами [9]), и незатененный сектор S, задаваемый серединным углом α и раствором $\Delta\alpha$.

Требуется выделить области затенения, т. е. области изображения, на которых изображена не радужка, а веко, ресница, блик и т. п. Строго говоря, нужно классифицировать все точки изображения, лежащие внутри кольца, аппроксимирующего радужку по двум классам — «радужка» и «затенение», т. е. получить бинарное изображение J:

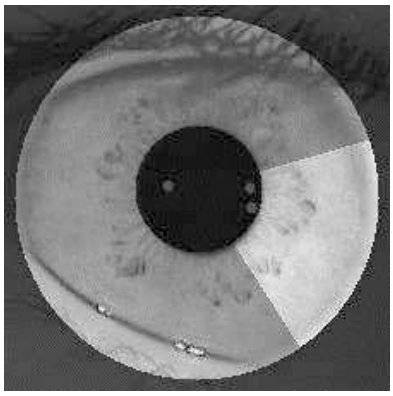
$$J(i,j) \in \{0,1\}; \quad i \in [1;H]; \;\; j \in [1,W], \tag{1}$$

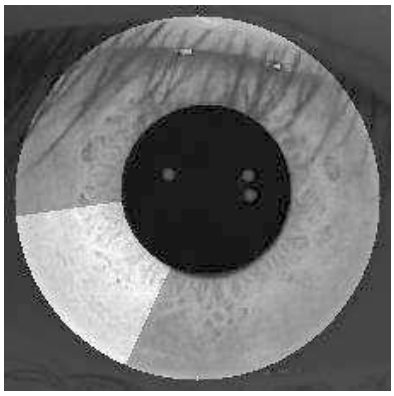
где $J(i,j) = 1 \Leftrightarrow$ точка $(i,j)^{\mathrm{T}}$ классифицирована как затенение.

3 Метод решения

Рассмотрим точки, находящиеся внутри сектора $S = \{\boldsymbol{x_n}\}_{n=1}^{N}$, где N — число точек, принадлежащих S. Локальные текстурные признаки этих точек будут являться обучающей выборкой для классификатора. На рис. 1 приведены примеры сектора S для различных изображений радужки.

Стоит заметить, что в качестве сектора S выбирается сектор с фиксированым раствором $\Delta\alpha$ такой, что его дисперсия минимальна, и, строго говоря, этот сектор может содержать затенения. В таких случаях точность распознавания не может быть высокой. Пример того, как наличие затенений в секторе S влияет на точность распознавания, рассмотрен в разд. 5 (пример 3, рис. 8). В данном примере сектор S содержит ресницы; таким


(*а*) S не содержит затенений


(*б*) S не содержит затенений

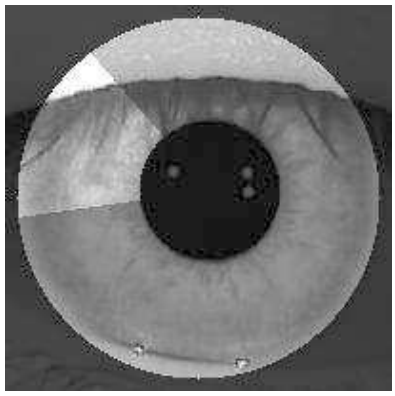

(*в*) S содержит затенения

Рис. 1 Примеры секторов с минимальной дисперсией яркости (выделены светлым)

образом, классификатор, обученный по этому сектору, не считает ресницы затенениями, что уменьшает точность. В дальнейшем планируется написать более точный алгоритм поиска незатененного сектора радужки.

В качестве локальных текстурных признаков выбираются следующие $K = 8$ признаков:

1. $B(\boldsymbol{x_n})$ — яркость в точке $\boldsymbol{x_n} = (x, y)^{\mathrm{T}}$:
$$B(\boldsymbol{x_n}) = I(x, y)\,.$$
2. $\overline{B}(\boldsymbol{x_n})$ — средняя яркость в окрестности точки $\boldsymbol{x_n}$ (в окне размера a).
В качестве окна размера a будем использовать квадрат со стороной $2a + 1$. Тогда количество точек в окне размера a равно $(2a + 1)^2$.
Обозначим момент яркости в окрестности размера a точки $\boldsymbol{x_n}$:
$$M_a^{(z)} = \frac{1}{(2a+1)^2} \sum_{i=-a}^{a} \sum_{j=-a}^{a} I^z(x + i, y + j).$$
Тогда
$$\overline{B}(\boldsymbol{x_n}) = M_a^{(1)}. \tag{2}$$
3. $\sigma(\boldsymbol{x_n})$ — среднеквадратичное отклонение яркости в окрестности точки $\boldsymbol{x_n}$ (в окне размера b):
$$\sigma(\boldsymbol{x_n}) = \sqrt{M_b^{(2)} - \left(M_b^{(1)}\right)^2}\,. \tag{3}$$
4. $\boldsymbol{C}(\boldsymbol{x_n})$ — вектор из пяти компонент дискретного косинусного преобразования в окрестности точки $\boldsymbol{x_n}$.
Как правило, дискретное косинусное преобразование считается в окне $w\times h$, где $w = h = 8$, но так как в данном случае, из соображений симметрии, желательно, чтобы окно было симметрично центрированным, используется окно 7×7 $(w = h = 7)$. Двумерное дискретное косинусное преобразование в окне $w\times h$ при $w = h = 7$ с центром в точке $(x, y)^{\mathrm{T}}$:
$$\widehat{C}_{p,q}(x, y) = \alpha_p \alpha_q \sum_{i=-3}^{3} \sum_{j=-3}^{3} I(x + i, y + j) \cos\frac{\pi(2i + 7)p}{2h} \cos\frac{\pi(2j + 7)q}{2w},$$
$$0 \leqslant p \leqslant h - 1; \quad 0 \leqslant q \leqslant w - 1,$$

где

$$\alpha_p = \begin{cases} \dfrac{1}{\sqrt{w}}, p = 0\,; \\ \dfrac{\sqrt{2}}{\sqrt{w}}, 1 \leqslant p \leqslant h-1\,; \end{cases} \qquad \alpha_q = \begin{cases} \dfrac{1}{\sqrt{h}}, q = 0\,; \\ \dfrac{\sqrt{2}}{\sqrt{h}}, 1 \leqslant q \leqslant w-1\,. \end{cases}$$

Из полученной матрицы коэффициентов 7×7 берутся первые 5 коэффициентов за исключением $\widehat{C}_{0,0}$ (рис. 2), так как коэффициент $\widehat{C}_{0,0}$ равен средней яркости в окрестности, которая уже используется в качестве признака:

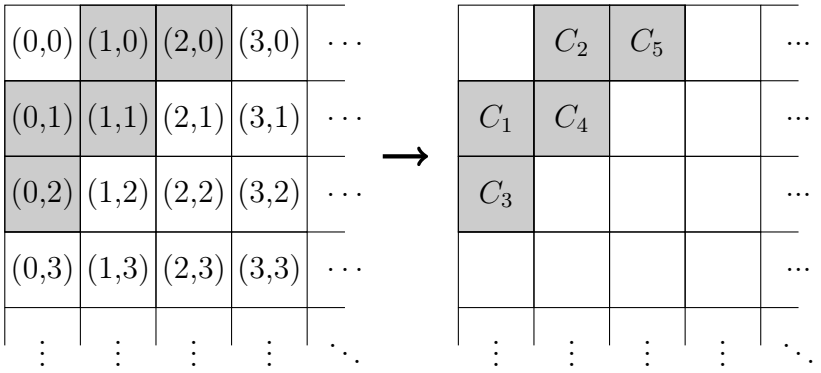


Рис. 2 Используемые коэффициенты дискретного косинусного преобразования

$$\boldsymbol{C}(\boldsymbol{x_n}) = (\widehat{C}_{0,1}, \widehat{C}_{1,0}, \widehat{C}_{0,2}, \widehat{C}_{1,1}, \widehat{C}_{2,0})^{\mathrm{T}}.$$

В формулах (2) и (3) размеры окрестностей a и b выступают в качестве параметров алгоритма.

Каждой точке сектора $S : \boldsymbol{x_n} = (x, y)^{\mathrm{T}}$ сопоставляем вектор из признаков:

$$\boldsymbol{p_n} = (B(\boldsymbol{x_n}), \overline{B}(\boldsymbol{x_n}), \sigma(\boldsymbol{x_n}), (\boldsymbol{C}(\boldsymbol{x_n}))^{\mathrm{T}})^{\mathrm{T}}\,.$$

Находим средний вектор:

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{p_n}\,; \qquad \hat{\boldsymbol{p}}_{\boldsymbol{n}} = \boldsymbol{p_n} - \boldsymbol{\mu}\,.$$

Из векторов $\hat{\boldsymbol{p}}_{\boldsymbol{n}}$ составляется матрица объект-признак M размерности $K \times N$:

$$M = \begin{pmatrix} \hat{\boldsymbol{p}}_{\boldsymbol{1}} & \hat{\boldsymbol{p}}_{\boldsymbol{2}} & \cdots & \hat{\boldsymbol{p}}_{\boldsymbol{n}} \end{pmatrix}.$$

Предполагаем, что объекты в пространстве, задаваемом данными параметрами, распределены по многомерному гауссиану. Гипотезу о принадлежности каждого пикселя классу «радужка» принимаем или отвергаем, основываясь на расстоянии Махаланобиса.

Матрица ковариаций, нормированная на средний вектор, задается следующей формулой:

$$C = MM^{\mathrm{T}}.$$

Расстояние Махаланобиса от $\boldsymbol{x}$ до $\boldsymbol{\mu}$:

$$D(\boldsymbol{x}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} C^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}\,.$$

Считаем, что пиксель $\boldsymbol{x}$ лежит в классе «радужка», если

$$\exp\left(-D^2(\boldsymbol{x})\right) > P_{\text{порог}}\,. \tag{4}$$

Таким образом, алгоритм задается следующими параметрами: $a, b, P_{\text{порог}}$, где a и b — размеры окон (окрестностей) для расчета 1-го и 2-го признаков (2) и (3), а $P_{\text{порог}}$ — пороговое значение в формуле (4). Для определения оптимальных значений указанных параметров был использован генетический алгоритм.

4 Подбор параметров

Важно заметить, что применение генетического алгоритма для подбора параметров является не обучением, а настройкой алгоритма, т. е. параметры выбираются единожды и фиксируются.

Для применения генетического алгоритма нужно ввести функцию ошибки алгоритма для оценки качества решения для конкретных параметров. Пусть результат работы алгоритма равен J, а экспертная разметка областей затенения равна $\hat{J}$, где J и $\hat{J}$ — бинарные изображения вида (1).

Пусть N_0 — общее количество классифицируемых точек, лежащих внутри заданной границы радужки, а $\hat{N}_0$ — количество точек, классифицируемых как затенения в экспертной разметке. Обозначим через $\delta_{\text{Iris}}(i,j)$ индикаторную функцию множества точек, лежащих внутри кольца, аппроксимирующего область радужки. Тогда

$$N_0 = |\{(i,j) \mid \delta_{\text{Iris}}(i,j) = 1\}|; \quad \hat{N}_0 = |\{(i,j) \mid \delta_{\text{Iris}}(i,j)\hat{J}(i,j) = 1\}|.$$

Тогда относительная ошибка первого рода равна

$$E_1 = \frac{1}{\hat{N}_0} \sum_{i=1}^{H} \sum_{j=1}^{W} \hat{J}(i,j)(1 - J(i,j))\delta_{\text{Iris}}(i,j),$$

а относительная ошибка второго рода равна

$$E_2 = \frac{1}{N_0 - \hat{N}_0} \sum_{i=1}^{H} \sum_{j=1}^{W} J(i,j)(1 - \hat{J}(i,j))\delta_{\text{Iris}}(i,j).$$

Функция ошибки определяется как сумма относительных ошибок первого и второго рода:

$$E = E_1 + E_2\,.$$

В данном генетическом алгоритме особь — это вектор, состоящий из трех параметров алгоритма: $v = (a, b, P_{\text{порог}})^{\mathrm{T}}$. Оценка качества особи производилась на наборе из $M = 10$ изображений с экспертной разметкой затенений: $\mathcal{I} = \{I_m\}_{m=1}^{M}$, т. е. функция качества особи определялась как средняя точность алгоритма на изображениях из $\mathcal{I}$:

$$f(v) = f\begin{pmatrix} a \\ b \\ P_{\text{порог}} \end{pmatrix} = \frac{1}{M} \sum_{m=1}^{M} \left(1 - E(I_m, a, b, P_{\text{порог}})\right),$$

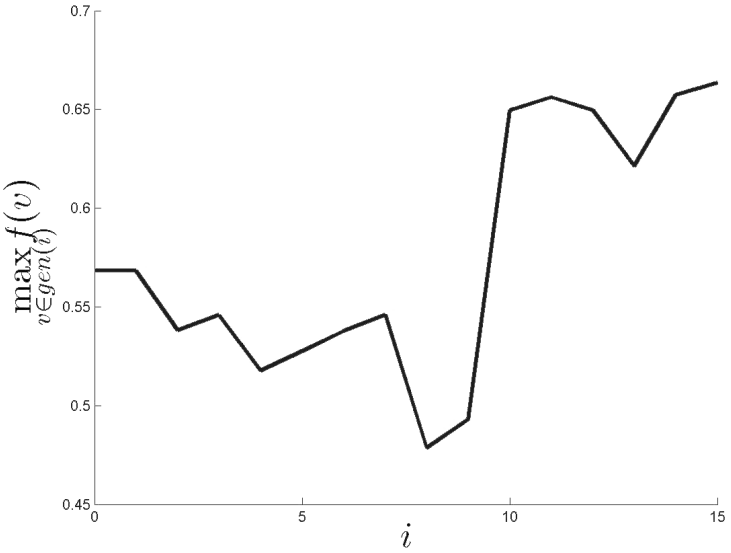


Рис. 3 Зависимость максимальной функции качества в популяции от номера поколения

где $E(I, a, b, P)$ — значение функции ошибки алгоритма с параметрами (a, b, P) на изображении I. Селекция выполнялась методом рулетки, т. е. вероятность выбора особи v в качестве родителя пропорциональна $f(v)$.

В генетическом алгоритме использовались следующие генетические операторы:

- скрещивание:

$$v \otimes w = \begin{pmatrix} a_1 \\ b_1 \\ P_1 \end{pmatrix} \otimes \begin{pmatrix} a_2 \\ b_2 \\ P_2 \end{pmatrix} = \begin{pmatrix} r(a_1, a_2) \\ r(b_1, b_2) \\ r(P_1, P_2) \end{pmatrix},$$

где $r(x, y)$ — случайная величина (распределенная по Бернулли):

$$r(x, y) = \begin{cases} x & \text{с вероятностью } p = \dfrac{1}{3}; \\ y & \text{с вероятностью } p = \dfrac{1}{3}; \\ \dfrac{x + y}{2} & \text{с вероятностью } p = \dfrac{1}{3}; \end{cases}$$

- мутация:

$$M(v) = M \begin{pmatrix} a \\ b \\ P \end{pmatrix} = \begin{cases} v & \text{с вероятностью } p = \dfrac{2}{3}; \\ (r_1, b, P)^{\mathrm{T}} & \text{с вероятностью } p = \dfrac{1}{9}; \\ (a, r_1, P)^{\mathrm{T}} & \text{с вероятностью } p = \dfrac{1}{9}; \\ (a, b, r_2)^{\mathrm{T}} & \text{с вероятностью } p = \dfrac{1}{9}, \end{cases}$$

где r_1 — дискретная случайная величина, равномерно распределенная на [1, 15]; r_2 — непрерывная случайная величина, равномерно распределенная на [0,6, 1]. Было проведено вычисление 15 поколений при размере популяции в 10 особей. На рис. 3 приведен

график зависимости максимальной функции качества в популяции от номера поколения в процессе выполнения генетического алгоритма. Лучшей особью в начальной популяции является вектор $v_0 = (5, 5, 0{,}7)^{\mathrm{T}}$.

В результате работы генетического алгоритма получены следующие параметры:

$$v^* = \begin{pmatrix} a \\ b \\ P \end{pmatrix} = \begin{pmatrix} 10 \\ 7 \\ 0{,}85 \end{pmatrix}; \quad f(v^*) = 0{,}664.$$

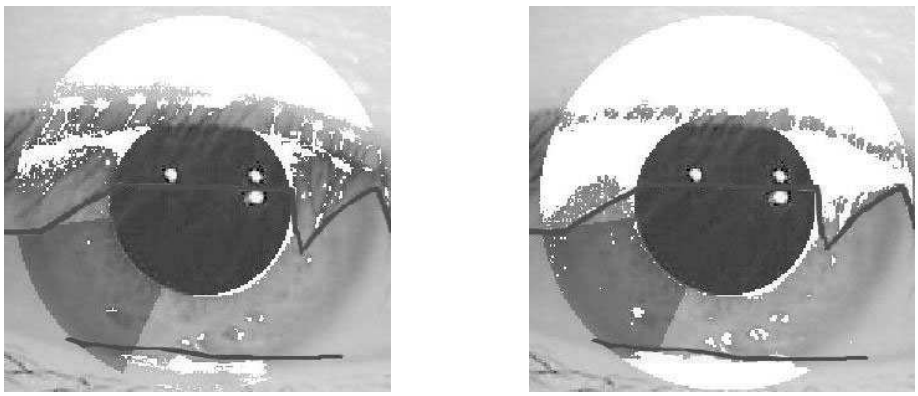
Строго говоря, утверждать, что данные параметры являются оптимальными, нельзя, однако на рис. 3 видно, что $f(v^*)$ больше, чем $f(v_0)$, т. е. подбор параметров генетическим алгоритмом улучшил результат. Это также видно на гистограммах на рис. 5 (см. разд. 5).

5 Вычислительный эксперимент

Метод был реализован в системе Matlab, код реализации находится в общем доступе [10]. К сожалению, возможности сравнить результаты с результатами других методов пока нет, ввиду отсутствия размеченных баз, на которых получены результаты других методов. Данная реализация метода была протестирована на базе изображений CASIA [11] с экспертной разметкой границ радужки, незатененного сектора и областей затенения (для оценки ошибки). Алгоритм запускался с параметрами: $v_0 = (5, 5, 0{,}7)^{\mathrm{T}}$ и v^*.

Примеры вывода алгоритмов с параметрами v_0 и v^* на одном и том же изображении приведены на рис. 4. На рис. 4 показаны две проблемы алгоритма:

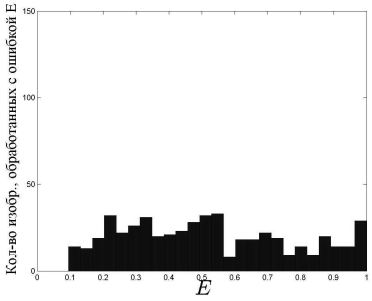
1) на изображениях виден шум (некоторые отдельные точки радужки классифицируются как затенения и наоборот);
2) существует проблема с распознаванием век — часть верхнего века мало отличается от радужки.



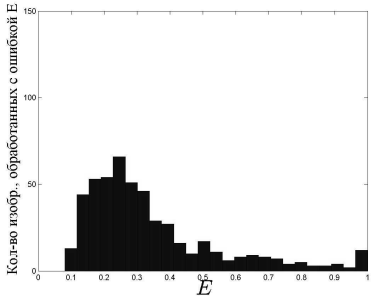
(*а*) Параметры v_0 (*б*) Параметры v^*

Рис. 4 Результат работы алгоритма на различных наборах параметров

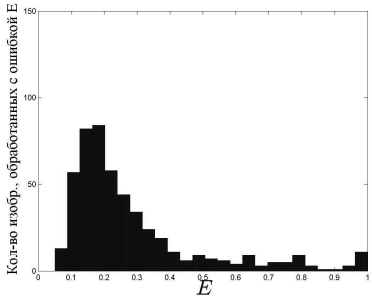
Для компенсации данных недостатков алгоритма применялась морфологическая постобработка. К логическому изображению вывода алгоритма применялись фильтры замыкания и размыкания. Фильтры считались в окнах 7×7. Точность алгоритма увеличилась после применения фильтра, это видно на рис. 5–8.



($а$) Параметры v_0; $\overline{E} = 0{,}524$



($б$) Параметры v^*; $\overline{E} = 0{,}344$



($в$) Параметры v^* + фильтр; $\overline{E} = 0{,}286$

Рис. 5 Гистограммы ошибок для различных модификаций алгоритма

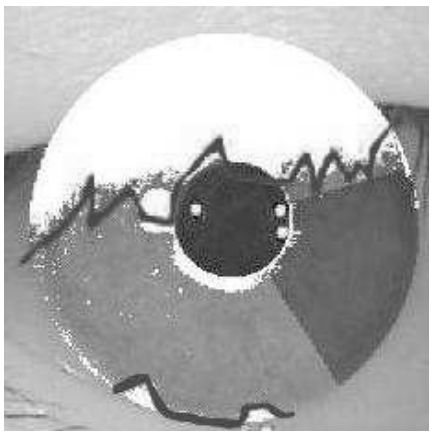
На рис. 5 представлены гистограммы распределения функционалов качества E для каждой из исследуемых модификаций алгоритмов на 500 изображениях из базы CASIA [11].

Стоит также заметить, что в экспертной разметке затенений не учитываются блики, в то время как данный алгоритм классифицирует блики как затенения. На некоторых изображениях это приводит к заведомо заниженной оценке точности распознавания.

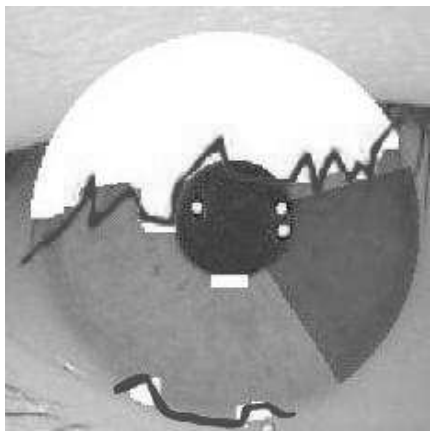
Далее приведены три примера работы алгоритма с параметрами v^* и применением фильтра.

Пример 1. На рис. 6 показаны результаты работы алгоритмов на файле 2003R05.bmp из базы CASIA, изображенном на рис. 6, *а*. Результат до применения фильтра изображен на рис. 6, *б*. Результат после применения фильтра изображен на рис. 6, *в*. Время работы: $t = 12{,}7$ с. Функция ошибки: $E = 0{,}102$.

($а$) Исходное изображение



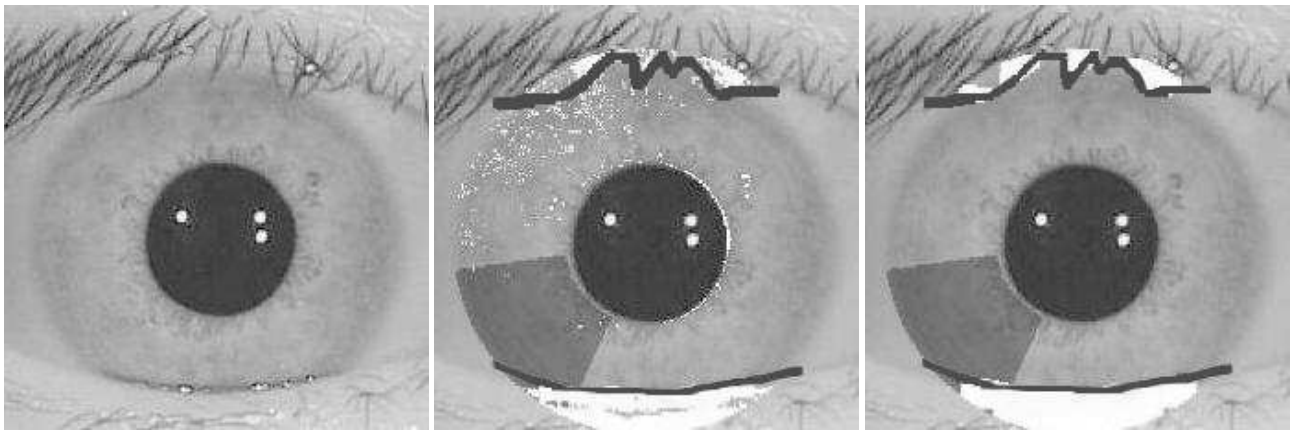
($б$) Результат без фильтра



($в$) Результат с фильтром

Рис. 6 Результаты работы на изображении 2003R05.bmp из базы CASIA [11]

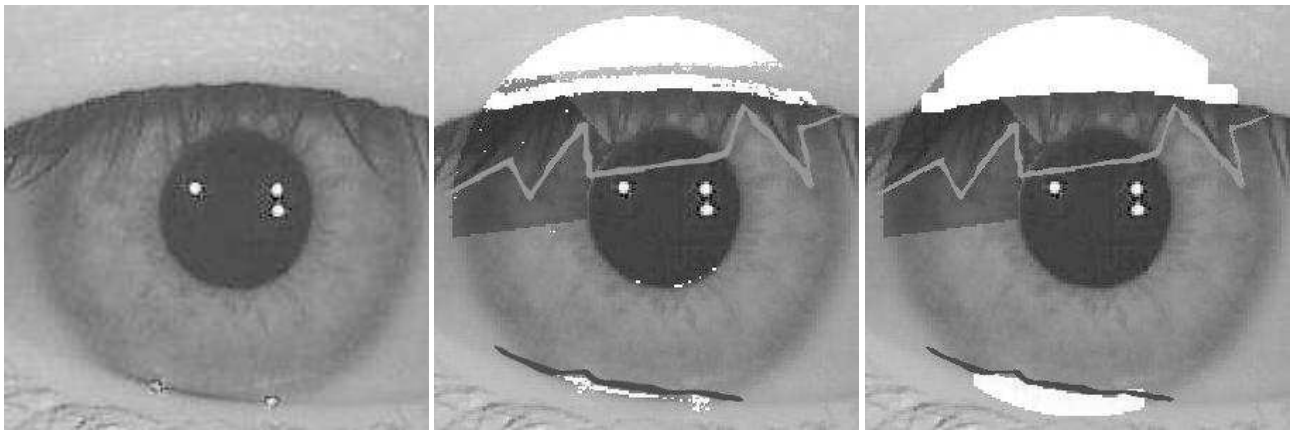
Пример 2. На рис. 7 показаны результаты работы алгоритмов на файле 2007R01.bmp из базы CASIA, изображенном на рис. 7, *а*. Результат до применения фильтра изображен на рис. 7, *б*. Результат после применения фильтра изображен на рис. 7, *в*. Время работы: $t = 10{,}3$ с. Функция ошибки: $E = 0{,}156$.



(*а*) Исходное изображение (*б*) Результат без фильтра (*в*) Результат с фильтром

Рис. 7 Результаты работы на изображении 2007R01.bmp из базы CASIA [11]

Пример 3. На рис. 8 показаны результаты работы алгоритмов на файле 2015R11.bmp из базы CASIA, изображенном на рис. 8, *а*. Результат до применения фильтра изображен на рис. 8, *б*. Результат после применения фильтра изображен на рис. 8, *в*. Время работы: $t = 10{,}7$ с. Функция ошибки: $E = 0{,}484$. В этом примере сектор S содержит затенения и, как следствие, алгоритм не распознал ресницы, однако тем не менее распознал веки.



(*а*) Исходное изображение (*б*) Результат без фильтра (*в*) Результат с фильтром

Рис. 8 Результаты работы на изображении 2015R11.bmp из базы CASIA [11]

6 Заключение

Представлен метод, позволяющий с высокой точностью выделять области затенения радужки. Основное преимущество метода состоит в том, что он является полностью автоматическим и не требует ручной разметки для обучающей выборки. Также отличительной чертой является тот факт, что метод строит новый классификатор для каждого изображения, что обеспечивает стабильность работы на изображениях с разной структурой радужки и с разной освещенностью. Скорость работы метода невелика (среднее время работы на одном изображении на выборке CASIA [11] составляет 10 с), однако это время можно уменьшить, используя более быстрые языки программирования.

Литература

[1] *Daugman J.* How iris recognition works // Conference (International) on Image Processing Proceedings. IEEE, 2002. Vol. 1. P. 33–36. doi: http://dx.doi.org/10.1109/ICIP.2002.1037952

[2] *Wildes R. P.* Iris recognition: An emerging biometric technology // Proc. IEEE, 1997. Vol. 85. No. 9. P. 1348–1363. doi: http://dx.doi.org/10.1109/5.628669

[3] *He Z. F., Tan T. N., Sun Z. A., Qiu X. C.* Robust eyelid, eyelash and shadow localization for iris recognition // 15th IEEE Conference (International) on Image Processing Proceedings. IEEE, 2008. P. 265–268.

[4] *Krichen E., Garcia-Salicetti S., Dorizzi B.* A new probabilistic iris quality measure for comprehensive noise detection // 1st IEEE Conference (International) on Biometrics: Theory, Applications, and Systems. IEEE, 2007. P. 1–6. doi: http://dx.doi.org/10.1109/BTAS.2007.4401906

[5] *Daugman J.* New methods in iris recognition // IEEE Trans. Syst. Man Cyb., 2007. Vol. 37. No. 5. P. 1167–1175. doi: http://dx.doi.org/10.1109/TSMCB.2007.903540

[6] *He Z. F., Tan T. N., Sun Z. A., Qiu X. C.* Toward accurate and fast iris segmentation for iris biometrics // IEEE Trans. Pattern Anal., 2009. Vol. 31. No. 9. P. 1670–1684.

[7] *Zhang D., Monro D. M., Rakshit S.* Eyelash removal method for human iris recognition // Conference (International) on Image Processing Proceedings. IEEE, 2006. P. 285–288. doi: http://dx.doi.org/10.1109/ICIP.2006.313181

[8] *Li Y., Savvides M.* A pixel-wise, learning-based approach for occlusion estimation of iris images in polar domain // ICASSP. IEEE, 2009. P. 1357–1360. doi: http://dx.doi.org/10.1109/ICASSP.2009.4959844

[9] *Ганькин К. А., Гнеушев А. Н., Матвеев И. А.* Сегментация изображения радужки глаза, основанная на приближенных методах с последующими уточнениями // Известия РАН. Теория и системы управления, 2014. № 2. С. 80–94. doi: http://dx.doi.org/10.1134/S1064230714020099

[10] *Соломатин И. А.* Реализация алгоритма выделения областей затенения радужки классификатором локальных текстурных признаков. 2015. http://svn.code.sf.net/p/mlalgorithms/code/Group274/Solomatin2015EESLocalization/code/.

[11] Chinese Academy of Sciences Institute of Automation. CASIA-IrisV3: CASIA-Iris-Lamp image database. http://www.cbsr.ia.ac.cn/IrisDatabase.htm.

Поступила в редакцию 14.06.15

Detecting visible areas of iris by qualifier of local textural features*

I. A. Solomatin and I. A. Matveev

ivan.solomatin@phystech.edu

Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Russia

Dorodnicyn Computer Centre of the Russian Academy of Sciences, 40 Vavilova st., Moscow, Russia

A person recognition by the image of the iris is an actual problem. To increase the accuracy of recognition, usually areas of occlusion are detected in addition to locating of the iris as an annular region. The problem of occlusion detection can be set as the classification of pixels from annular region into two classes: “iris” and “occlusion.” In the annular region, the segment with minimum dispersion of brightness is selected, which usually contains no occlusion (in this article, this segment is not calculated and supposed to be a part of the input data).

*This work was done under financial support of the Russian Foundation for Basic Research (grant 15-01-05552)

Then, a classifier based on multivariate Gaussian is built and after that, it is trained on the training set, which is set by local textural features of the pixels from this sector. The parameters of the classifier were optimized using genetic algorithm. The problems with noise and errors of classification in particular areas of the image are solved by applying morphological postprocessing. A computational experiment was carried out and it allowed to obtain the distribution of the functional of quality of the algorithm.

Keywords: *image processing; pattern recognition; biometric identification by iris*

DOI: 10.21469/22233792.1.14.03

References

[1] Daugman, J. 2002. How iris recognition works. *Conference (International) on Image Processing Proceedings*. IEEE. 1:33–36. doi: http://dx.doi.org/10.1109/ICIP.2002.1037952

[2] Wildes, R. P. 1997. Iris recognition: An emerging biometric technology. *IEEE Proc.* 85(9):1348–1363. doi: http://dx.doi.org/10.1109/5.628669

[3] He, Z. F., T. N. Tan, Z. A. Sun, and X. C. Qiu. 2008. Robust eyelid, eyelash and shadow localization for iris recognition. *15th IEEE Conference (International) on Image Processing Proceedings*. IEEE. 265–268.

[4] Krichen, E., S. Garcia-Salicetti, and B. Dorizzi. 2007. A new probabilistic iris quality measure for comprehensive noise detection. *1st IEEE Conference (International) on Biometrics: Theory, Applications, and Systems*. 1–6. doi: http://dx.doi.org/10.1109/BTAS.2007.4401906

[5] Daugman, J. 2007. New methods in iris recognition. *IEEE Trans. Syst. Man Cyb.* 37(5):1167–1175. doi: http://dx.doi.org/10.1109/TSMCB.2007.903540

[6] He, Z. F., T. N. Tan, Z. A. Sun, and X. C. Qiu. 2009. Toward accurate and fast iris segmentation for iris biometrics. *IEEE Trans. Pattern Anal.* 31(9):1670–1684.

[7] Zhang, D., D. M. Monro, and S. Rakshit. 2006. Eyelash removal method for human iris recognition. *Conference (International) on Image Processing Proceedings*. 285–288. doi: http://dx.doi.org/10.1109/ICIP.2006.313181

[8] Li, Y., and M. Savvides. 2009. A pixel-wise, learning-based approach for occlusion estimation of iris images in polar domain. *ICASSP*. IEEE. 1357–1360. doi: http://dx.doi.org/10.1109/ICASSP.2009.4959844

[9] Gankin, K., A. Gneushev, and I. Matveev. 2014. Iris image segmentation based on approximate methods with subsequent refinements. *J. Comput. Sys. Sc. Int.* 53(2):224–238. doi: http://dx.doi.org/10.1134/S1064230714020099

[10] Solomatin, I. A. 2015. Algorithm, detecting visible areas of iris by qualifier of local textural features. Available at: http://svn.code.sf.net/p/mlalgorithms/code/Group274/Solomatin2015EESLocalization/code/ (accessed January 11, 2016).

[11] CASIA. Chinese Academy of Sciences Institute of Automation. CASIA-IrisV3: CASIA-IrisLamp image database. Available at: http://www.cbsr.ia.ac.cn/IrisDatabase.htm (accessed January 11, 2016).

Received June 14, 2015

2-симплекс призма — когнитивное средство принятия и обоснования решений в интеллектуальных динамических системах*

А. Е. Янковская[1,2,3,4]***, А. В. Ямшанов***[4]***, Н. М. Кривдюк***[4]
ayyankov@gmail.com, yav@keva.tusur.ru, skratnat@gmail.com
[1]Томский государственный архитектурно-строительный университет, Россия, г. Томск, пл. Соляная, 2
[2]Томский государственный университет, Россия, г. Томск, пр. Ленина, 36
[3]Сибирский государственный медицинский университет, Россия, г. Томск, Московский тракт, 2
[4]Томский государственный университет систем управления и радиоэлектроники, Россия, г. Томск, пр. Ленина, 40

Для ряда проблемных и междисциплинарных областей, таких как медицина, биомедицина, экогеология, образование, дорожное строительство, впервые в интеллектуальных динамических системах предлагается для принятия и обоснования решений применять оригинальное когнитивное средство — 2-симплекс призму. Идея применения n-симплексов, теорема для принятия и обоснования решений на основе n-симплексов и использование ее в интеллектуальных системах (ИС) впервые предложены А. Е. Янковской в 1990 г. Описывается применение 2-симплекс призмы для принятия и обоснования решений в интеллектуальных динамических системах, основанных на тестовых методах распознавания образов, нечеткой и пороговой логиках.

Ключевые слова: *2-симплекс призма; n-симплекс; когнитивное средство; принятие решений; обоснование решений; интеллектуальные системы; интеллектуальные динамические системы; тестовые методы распознавания образов; нечеткая логика; пороговая логика; области применения 2-симплекс призмы*

DOI: 10.21469/22233792.1.14.04

1 Введение

Когнитивная графика как научное направление, связанное с принятием и обоснованием решений в ИС, начала развиваться с 1970-х гг. Значительный вклад в развитие когнитивных средств внесли R. Axelrod [1], R. G. Basaker и T. L. Saati [2], Д. А. Поспелов [3, 4], Д. А. Поспелов с соавт. [5], А. А. Зенкин [6, 7], В. А. Албу и В. Ф. Хорошевский [8], Б. А. Кобринский [9], А. Е. Янковская [10], А. Е. Янковская с соавт. [11, 12].

Растущий интерес к разработке и использованию прикладных ИС стимулирует спрос на создание графических, включая когнитивные, средств визуализации различных информационных структур, выявленных закономерностей, а также принятия и обоснования решений. Графические изображения на экране компьютера упрощают восприятие и понимание закономерностей в данных и знаниях.

В настоящее время средства когнитивной графики широко используются в различных ИС для решения разнообразных задач для проблемных и междисциплинарных областей: медицина, образование, геология, проектирование, радиоэлектроника, социология, психология, психиатрия, экобиомедицина, экогеология и др. Средства когнитивной графики

*Работа поддержана грантами РФФИ (проекты №№ 13-07-00373, 13-07-98037 и 14-07-00673) и частично РГНФ (проект № 13-06-00709).

широко применяются в разнообразных ИС: для анализа информационных структур, выявления различного рода закономерностей в данных и знаниях, принятия и обоснования принятия результатов решения; в интеллектуальных обучающе-тестирующих системах: для визуализации и прогнозирования результатов процесса обучения, для оптимизации учебного процесса и др.

Весьма актуальным является применение когнитивных средств для анализа и визуализации динамических процессов при принятии и обосновании решений. Например, преподавателю необходимо учитывать полученные респондентами в ходе обучения результаты, для постановки итоговой оценки или врачу для отслеживания осуществленных мероприятий в ходе лечения пациентов в целях успешного лечения. Применение когнитивной графики при принятии решений существенно упрощает анализ информации и способствует принятию оптимального решения.

Предлагаемая статья является продолжением исследования применения когнитивных средств, основанных на n-симплексе [13, 14]. Несомненным преимуществом когнитивных средств, основанных на n-симплексе, является инвариантность к проблемным областям. Отметим, что целесообразно создание средств когнитивной графики с использованием современных технологий и с учетом всех доступных платформ: desktop-приложения (приложения для настольного персонального компьютера), приложения для смартфонов и планшетов, WEB-приложения.

Ниже описываются математические основы представления исследуемого объекта в 2-симплексе и основы представления исследуемого объекта (процесса) на базе 2-симплекс призмы; приводятся примеры применения 2-симплекс призмы в разработанных, а также разрабатываемых ИС; предлагаются дальнейшие направления исследований.

2 Математические основы представления исследуемого объекта в 2-симплекс призме

В основе принятия и обоснования решений, а также графической визуализации правильного n-симплекса лежит следующая теорема, сформулированная в [15, 16].

Теорема. Для любого набора одновременно не равных нулю чисел $a_1, a_2, \ldots, a_{n+1}$, где n — размерность правильного n-симплекса, можно найти одну и только одну такую точку, что $h_1 : h_2 : \cdots : h_{n+1} = a_1 : a_2 : \cdots : a_{n+1}$, где $h_i, (i = 1, \ldots, n+1)$ — расстояние этой точки до i-й грани [15, 16]. Коэффициент $a_i, (i = 1, \ldots, n+1)$ представляет собой степень условной близости исследуемого объекта к i-му образу.

Эта теорема использовалась при построении более чем тридцати прикладных интеллектуальных системах и трех инструментальных средств, предназначенных для выявления различного рода закономерностей и принятия диагностических, организационно-управленческих и классификационных решений, для принятия решения и их обоснования.

Далее приведем соотношения между коэффициентами и высотами для 2-симплекса. Поскольку 2-симплекс обладает свойством постоянства суммы расстояний (h) из любой точки до его граней и свойством сохранения отношений между этими расстояниями $h_1 : a_1 = h_2 : a_2 = h_3 : a_3$, то расстояния h_1, h_2, h_3 от точки до сторон треугольника вычисляются на основе коэффициентов $a_i (i \in 1, 2, 3)$ и операции нормализации исходя из следующих соотношений:

$$H = \sum_{i=1}^{3} h_i;$$
$$H = A \sum_{i=1}^{3} a_i;$$
$$\frac{h_1}{a_1} = \frac{h_2}{a_2} = \frac{h_3}{a_3}$$

по формуле:

$$h_i = Aa_i, \quad i \in \{1, 2, 3\}.$$

Когнитивное средство 2-симплекс призма представляет собой правильную треугольную призму, содержащую в основаниях и срезах 2-симплексы, зафиксированные в конкретные моменты времени. В целях вычисления расстояния от основания призмы до 2-симплекса h_2 введем следующие параметры: H_2 — высота 2-симплекс призмы, задаваемая пользователем и сопоставленная продолжительности исследования; t — момент фиксации параметров исследуемого объекта; $T_{\min}$ — момент первой фиксации параметров исследуемого объекта; $T_{\max}$ — момент последней фиксации параметров исследуемого объекта. Расстояние h_2 вычисляется по следующей формуле:

$$h_2 = H_2 \frac{t - T_{\min}}{T_{\max} - T_{\min}}.$$

3 Применение 2-симплекс призмы в различных проблемных областях

Когнитивное средство 2-симплекс призма предназначено для применения в различных проблемных и междисциплинарных областях. Рассмотрим примеры применения 2-симплекс призмы в различных областях: (1) диагностики и интервенции организационного стресса (ОС); (2) экспресс-диагностики и профилактики депрессии; (3) образовании.

Рассмотрим пример диагностики и интервенции ОС с использованием 2-симплекс призмы в ИС экспресс-диагностики и интервенции ОС (ДИОС) [17], основанной на идее трехступенчатой диагностики и выбора интервенции на базе сконструированного авторами опросника [17, 18] по каждой из трех стадий ОС (1 — напряжения; 2 — адаптации; 3 — истощения) и базируемой на пороговой и нечеткой логиках [19, 20]. Идея трехступенчатой диагностики ОС позволяет в более короткие сроки оказывать дифференцированную помощь при наличии ОС у обследуемых.

Для экспресс-диагностики ОС используется опросник, включающий вопросы по трем стадиям ОС: 1 — напряжения (возбуждения); 2 — адаптации (сохранения энергии); 3 — истощения. В основе опросника лежит концепция Г. Селье [21]. В 1-й и 2-й стадиях используется 7 признаков (симптомов) для выявления ОС; в 3-й стадии — 8 признаков, причем количество значений каждого признака равно 5 (никогда — 0; редко — 0,25; иногда — 0,5; часто — 0,75; постоянно — 1).

После проведения тестирования система обрабатывает анкету исследуемого и выдает результаты диагностики ОС по каждой из стадий. Полученные результаты передаются в модуль когнитивной графики. В случае если не предусмотрено исследование динамики развития ОС, то для принятия и обоснования используется когнитивное средство 2-симплекс. Если необходимо наблюдение в динамике за обследуемым, то используется 2-симплекс призма. Отображаемая динамика результатов по диагностике ОС с применением

2-симплекс призмы позволяет определить степень близости к тому или иному диагнозу на каждом этапе лечения.

2-симплекс призма одновременно отображает только три образа, однако ИС ДИОС выявляет как 3 стадии ОС (возбуждения, адаптация, истощение), так и отсутствие стресса. В связи с этим предлагается для отображения динамики использовать две 2-симплекс призмы: первая предназначена для отображения динамики для трех стадий ОС (рис. 1), вторая — для отображения двух первых стадий и факта отсутствия стресса (рис. 2). Отметим, что если у обследуемого история развития болезни укладывается в рамки диагнозов, сопоставленных сторонам одной из двух 2-симплекс призм, то целесообразно использовать только одну призму. Также заметим, что в 2-симплекс призме используются когнитивные свойства, представленные цветовой палитрой для отображения опасности диагнозов и сопоставляемых им образов, например, красным цветом отображается самая тяжелая стадия ОС — истощение. Кроме того, поскольку рис. 1 и 2 используют перспективную проекцию, может создаться ощущение, что линии, соединяющие точки и грани, не перпендикулярны граням.

Представим иллюстративный пример, когда история развития болезни не укладывается в рамки диагнозов, сопоставленных сторонам одной 2-симплекс призмы, а необходимо использовать две 2-симплекс призмы. При этом проведено 5 тестов по определению стадии ОС. В первой 2-симплекс призме отображены результаты 1-, 2-, 3- и 4-го тестов, во второй 2-симплекс призме отображены результаты 3-, 4- и 5-го тестов. На рис. 1 представлена 2-симплекс призма, отображающая первую часть лечения обследуемого со стадии истощения (3) до стадии возбуждения (1).

Первый тест (T_1) выявил у обследуемого стадии адаптации (2) и истощения (3), причем стадия истощения преобладает над стадией адаптации и существенно преобладает над стадией возбуждения (1). Второй тест (T_2) выявил, что болезнь развивается от стадии истощения (3) к стадии адаптации (2). Третий тест (T_3) выявил, что болезнь развилась до уровня, находящегося между стадией адаптации (2) и возбуждения (1). Четвертый тест (T_4) выявил преобладание стадии возбуждения (1).

Вторая 2-симплекс призма (см. рис. 2) отображает процесс перехода от стадии адаптации (2) к отсутствию стресса (0).

Пятый тест (T_5) выявил отсутствие ОС.

Рассмотрим отображение результатов тестирования качества знаний респондентов в модуле интерпретации результатов обучающе-тестирующей системы с использовани-

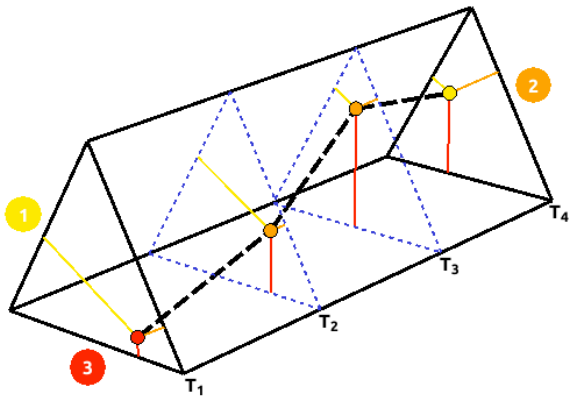


Рис. 1 Отображение результатов четырех тестов диагностики ОС в 2-симплекс призме

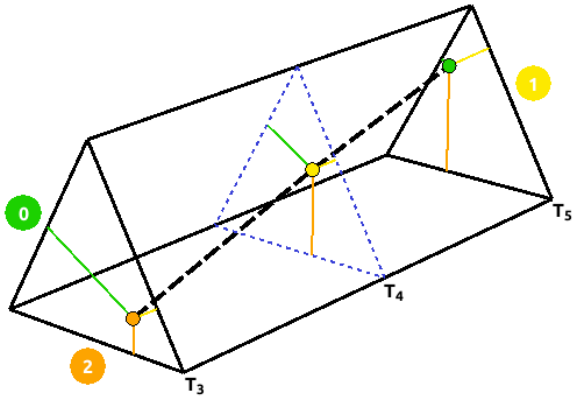


Рис. 2 Отображение результатов пяти диагностик ОС в 2-симплекс призме

ем оценочных коэффициентов, определяющих насколько хорошо респондент справляется с различными заданиями на основе способностей (навыков) [22]. В разрабатываемой авторами обучающе-тестирующей системе респондент после изучения выбранной дисциплины проходит смешанный диагностический тест, представляющий собой оптимальное сочетание условной и безусловной составляющей [23]. Во время прохождения смешанного диагностического теста (СДТ) формируется карта действий респондента (КДР). После прохождения респондентом СДТ КДР проецируется в набор оценочных коэффициентов на основе следующих способностей (навыков): (1) запоминание и воспроизведение учебного материала в неизмененном виде; (2) воспроизведение учебного материала в измененном виде; (3) извлечение новых знаний на основе изученного учебного материала; (4) решение новых задач и т. д.

При разработке клиент-серверной программной системы с мультимедийными возможностями целесообразен перевод набора оценочных коэффициентов в следующие показатели: (1) решение задач, требующих большой сосредоточенности; (2) решение нетривиальных задач; (3) быстрая обучаемость и знание большого количества технологий.

Отметим, что при оценке способностей респондентов на основе оценочных коэффициентов возможна ситуация, когда респондент одинаково владеет или не владеет знаниями по оцениваемым способностям. Это отображается кругом в центре 2-симплекса, что затрудняет визуально определить значение показателей. С целью разделения этих двух состояний используется цветовая индикация — насыщенность цвета для круга, отображающего прохождение теста, сопоставлена набранному количеству баллов по тесту (рис. 3).

На основе 2-симплекс призмы предлагается осуществлять анализ динамики развития способностей как респондента, так и группы респондентов. Используя 2-симплекс призму, как и любое графическое средство, необходимо учитывать, что отображение результатов большой группы респондентов резко увеличивает сложность анализа.

Применение 2-симплекс призмы не ограничивается вышеописанными примерами.

4 Заключение

Одним из важных свойств отображения информации в 2-симплекс призме является возможность анализировать динамику положения исследуемого объекта на заданном временно́м интервале, что позволяет пользователям принимать и обосновывать принимаемые решения, анализируя изменение параметров исследуемого объекта.

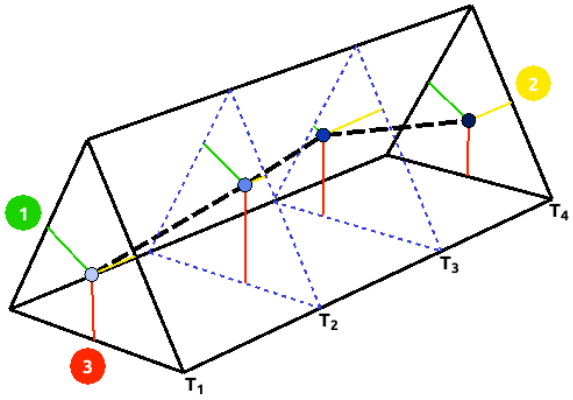


Рис. 3 Пример использования 2-симплекс призмы для обучающе-тестирующих систем

Применение когнитивных средств целесообразно для любой проблемной и междисциплинарной области, в которой необходимо проводить принятие и обоснование решения об отношении исследуемого объекта к тому или иному образу (классу) в фиксированный момент времени или на заданном промежутке времени. В отличие от ранее разработанных когнитивных средств, основанных на n-симплексе [24, 25], 2-симплекс призма позволяет исследовать объект динамически на заданном пользователем временно́м интервале.

Литература

[1] *Axelrod R.* The structure of decision: Cognitive maps of political elites. — Princeton University Press, 1976. 395 p.

[2] *Basaker R. G., Saati T. L.* Finite graphs and networks: An introduction with applications. — New York, NY – London – Toronto: Research Analysis Corp., Mc Graw Hill Co., 1965. 294 p.

[3] *Поспелов Д. А.* Когнитивная графика — окно в новый мир // Программные продукты и системы, 1992. Т. 2. С. 4–6.

[4] *Поспелов Д. А.* Десять «горячих точек» в исследованиях по искусственному интеллекту // Интеллектуальные системы (МГУ), 1996. Т. 1. Вып. 1-4. С. 47–56.

[5] *Поспелов Д. А., Литвинцева Л. В.* Как совместить левое и правое? // Новости искусственного интеллекта. 1996. № 2. С. 66–71.

[6] *Зенкин А. А.* Когнитивная компьютерная графика. — М.: Наука, 1991. 192 с.

[7] *Зенкин А. А.* Знание-порождающие технологии когнитивной реальности // Новости искусственного интеллекта. 1996. № 2. С. 72–78.

[8] *Албу В. А., Хорошевский В. Ф.* КОГР система когнитивной графики: разработка, реализация и применения // Изв. АН СССР. Техническая кибернетика, 1990. № 5. С. 105–118.

[9] *Кобринский Б. А.* К вопросу учета образного мышления и интуиции в экспертных медицинских системах // V Национальная конф. с междунар. участием «Искусственный интеллект-96»: Сб. науч. тр., 1996. Т. 2. С. 110–117.

[10] *Янковская А. Е.* Принятие и обоснование решений с использованием методов когнитивной графики на основе знаний экспертов различной квалификации // Известия РАН. Теория и система управления, 1997. № 5. С. 125–126.

[11] *Yankovskaya A., Galkin D.* Cognitive computer based on n-m multiterminal networks for pattern recognition in applied intelligent systems // Conference GraphiCon'2009 Proceedings. — Moscow: Maks Press, 2009. P. 299–300.

[12] *Yankovskaya A. E., Galkin D. V., Chernogoryuk G. E.* Computer visualization and cognitive graphics tools for applied intelligent systems // IASTED Conference (International) on Automation, Control and Information Technology Proceedings, 2010. Vol. 1. P. 249–253. doi: http://dx.doi.org/10.2316/P.2010.691-081

[13] *Янковская А. Е., Тетенев Ф. Ф., Черногорюк Г. Э.* Отражение образного мышления специалиста в интеллектуальной распознающей системе патогенеза заболевания // Компьютерная хроника, 2000. № 6. С. 77–92.

[14] *Yankovskaya A. E., Mozheiko V. I.* Optimization of a set of tests selection satisfying the criteria prescribed // 7th Conference (International) on Pattern Recognition and Image Analysis: New Information Technologies Proceedings. — St. Petersburg: SPbETU, 2004. Vol. I. C. 145–148.

[15] *Янковская А. Е.* Преобразование пространства признаков в пространство образов на базе логико-комбинаторных методов и свойств некоторых геометрических фигур // Распознавание образов и анализ изображений: новые информационные технологии: Тез. докл. I Всесоюзной конф. — Минск, 1991. Ч. II. С. 178–181.

[16] *Кондратенко С. В., Янковская А. Е.* Система визуализации TRIANG для обоснования принятия решений с использованием когнитивной графики // Тез. докл. III Конф. по искусственному интеллекту. — Тверь, 1992. Т. I. С. 152–155.

[17] *Янковская А. Е., Китлер С. В., Силаева А. В.* Интеллектуальная система диагностики и интервенции организационного стресса: ее развитие и апробация // Открытое образование, 2012. № 2(91). С. 61–69.

[18] *Корнетов Н. А., Янковская А. Е., Китлер С. В., Силаева А. В., Шагалова Л. В.* К вопросу динамики развития представлений об организационном стрессе и подходов к его оценке // Фундаментальные исследования. 2011. № 10. Ч. 3. С. 598–603.

[19] *Beck A. T., Ward C., Mendelson M.* Beck Depression Inventory (BDI) // Arch. Gen. Psychiat., 1961. Vol. 4. No. 6. P. 561–571.

[20] *Zadeh, L. A.* Fuzzy logic, neural networks, and soft computing // Commun. ACM, 1994. Vol. 37. No. 3. P. 77–84. doi: http://dx.doi.org/10.1145/175247.175255

[21] *Selye H.* A syndrome produced by diverse nocuous agents // Nature, 1936. Vol. 138. P. 32. doi: http://dx.doi.org/10.1038/138032a0

[22] *Янковская А. Е., Шурыгин Ю. А., Ямшанов А. В., Кривдюк Н. М.* Определение уровня усвоенных знаний по обучающему курсу, представленному семантической сетью // Открытые семантические технологии проектирования интеллектуальных систем: Мат-лы V Междунар. науч.-техн. конф. — Минск: БГУИР, 2015. С. 331–339.

[23] *Янковская А. Е.* Логические тесты и средства когнитивной графики. — LAP LAMBERT Academic Publishing, 2011. 92 с.

[24] *Yankovskaya A., Krivdyuk N.* Cognitive graphics tool based on 3-simplex for decision-making and substantiation of decisions in intelligent system // IASTED Conference (International) on Technology for Education and Learning Proceedings, 2013. P. 463–469. doi: http://dx.doi.org/10.2316/P.2013.808-017

[25] *Янковская А. Е. Ямшанов А. В. Кривдюк Н. М.* Средства когнитивной графики в интеллектуальных обучающе-тестирующих системах // Открытые семантические технологии проектирования интеллектуальных систем: Мат-лы IV Междунар. науч.-техн. конф. — Минск: БГУИР, 2014. С. 303–308.

Поступила в редакцию 26.06.15

2-simplex prism — a cognitive tool for decision-making and its justifications in intelligent dynamic systems*

***A. E. Yankovskaya*[1,2,3,4], *A. V. Yamshanov*[4], *and N. M. Krivdyuk*[4]**

ayyankov@gmail.com, yav@keva.tusur.ru, skratnat@gmail.com

[1]Tomsk State University of Architecture and Building, 2 Solyanaya Sq., Tomsk, Russia

[2]Tomsk State University, 36 Pr. Lenina, Tomsk, Russia

[3]Siberian State Medical University, 2 Moskovskiy trakt, Tomsk, Russia

[4]Tomsk State University of Control Systems and Radioelectronics, 40 Pr. Lenina, Tomsk, Russia

The cognitive tool 2-simplex prism is first proposed for application at decision-making and its justification in intelligent dynamic systems for different problem areas: medicine, biomedicine, ecogeology, education, road building etc. The idea of n-simplex application, the theorem for decision-making and its justification for intelligent systems proposed by A. Yankovskaya in 1990 year. Usage of 2-simplex prism for application at decision-making and its justification in intelligent dynamic systems based on test methods of pattern recognition and methods of fuzzy and threshold logics is described.

Keywords: *2-simplex prism; n-simplex; cognitive tool; decision-making; justification; intelligent systems; intelligent dynamic systems; test methods of pattern recognition; fuzzy logic; threshold logic; areas of 2-simplex prism application*

DOI: 10.21469/22233792.1.14.04

References

[1] Axelrod, R. 1976. *The structure of decision: Cognitive maps of political elites*. Princeton University Press. 395 p.

[2] Basaker, R. G., and T. L. Saati. 1965. *Finite graphs and networks: An introduction with applications*. New York, NY – London – Toronto: Research Analysis Corp., Mc Graw Hill Co. 294 p.

[3] Pospelov, D. A. 1992. Cognitive graphics — a window into the new world. *Software Products Systems* 2:4–6. (In Russian.)

[4] Pospelov, D. A. 1996. Ten "hot spots" in research on artificial intelligence *Intelligent Systems (MSU)* 1(1-4):47–56. (In Russian.)

[5] Pospelov, D. A., and L. V. Litvintseva. 1996. How to combine left and right? *News Artificial Intelligence* 2:66–71. (In Russian.)

[6] Zenkin, A. A. 1991. *Cognitive computer graphics*. Moscow: Nauka. 192 p. (In Russian.)

[7] Zenkin, A. A. 1996. Knowledge-generating technologies of cognitive reality. *News Artificial Intelligence* 2:72–78. (In Russian.)

[8] Albu, V. A., and V. F. Khoroshevskiy. 1990. COGR — Cognitive graphics system, design, development, application. *Russ. Acad. Sci. Bull. Technical Cybernetics* 5:105–118. (In Russian.)

[9] Kobrinskiy, B. A. 1996. Why should we take into account imaginary thinking and intuition in medical expert systems. *5th National Conference with International Participation "Artificial Intelligence-96" Proceedings*. Kazan. 2:110–117. (In Russian.)

[10] Yankovskaya, A. E. 1997. Decision-making and decision-justification using cognitive graphics methods based on the experts of different qualification. *Russ. Acad. Sci. Bull., Theory and Control Systems* 5:125–126. (In Russian.)

*The work is supported by the Russian Foundation for Basic Research (projects 13-07-00373, 13-07-98037 and 14-07-00673) and partially supported by the Russian Humanitarian Scientific Foundation (project 13-06-00709).

[11] Yankovskaya, A., and D. Galkin. 2009. Cognitive computer based on *n-m* multiterminal networks for pattern recognition in applied intelligent systems. *Conference GraphiCon'2009 Proceedings*. Moscow: Maks Press. 299–300.

[12] Yankovskaya, A. E., D. V. Galkin, and G. E. Chernogoryuk. 2010. Computer visualization and cognitive graphics tools for applied intelligent systems. *IASTED Conference (International) on Automation, Control and Information Technology Proceedings*. 1:249–253. doi: `http://dx.doi.org/10.2316/P.2010.691-081`

[13] yank2000mirror Yankovskaya, A. E., F. F. Tetenev, and G. E. Chernogoryuk. 2000. Reflection of creative thinking expert in intellectual pattern recognition system disease pathogenesis. *Computer Chronicle* 6:77–92. (In Russian.)

[14] Yankovskaya, A. E., and V. I. Mozheiko. 2004. Optimization of a set of tests selection satisfying the criteria prescribed. *7th Conference (International) on Pattern Recognition and Image Analysis: New Information Technologies Proceedings*. St. Petersburg: SPbETU. I:145–148.

[15] Yankovskaya, A. E. 1991. Transformation of features space in patterns space on the base of the logical-combinatorial methods and properties of some geometric figures. *Conference (International) on Pattern Recognition and Image Analysis: New Information: Abstracts*. Minsk. II:178–181. (In Russian.)

[16] Kondratenko, S. V., and A. E. Yankovskaya. 1992. System of visualization TRIANG for decision-making substantiation with use of cognitive graphics. *3rd Conference on Artificial Intelligence Proceedings*. Tver. I:152–155. (In Russian.)

[17] Yankovskaya, A. E., S. V. Kitler, and A. V. Silaeva. 2012. Intelligent system of diagnostics and intervention of organizational stress: Its development and testing. *Open Education* 2(91):61–69. (In Russian.)

[18] Kornetov, N. A., A. E. Yankovskaya, S. V. Kitler, A. V. Silaeva, and L. V. Shagalova. 2011. About development dynamics of representations about organizational stress and approaches to its evaluation. *Fundamental Research* 10(3):598–603. (In Russian.)

[19] Beck, A. T., C. Ward, and M. Mendelson. 1961. Beck Depression Inventory (BDI). *Arch. Gen. Psychiat.* 4(6):561–571.

[20] Zadeh, L. A. 1994. Fuzzy logic, neural networks, and soft computing. *Commun. ACM* 37(3):77–84. doi: `http://dx.doi.org/10.1145/175247.175255`

[21] Selye, H. 1936. A syndrome produced by diverse nocuous agents. *Nature* 138:32. doi: `http://dx.doi.org/10.1038/138032a0`

[22] Yankovskaya, A. E., Y. A. Shurigin, A. V. Yamshanov, and N. M. Krivdyuk. 2015. Determination of the student knowledge level on the base of a training course which is presented by a semantic network. *Open Semantic Technologies for Intelligent Systems (OSTIS-2015) Proceedings*. Minsk: BSUIR. 331–339. (In Russian.)

[23] Yankovskaya, A. E. 2011. *Logic tests and cognitive graphic tools*. LAP LAMBERT Academic Publishing. 92 p.

[24] Yankovskaya, A. E., and N. M. Krivdyuk. 2013. Cognitive graphics tool based on 3-simplex for decision-making and substantiation of decisions in intelligent system. *IASTED Conference (International) Technology for Education and Learning Proceedings*. 463–469. doi: `http://dx.doi.org/10.2316/P.2013.808-017`

[25] Yankovskaya, A. E., A. V. Yamshanov, and N. M. Krivdyuk. 2014. Cognitive graphic tools in intelligent training-testing systems. *Open Semantic Technologies for Intelligent Systems (OSTIS-2014) Proceedings*. Minsk: BSUIR. 303–308. (In Russian.)

Received June 26, 2015

Panel matrix and ranking model recovery using mixed-scale measured data

O. Y. Bakhteev
bakhteev@phystech.edu
Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Russia

A decision-making problem is solved in the field of operational research education. The paper presents a method for recovery of changes in ratings of student employees. These ratings are based on interviews at the information technology (IT) training center. A dataset consisting of expert estimates for assessments for different years and overall rating for these students is considered. The scales of the expert estimates vary from year to year, but the scale of the rating remains stable. One should recover the time-independent ranking model. The problem is stated as the object–feature–year panel matrix recovery. It is a map from student descriptions (or their generalized portraits) to expected ratings for all years. Also, a stability of the ranking model produced by the panel matrix is studied. A new method of panel matrix recovery is suggested. It is based on a solution of multidimensional assignment problem. To construct a ranking model, an ordinal classification algorithm with partially ordered feature sets and an algorithm based on support vector machine have been used. The problem is illustrated by the dataset containing the expert assessment of the student interviews at the IT center.

Keywords: *operational research education; business analytics; knowledge extraction; ratings; expert estimates; clustering; mixed scales*

DOI: 10.21469/22233792.1.14.05

1 Introduction

The paper presents a solution for the panel matrix recovery problem, where the panel matrix is a multidimensional object–feature–year [1] matrix. The objects of the matrix are represented by vectors containing different object features for several years. This algebraic structure is used to recover the ranking model and estimate its stability: whenever the parameters of the model remain stable in different years, is considered to be stable. The original dataset is represented by the design matrix, namely, the object–feature matrix, which contains all the object descriptions during all the timestamps.

The main goal of this paper is to develop an algorithm of panel matrix recovery and to recover the ranking model. Let the panel matrix $\mathbf{Z}$ be the matrix, where the entry z_{ijt} is the feature j of the student i in the year t.

The problem of the panel matrix recovery can be found in the pattern recognition, when it is required to recover the tracks of different targets received by sensors [2]. In this paper, another application problem is considered that can be met in business-analytics: an employee selection problem. The dataset containing expert assessments, which were received during the interview at an educational IT-center in 2006–2009, is considered. The purpose is to recover the ranking model and to estimate the stability of this ranking model during all the years. It is proposed to construct some generalized "portraits" of these students and to recover the panel matrix $\mathbf{Z}$ based on these portraits. Note that in this paper, a special case of the panel matrix recovery is considered when the features (answers from assessment) of the portraits remain stable and the only elements that are changeable are the classes of students. The scheme of the panel matrix recovery is shown in Fig. 1.

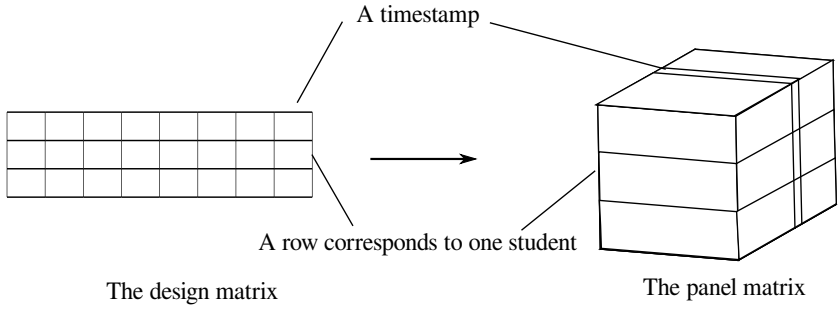


Figure 1 The panel matrix recovery. The generalized student "portraits" that remain stable during all the time are found and considered to be the panel matrix objects

The problem is stated as the multidimensional assignment problem. It requires to find a bijection between object descriptions in different years. The main difficulty in solving this problem is that the multidimensional assignment problem is NP-hard (nondeterministic polynomial-time hard) [3]; therefore, it requires to use heuristic algorithms to solve it. There are several solutions for this problem and related problems [4–6]. The papers [3, 7] propose to use linear programming and randomization algorithms. The methods proposed in the present paper are based on a hypergraph construction. One can use a genetic algorithm [8]. As an alternative, the problem is stated as the common min-cost max flow problem [9].

Define some terms that will be used for the dataset description.

Definition 1. A scale $\mathbb{L}$ is an algebraic structure [10] with a fixed set of operations, relations, and a fixed set of axioms.

Definition 2. A nominal scale $\mathbb{C}$ is a scale with a fixed binary relation:

1) $x = y \vee x \neq y$;
2) $x, y : x = y \Rightarrow y = x$; and
3) $x, y, z : x = y \wedge y = z \Rightarrow x = z$

where x, y, and z are the objects from the scale $\mathbb{C}$: $x, y, z \in \mathbb{C}$.

Definition 3. An oridnal scale $\mathbb{O}$ is a nominal scale with a fixed relation:

1) xRx;
2) $xRy \wedge yRx \Rightarrow x = y$; and
3) $xRy \wedge yRz \Rightarrow xRz$

where $x, y, z \in \mathbb{O}$.

Definition 4. A linear scale $\mathbb{W}$ is an ordinal scale with total order and addition and subtraction operations defined on it.

Definition 5. A ranking function f is a mapping from the object space $\mathbb{X}$ to the finite set of classes $\mathbb{Y}$ with a total order defined on it [11].

The ranking model recovery problem can be met not only in the employee selection but also in information technologies [12], agriculture [13], and energy management [14]. The type of the ranking model recovery algorithm can be chosen with respect to the dataset scale [15–17]. In this paper, a pairwise dominating matrix algorithm is considered for the feature set with a partial order defined on each feature [18]. Another algorithm considered in the present paper

is an algorithm RankSVM, which is a generalization of a classification algorithm based on the support vector machine (SVM) [19].

The ranking model is recovered by the dataset [20] containing students that attempt to pass the interview at the educational center during 2006–2009. The data can contain missing values. The dataset feature descriptions are shown in Table 1.

The expert proposes that each feature should give a positive contribution into the rating. The higher score student gets the higher rating he receives according to the "bigger is better" [21] principle. The nominal feature "Student's interests" is not used in the ranking model recovery, but it is used in the panel matrix recovery in order to cluster students. The expert also recommends to round the feature "Student's interests" in order to get three discrete values.

One of the steps of the panel matrix recovery is a clustering, which requires a distance function. This function determines how close to each other the students estimates are. A generalized Heterogeous Euclidean-Overlap Metric (HEOM) function [22] and Heterogeous Manhattan-Overlap Metric (HMOM) function [23] are proposed for a mixed-scale dataset (a dataset containing linear, ordinal [24], and nominal scales). Extracting significant information from such datasets is a challenging high priority issue for many organizations in the business analytics.

2 The problem formulation

In this section, a formal definition of the panel matrix $\mathbf{Z}$ and ranking model recovery problem are presented.

Definition 6. The panel matrix $\mathbf{Z}$ is a matrix where the entry z_{ijt} is the feature j (answers from assessment) of student i in year t.

The dataset contains the set of pairs of mixed-scale data:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i) : i \in \mathcal{I}\}, \quad \text{the object index } i \in \mathcal{I} = \{1, \ldots, m\},$$

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^{\mathrm{T}}, \quad y_i \in \mathbf{y}$$

with metric

$$d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$$

Table 1 Dataset feature descriptions

Feature	Scale type	Scale cardinality
Average score during university education	Linear, $\mathbb{W}$	Rational number in [3;5]
Average score for the last term		
Acceptance preference (expert estimation)	Ordinal, $\mathbb{O}$	Rational number, the cardinality changes during some years
Student's interests: programming, telecommunication development, or both	Nominal, $\mathbb{C}$	The experts used 3 discrete values {programming, both, telecommunication} in 2006; later, the experts used rational number
Students' responsibility	Ordinal, $\mathbb{O}$	Rational number, the cardinality changes during some years
Level of knowledge		
Motivation		
Student's class — the final rating in the assessment		

where $\mathbb{X} = \mathbb{L}_1 \times \cdots \times \mathbb{L}_n$ is the object space; $\mathbf{X}$ is the object–feature matrix for the dataset; $\mathbf{x}_i \subset \mathbb{X}$; and $\mathbf{y}$ is the vector of classes for each object in dataset such that its elements are in $\mathbb{Y}$. In this paper, the generalized HEOM distance and HMOM distance functions are used as the functions of d (see Eqs. (12) and (14) below). Define a total order on the set of classes:

$$\mathbb{Y} = \{\text{“1”}, \text{“2”}, \text{“3”}, \text{“4”}, \text{“5”}\} \tag{1}$$

where “1” $\prec$ “2” $\prec$ “3” $\prec$ “4” $\prec$ “5”.

Let $\mathcal{T} = \{t\}$ be the set of timestamps of the estimations. In this paper, the set $\mathcal{T}$ contains 4 elements, corresponding to 2006–2009. Let $\mathbf{X}^t$ be the matrix of the objects $\mathbf{X}$ of the year t. Let $\mathbf{D}^t$ be the distance matrix for all pairs of objects per year t:

$$d_{iq}^t = d(\mathbf{x}_i^t, \mathbf{x}_q^t), \quad \mathbf{x}_i^t, \mathbf{x}_q^t \in \mathbf{X}^t.$$

The panel matrix recovery procedure consists of the following parts:

11) a dendrogram constructing algorithm $\mathfrak{d}$;
21) a clustering algorithm $\mathfrak{c}$;
31) a class recovery algorithm $\mathfrak{r}$;
41) a bijection recovery algorithm $\mathfrak{m}$ that finds a bijection between cluster centroids $\mathbf{M}^t$ of different years; and
51) an algorithm $\mathfrak{a}$ of averaging cluster centroids.

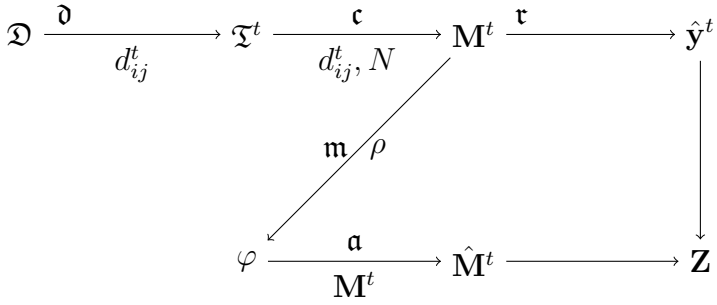


Figure 2 Panel matrix recovery procedure

The panel matrix recovery procedure is shown in Fig. 2: for each year t, the algorithm $\mathfrak{d}$ constructs the dendrogram $\mathfrak{T}^t$. Then, calculate the optimal number of clusters N and the algorithm $\mathfrak{c}$ proceeds clustering. For each cluster $\boldsymbol{\mu}$ from the set of cluster centroids $\mathbf{M}^t$, the algorithm $\mathfrak{r}$ recovers its class $\hat{y}^t \in \hat{\mathbf{y}}^t$. After that, the algorithm $\mathfrak{m}$ finds a bijection φ that matches clusters from different years $\mathbb{Y}$. As a result, get the panel matrix $\mathbf{Z}$ from the averaged centroids $\hat{\mathbf{M}}$, which correspond to the student portraits, and the vector of recovered classes $\hat{\mathbf{y}}^t$.

The algorithm $\mathfrak{c}$ clusters the objects of the dataset for each year t. Let $\mathbf{M}^t \subset \mathbb{X}$ be the set of N cluster centroids for year t, $\mathbf{M}^t = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N]^T$.

For each cluster centroid $\boldsymbol{\mu}_k^t$, recover its class $\hat{y_k}^t \in \mathbb{Y}$ (1). Here, for this purpose, median function has been used:

$$\hat{y}_k^t = \text{median}\{y_i^t : \text{cluster}(\mathbf{x}_i^t) = k\}$$

where $\text{cluster}(\mathbf{x})$ is the function which returns the index of cluster that contains element $\mathbf{x}$.

Let the distance function be given by

$$\rho : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+. \tag{2}$$

This function is used in algorithm $\mathfrak{m}$ to find the mapping that satisfies criteria (5) and (4) (see below). The distance function used as the function of ρ is also described below (15).

The algorithm $\mathfrak{m}$ of the bijection recovery between clusters of different years finds the permutation of cluster indexes:

$$\varphi : \{1, \ldots, N\} \to \{1, \ldots, N\} \tag{3}$$

such that for each year t the mapping is a bijection. Let use the distance function ρ (2) to find this mapping. A set of cluster centroids is called $\mathbf{G}_k = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{|\mathcal{T}|}]$ if it contains all the centroids that φ returns k for them:

$$\mathbf{G}_k = \{\boldsymbol{\mu} \in \mathbb{X} : \varphi(\text{index}(\boldsymbol{\mu})) = k\}$$

where $\text{index} : \mathbf{M}^t \to \{1, \ldots, N\}$ is the function, which returns the index for each cluster.

Let us select φ that minimizes the following criteria:

1. The clustering criterion $C_{\mathcal{C}}$: φ should minimize the average value of R where R is the ratio from the average distance between objects $\boldsymbol{\mu}_{k_1}$ and $\boldsymbol{\mu}_{k_2}$ from cluster set $\mathbf{G}_k$ to the average distance between cluster centroids $\mathbf{G}_{k_1}$ and $\mathbf{G}_{k_2}$:

$$C_{\mathcal{C}} = \text{mean}_{k \in \{1,\ldots,N\}} R(\mathbf{G}_k), \quad R(\mathbf{G}_k) = \frac{\text{mean}_{\boldsymbol{\mu}_{k_1}, \boldsymbol{\mu}_{k_2} \in \mathbf{G}_k} d(\boldsymbol{\mu}_{k_1}, \boldsymbol{\mu}_{k_2})}{\text{mean}_{\mathbf{G}_{k_1}, \mathbf{G}_{k_2}} d(\mathbf{G}_{k_1}, \mathbf{G}_{k_2})}. \tag{4}$$

2. The stability criterion $C_{\mathcal{S}}$: φ should minimize the difference in classes $\hat{y}_{k_1}$ and $\hat{y}_{k_2}$ of cluster set $\mathbf{G}_k$:

$$C_{\mathcal{S}} = \sum_{k=1}^{N} \sum_{\boldsymbol{\mu}_{k_1}, \boldsymbol{\mu}_{k_2} \in \mathbf{G}_k} |\hat{y}_{k_1} - \hat{y}_{k_2}|. \tag{5}$$

The resulting optimization problem is the following:

$$\begin{cases} \varphi = \arg\min\limits_{\varphi' \in \Phi} C_{\mathcal{C}}; \\ \varphi = \arg\min\limits_{\varphi' \in \Phi} C_{\mathcal{S}} \end{cases}$$

where Φ is the set of mappings from the index set $\{1, \ldots, N\}$ to itself such that for each year t, the mapping is bijective.

As an averaging algorithm $\mathfrak{a}$, one gets averaged cluster centroids using the following function:

$$\text{avg}\{\hat{\boldsymbol{\mu}}_{kj}\} = \begin{cases} \text{mean}\{\mu_{qj} : \boldsymbol{\mu}_q \in \boldsymbol{G}_k\} & \text{whenever } \mathbb{L}_j \text{ is linear scale;} \\ \text{median}\{\mu_{qj} : \boldsymbol{\mu}_q \in \boldsymbol{G}_k\} & \text{whenever } \mathbb{L}_j \text{ is ordered scale;} \\ \text{mode}\{\mu_{qj} : \boldsymbol{\mu}_q \in \boldsymbol{G}_k\} & \text{whenever } \mathbb{L}_j \text{ is nominal scale} \end{cases} \tag{6}$$

where $\hat{\boldsymbol{\mu}} \in \hat{\mathbf{M}}$ is the averaged cluster centroid from G_k; and $\hat{\mathbf{M}}$ is the set of averaged cluster centoids. Let use $\hat{\mathbf{M}}$ as an object set for the panel matrix $\mathbf{Z}$.

As a result of the panel matrix recovery procedure, obtain the matrix $\mathbf{Z}$ that contains the set of the averaged centroids $\hat{\mathbf{M}}$ and the vector of recovered classes $\hat{y}_i^t \in \mathbf{y}^t$.

Ranking model recovery

To solve the ranking model recovery problem, one should find a mapping:

$$f : \mathbb{X} \to \mathbb{Y}$$

which minimizes error function $Q(\mathbf{X})$. In this paper, Kendall correlation coefficient [25] has been used:

$$Q(\mathbf{X}) = 1 - \text{KendallTau}(\mathbf{y}, \hat{\mathbf{y}})$$

where $\hat{\mathbf{y}}$ is the vector of classes which is returned for objects $\mathbf{X}$ by the function f; and the Kendall correlation coefficient is

$$\text{KendallTau} = \frac{4|\{(i,q) : y_i > y_q, \hat{y}_i > \hat{y}_q\}|}{m(m-1)} - 1. \tag{7}$$

3 Calculating optimal number of clusters

In the previous section, the number of clusters N was considered to be fixed. One can select the value for N using expert estimates. The other way is to optimize the number of clusters using heuristics. This section describes the optimization problem, which can be used as the one way to find the optimal number of clusters N. Assume the number N of clusters remains stable for each year from the set $\mathcal{T}$. The reason of this assumption is the wish to recover the ranking model for each year of the panel matrix and to estimate correlation between rankings of different years. If the number N differs for different years, this problem is incorrect.

Optimize the number of clusters N using dendrogram constructing algorithm $\mathfrak{d}$.

Definition 7. The dendrogram $\mathfrak{T}^t$ is a tree that is built using the distance matrix $\mathbf{D}^t$ which shows the relationships between clusters.

Describe the dendrogram constructing method. Suppose one has a linkage algorithm:

$$A_{\mathcal{L}} : {\mathbb{R}_+}^m \times {\mathbb{R}_+}^n \to \mathbb{X} \times \mathbb{X}\,. \tag{8}$$

It defines the pair of elements $\mathbf{x}_i$ and $\mathbf{x}_q$ to merge into one cluster $\boldsymbol{\mu}_k$. Let merge this pair and then recalculate the distance matrix $\mathbf{D}^t$ using information about the merged elements.

At the end of dendrogram constructing algorithm, one receives a tree $\mathfrak{T}^t$. Its root contains two last elements merged at the final step.

The example of dendrogram is shown in Fig. 3. The elements A, B, and C are clustering until one cluster remains.

Theorem 1. For each $N \in \{1, \ldots, m\}$, a clustering with a set of N clusters is constructible, where m is the number of objects.

Proof. Each step, the number of clusters is reduced by one. Therefore, after $m - N$ steps, one gets the set of clusters with cardinality equal to N. ∎

Let us construct the dendrogram $\mathfrak{T}^t$ for each year t for optimal N calculating. The number of clusters N is optimal whenever it satisfies the following criteria.

1. The uniform class criterion $C_{\mathcal{U}}$: the number of cluster centroids $\mathbf{M}^t$ of different classes should be equal. N should minimize the deviation of number of different classes of clusters:

$$C_{\mathcal{U}}(\mathbf{M}^t) = \sigma\{|\mathbf{M}^t_y|, y \in \mathbb{Y}\}$$

where $|\mathbf{M}^t_i|$ is the cardinality of the set $\mathbf{M}^t$ with class $y \in \mathbb{Y}$; and σ is the standard deviation.

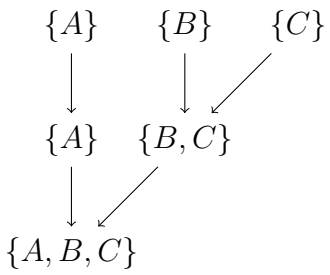


Figure 3 The example of dendrogram

2. Mixing class criterion $C_{\mathcal{M}}$: the number of clusters N should decrease the difference of classes inside clusters:

$$C_{\mathcal{M}}(\mathbf{M}^t) = \text{mean}_{\boldsymbol{\mu}_k \in \mathbf{M}^t} \sigma(\{y_i : \text{cluster}(\mathbf{x}_i) = k\}).$$

The number of clusters N should be less than or equal to the minimum number of objects in the sets: $N \leqslant \min\limits_{t \in \mathcal{T}} |\mathbf{X}^t|$. Also, let construct a clustering that contains a representative of each class; therefore, N should be greater than or equal to the cardinality of $\mathbb{Y}$. The final formula for the optimization problem is the following:

$$\begin{cases} N = \arg\min\limits_{N}(\text{mean}_{t \in \mathcal{T}}(\delta_{\mathcal{U}}(\mathbf{M}^t))); \\ N = \arg\min\limits_{N}(\text{mean}_{t \in \mathcal{T}}(\delta_{\mathcal{M}}(\mathbf{M}^t))); \\ N \geqslant 5, \quad N \leqslant \min\limits_{t \in \mathcal{T}} |\mathbf{X}^t|. \end{cases} \tag{9}$$

Some heuristics have been proposed to select N. Let us construct two dendrograms $\mathfrak{T}^t_{\mathcal{U}}$ for each year t. They use the linkage algorithms (8) $A_{\mathcal{LE}}$ and $A_{\mathcal{LM}}$ to estimate functionals δ_Y and δ_E.

In order to estimate $C_{\mathcal{U}}$, let us use the following linkage algorithm:

$$A_{\mathcal{LU}} = \underset{\substack{\boldsymbol{\mu}_{k_1}, \boldsymbol{\mu}_{k_2} \in \mathbf{M}^t, \\ \hat{y}_{k_1} = \hat{y}_{k_2} = \max\limits_{i = \{1,\ldots,5\}} |\mathbf{M}^t_i|}}{\arg\min} D_{k_1 k_2}.$$

Select a pair of the closest objects of the most common class (the class which has the most number of representatives). Each step, the cardinality of the largest set $\mathbf{M}^t_i$ of cluster centroids of the fixed class y has been reduced. The difference in cardinality between these sets decreases. Therefore, the dendrogram $\mathfrak{T}^t_E$ is quite close to be optimal with respect to δ_Y for the fixed N.

In order to estimate $C_{\mathcal{M}}$, the following linkage algorithm has been used:

$$A_{\mathcal{LM}} = \underset{\substack{\boldsymbol{\mu}_{k_1}, \boldsymbol{\mu}_{k_2} \in \mathbf{M}^t, \\ |\hat{y}_{k_1} - \hat{y}_{k_2}| = \max}}{\arg\min} ||\text{members}(\boldsymbol{\mu}_{k_1})| - |\text{members}(\boldsymbol{\mu}_{k_2})||_2,$$

where members $\mathbf{M}^t \to 2^{\mathbf{X}^t}$ are the functions that return a set of objects assigned to the cluster. This linkage algorithm selects the pair $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ of clusters with the largest difference in classes and with the smallest difference in cardinality. Each step, the difference in classes inside some

cluster has been maximized; therefore, the dendrogram is quite close to be the worst with respect to δ_M for the fixed N.

To find a compromise between two criteria $C_{\mathcal{U}}$ and $C_{\mathcal{M}}$, these criteria for each N have been estimated and ranked. Consider the optimal number of cluster gets minimum of these ranks:

$$N = \arg\min_N(\text{rank}(C_{\mathcal{U}}, N) + \text{rank}(C_{\mathcal{M}}, N))$$

where rank is the function that gives rank for each estimation for current N.

4 Distance functions for mixed-scale data

In this section, distance functions are described for different scale types — linear (10), ordinal (11), nominal (13), and mixed (12) and (14). The distance function for mixed-scale dataset is proposed below.

4.1 Distance function for linear-scale data

Consider the generalized distance function for a linear-scale dataset:

$$r(\mathbf{x}_i, \mathbf{x}_q) = \left(\left(|\mathbf{x}_i - \mathbf{x}_q|^p \right)^T \mathbf{S}^{-1} |\mathbf{x}_i - \mathbf{x}_q|^p \right)^{1/(2p)} \tag{10}$$

where p is the number; $\mathbf{S}$ is the symmetric nonnegative definite matrix (for example, identity matrix $\mathbf{I}$); and exponentiation is proceeded per component: $\mathbf{x}^p = [x_1^p, \ldots, x_n^p]^{\mathrm{T}}$. The Euclidean metric corresponds to this formula with $\mathbf{S} = \mathbf{I}$ and $p = 1$:

$$r(\mathbf{x}_i, \mathbf{x}_q) = \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x}_q)^2)^{1/2}.$$

The Manhattan distance corresponds to this formula with $\mathbf{S} = \mathbf{I}$ and $p = 0.5$:

$$r(\mathbf{x}_i, \mathbf{x}_q) = \sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{x}_q|.$$

4.2 Distance for ordinal-scaled data

Define matrix functions $\mathbf{H}^{j+}$ and $\mathbf{H}^{j-}$ for projection the object set $\mathbf{X}$ on feature j where the scale $\mathbb{L}_j$ is ordinal. Each component of vectors $\mathbf{H}_i^{j+}$ and $\mathbf{H}_i^{j-}$ determine the order between feature j of object i and other objects:

$$(\mathbf{H}_i{}^{j+})_l = \begin{cases} 1 & \text{whenever } x_{ij} \succ x_{lj}; \\ 0 & \text{otherwise;} \end{cases}$$

$$(\mathbf{H}_i{}^{j-})_l = \begin{cases} 1 & \text{whenever } x_{lj} \succ x_{ij}; \\ 0 & \text{otherwise.} \end{cases}$$

Let the distance function pdist be given by:

$$\text{pdist}(x_{ij}, x_{qj}) = \frac{m - \left(\langle \mathbf{H}_i{}^{j+}, \mathbf{H}_q{}^{j+} \rangle + \langle \mathbf{H}_i{}^{j-}, \mathbf{H}_q{}^{j-} \rangle \right)}{m} \tag{11}$$

where m is the number of objects in the dataset.

Theorem 2. If $\mathbb{L}_j$ is a totally ordered set, then pdist is a metric.

Proof. At first, let us prove that the range of the function is in $[0;1]$. Let x_{ij} be less than or equal to x_{qj} : $x_{ij} \leqslant x_{qj}$. Then

$$\langle \mathbf{H}_i{}^{j+}, \mathbf{H}_q{}^{j+} \rangle = ||\mathbf{H}_i^{j+}||_2^2, \quad \langle \mathbf{H}_i{}^{j-}, \mathbf{H}_q{}^{j-} \rangle = ||\mathbf{H}_q^{j-}||_2^2;$$

$$\text{pdist}(x_{ij}, x_{qj}) = \frac{m - ||\mathbf{H}_i^{j+}||_2^2 - ||\mathbf{H}_q^{j-}||_2^2}{m}.$$

The maximum of the function is not more than 1. The function pdist gets minimum whenever $x_{ij} = x_{qj}$, $\text{pdist}(x_{ij}, x_{ij}) = 0$. The function is symmetric. Let us prove that the function satisfies the subadditivity condition for each $\mathbf{x}_w \in \mathbb{X}$:

$$\text{pdist}(x_{ij}, x_{qj}) \leqslant \text{pdist}(x_{ij}, x_{wj}) + \text{pdist}(x_{wj}, x_{qj}).$$

The proof contains 3 cases:

$$x_{ij} \leqslant x_{qj} \leqslant x_{wj}; \quad x_{wj} \geqslant x_{ij} \geqslant x_{qj}; \quad x_{ij} \leqslant x_{wj} \leqslant x_{ij}.$$

Consider the first case, other cases can be proved similarly:

$$\text{pdist}(x_{ij}, x_{wj}) + \text{pdist}(x_{qj}, x_{wj}) = \frac{2m - ||\mathbf{H}_i^{j+}||_2^2 - ||\mathbf{H}_q^{j+}||_2^2 - 2||\mathbf{H}_w^{j-}||_2^2}{m}$$

$$\geqslant \frac{2m - ||\mathbf{H}_i^{j+}||_2^2 - ||\mathbf{H}_q^{j+}||_2^2 - 2||\mathbf{H}_q^{j-}||_2^2}{m}$$

$$= \frac{2m - ||\mathbf{H}_i^{j+}||_2^2 - m + ||\mathbf{H}_q^{j-}||_2^2 - 2||\mathbf{H}_q^{j-}||_2^2}{m} = \frac{m - ||\mathbf{H}_i^{j+}||_2^2 - ||\mathbf{H}_q^{j-}||_2^2}{m} = \text{pdist}(x_{ij}, x_{qj}). \quad \blacksquare$$

4.3 The generalization of HEOM and HMOM distance functions

Supplement the HEOM [22] function for ordinal-scale datasets:

$$d_1(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{n} r(x_{ij}, x_{qj})^2 \right)^{1/2} \tag{12}$$

where

$$\left.\begin{aligned} r(x_{ij}, x_{qj}) &= \begin{cases} \text{overlap}(x_{ij}, x_{qj}) & \text{whenever } \mathbb{L}_j \text{ is a nominal scale;} \\ \text{pdist}(x_{ij}, x_{qj}) & \text{whenever } \mathbb{L}_j \text{ is an ordinal scale;} \\ \text{diff}(x_{ij}, x_{qj}) & \text{otherwise;} \end{cases} \\ \text{overlap}(x_{ij}, x_{qj}) &= \begin{cases} 1 & \text{whenever } x_{ij} \neq x_{qj}; \\ 0 & \text{otherwise;} \end{cases} \\ \text{diff}(x_{ij}, x_{qj}) &= \frac{|x_{ij} - x_{qj}|}{\max\limits_{\mathbb{L}_j} - \min\limits_{\mathbb{L}_j}}, \end{aligned}\right\} \tag{13}$$

the function $\text{diff}(x_{ij}, x_{qj})$ is determined by normalized difference between two values of feature j.

The range of the resulting function d is less than or equal to the square root of the feature number: $d(\mathbf{x}_i, \mathbf{x}_j) \leqslant \sqrt{n}$.

The difference between HEOM and HMOM modifications is only in lack of exponentiation:

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{n} r(x_{ij}, x_{qj}). \tag{14}$$

5 Panel matrix recovery procedure

5.1 Clustering algorithm $\mathfrak{c}$

Let use a modification of k-means [26] algorithm as the clustering algorithm $\mathfrak{c}$. This algorithm is iterative. At first, select N cluster centroids $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ randomly. Each iteration assign each object $\mathbf{x}_i$ from dataset $\mathbf{X}^t$ to the closest cluster in the sense of the distance function d:

$$\text{cluster}(\mathbf{x_i}) = \arg\min_{k \in \{1,\ldots,N\}} d(\mathbf{x}_i, \boldsymbol{\mu}_k)$$

where $\text{cluster}(\mathbf{x})$ is the function that returns a cluster index for each object $\mathbf{x}$. After that, recalculate cluster centroids:

$$\mu_{kj} = \text{avg}\{x_{ij}, \text{cluster}(\mathbf{x}_i) = k\}.$$

Use avg function (6) such that corresponds to scale types instead of arithmetic mean recommended in the k-means algorithm:

$$\text{avg}\{\mathbf{x}_{i_1 j}, \ldots, \mathbf{x}_{i_p j}\} = \begin{cases} \text{mean}\{\mathbf{x}_{i_1 j}, \ldots, \mathbf{x}_{i_p j}\} & \text{whenever } \mathbb{L}_j \text{ is a linear scale;} \\ \text{median}\{\mathbf{x}_{i_1 j}, \ldots, \mathbf{x}_{i_p j}\} & \text{whenever } \mathbb{L}_j \text{ is an ordinal scale;} \\ \text{mode}\{\mathbf{x}_{i_1 j}, \ldots, \mathbf{x}_{i_p j}\} & \text{whenever } \mathbb{L}_j \text{ is a nominal scale.} \end{cases}$$

5.2 Bijection recovery algorithm $\mathfrak{m}$

In this section, two methods of the function φ (3) finding are considered: the reducing the problem to the transport problem and the genetic algorithm.

Let state the problem of finding φ as multidimensional assignment problem [7]. Construct $|\mathcal{T}|$-partite hypergraph $\langle V, E \rangle$, $V = V^1 \cup \cdots \cup V^{|\mathcal{T}|}$ where $\mathcal{T}$ is the set of years. The vertices of each partite sets V^t correspond to the set of cluster centroids $\mathbf{M}^t$ for year t. The hyperedges of the hypergraph correspond to the all subsets of cluster centroids that contain $|\mathcal{T}|$ cluster centroids and correspond to the condition that each hyperedge $e \in E$ contains only one cluster centroid for each year. Let the weight of each hyperedge be given by:

$$w_e = \sum_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in e} \rho(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2).$$

It is required to find a maximal set of hyperedges where each pair of this set does not intersect and where the sum of the hyperedge weights is minimal.

5.3 Reducing the φ finding problem to the transport problem

Consider a two-dimensional assignment problem, where it is required to find the bijection between two sets of objects. This problem can be stated as the min-cost max flow problem by constructing a transport directed graph [27]. The vertices of this graph correspond to the cluster centroids $\boldsymbol{M}^t$ with capacity equal to one and edge weights equal to the distance between cluster centroids: $\rho(\boldsymbol{\mu}_i^{t_1}, \boldsymbol{\mu}_j^{t_2})$. After reducing the problem, it is required to find the maximal flow with minimal edge weight sum, which is called the cost of the flow. In the considered case, one can construct a hypergraph $\langle V, E \rangle$ instead of the directed graph whose hyperedge configuration was described above.

In order to find the maximal flow of minimal cost of the hypergraph $\langle V, E \rangle$, let transform the hypergraph into a directed graph $\langle V', E' \rangle$ and use common algorithms for directed graphs.

There are some heuristic algorithms of hypergraph to directed graph transformation that can be used for this case [28, 29].

5.4 Genetic algorithm

As an alternative method of finding the function φ, let use the genetic algorithm [8]. Each solution of the problem is represented by a hypergraph with N hyperedges such that each pair of hyperedge does not intersect and each hyperedge contains only one cluster centroid for each year. Let $\mathbf{S}^{qk}$ be a matrix for the solution k of the generation q. The entry S_{ij}^{qk} is the index number of cluster of the year j in the hyperedge i:

$$S_{ij}^{qk} = \text{whenever } \boldsymbol{\mu}_l^j \in e_i$$

where $e_i \in E$. The starting population $\mathbf{S}^1$ is generated randomly, its cardinality s_1 is a structural parameter. Each new generation is generated from the older one by application the special procedures: mutation, crossover, and selection.

As the crossover of the generation q, the following procedure has been used. Select two solutions $\mathbf{S}^{qk_1}$ and $\mathbf{S}^{qk_2}$ from this generation randomly. Also, select a row l_1 from the first matrix and a row l_2 from the second matrix, the number of columns to modify col, and a set of column indexes $\{c_{\text{perm}_{(1)}}, \ldots, c_{\text{perm}_{(\text{col})}}\}$, where perm is a random permutation. For each column c_i in $\{c_{\text{perm}_{(1)}}, \ldots, c_{\text{perm}_{(\text{col})}}\}$ and for both matrices, proceed the permutation given by $\mathbf{S}_{l_1 c_i}^{qk_1} \leftrightarrow \mathbf{S}_{l_2 c_i}^{qk_2}$.

After crossover procedure, mutations. Select a solution $\mathbf{S}^{qk}$ and a column c randomly. After that, proceed random permutation on all the elements of column c. Such procedure helps one to avoid stopping algorithm in local extrema. After mutations and crossovers, select the best solution generation $\mathbf{S}^{q+1}$ in the sense of the distance function ρ. The number of mutations per generation f_{mutation}, the number of crossovers per generation $f_{\text{crossover}}$, and the generation cardinalities s_q are the structural parameters of the algorithm. The algorithm stops whenever the generation satisfies the stopping criterion $C_{\mathcal{F}}$. In this paper, stopping criterion is used:

$$C_{\mathcal{F}} = (K_{\text{av}} > \hat{K}_{\text{av}}) \text{ or } (K_{\text{av}} \text{ does not change after few iterations})$$

where

$$K_{\text{av}} = \text{mean}_{t_1,t_2 \in \mathcal{T}, t_1 \neq t_2}(\text{KendallTau}(\mathbf{S}_{1,\ldots,N,t_1}^{q(1)}, \mathbf{S}_{1,\ldots,N,t_2}^{q(1)})),$$

$\mathbf{S}^{q(1)}$ is the best solution of the current generation in the sense of ρ, $\mathbf{S}_{1,\ldots,N,t}^{qi}$ is the column t of the matrix $\mathbf{S}^{qi}$. $\hat{K}_{\text{av}}$ is a structural parameter, which represents required average Kendall correlation coefficient in the panel matrix $\mathbf{Z}$.

5.5 Defining hyperedge weight

Use the sum of the generalized distances (12) and (14) between cluster centroids as the hyperedge weights:

$$\rho_1(\mathbf{x}_i, \mathbf{x}_j) = \Big(\frac{\sum_{k=1}^n d(x_{ik}, x_{jk})^2}{n} + \text{pdist}^2(y_i, y_j) \cdot \text{coef}\Big)^{1/2}; \tag{15}$$

$$\rho_2(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k=1}^n d(x_{ik}, x_{jk})}{n} + \text{pdist}(y_i, y_j) \cdot \text{coef} \tag{16}$$

where coef is the parameter which regulates the balance between priority of the stability criterion (5) and the clustering criterion (4). Whenever coef = 1, these criteria priorities are equal. Let use (15) in the experiment with the generalized HEOM metric and (16) in the experiment with the generalized HMOM metric.

5.6 Complexity analysis of the algorithm

The clustering algorithm complexity can be bounded to $O(Nnm \cdot \text{iter})$ where iter is the number of iterations of the clustering algorithm.

The complexity of one crossover series can be bounded to $O(f_{\text{crossover}}|\mathcal{T}|)$. The complexity of a mutation series is $O(f_{\text{mutation}}N)$; so, the naive estimation of the genetic algorithm iteration is $O(f_{\text{crossover}}|\mathcal{T}| + f_{\text{mutation}}N)$.

6 The ranking model recovery

In this section, the methods of the ranking model recovery used in this paper are described. Consider three ranking algorithms: the ordinal classification algorithm using partially ordered feature sets [18] and rankSVM [30], the algorithm based on the SVM [19], and an algorithm based on the method of least squares in order to compare the results of the ranking model recovery.

6.1 The ordinal classification algorithm using partially ordered feature sets

In this subsection, suppose that the class of the object is also a feature with number 0: $\mathbb{Y} = \mathbb{L}_0, x_{i0} = y_i, \mathbf{x}_i \in \mathbf{X}$. For each feature $\mathbb{L}_q$, construct a matrix $\mathbf{U}_q$ which determines the order of the feature q:

$$U_q(i,j) = \begin{cases} 1 & \text{if } x_{iq} \prec x_{jq}; \\ 0 & \text{otherwise}. \end{cases}$$

Estimate the matrix $\boldsymbol{\psi}$ using feature matrices $\mathbf{U}_q, q \in \{0, \ldots, m\}$. This matrix $\boldsymbol{\psi}$ is called a pairwise dominance matrix:

$$\hat{\psi_{ij}} = \sum_{k=1}^{n} w_k U_k(i,j);$$

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{i=1}^{m}\sum_{k=1}^{m} \left(U_0(i,k) - \sum_{j=1}^{n} w_j U_j(i,k) \right)^2$$

where $\mathbf{w}$ is the weight vector for feature matrices.

After that, estimate the class $\hat{y}$ of the object using the pairwise dominance matrix:

$$\hat{y} = f(\hat{\psi}, \lambda), \quad \lambda = \arg\min_{\lambda} ||y - \hat{y}||_2.$$

In this paper, a logistic regression is used for $\mathbf{w}$ and λ estimations. Also, propose that $\lambda_i = \lambda_j$ whenever $y_i = y_j$.

6.2 The RankSVM algorithm

This algorithm is a generalization of the classification algorithm based on SVM [19]. The optimization problem for this algorithm is given by

$$||\mathbf{w}||_2 + C \sum_{i,j} \xi_{ij} \to \min,$$

$$\text{for each } \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}, y_y > y_j : K(\mathbf{w}, \mathbf{x}_i) \geqslant K(\mathbf{w}, \mathbf{x}_j) + 1 - \xi_{ij}$$

where

$$K : \mathbb{R}^n \times \mathbb{X} \to \mathbb{R} \tag{17}$$

is the kernel function, commonly the dot product; ξ_{ij} and C are the parameters. This optimization problem can be reduced to the classification SVM optimization problem [30] and solved by standard methods [19].

The most interesting feature of this algorithm is the use of different kernel functions K instead of dot product. This modifies original object space and makes it more similar to linearly separable space. In this paper, the following kernel functions have been used:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j^{\mathrm{T}}; \tag{18}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\frac{1}{n}\mathbf{x}_i \cdot \mathbf{x}_j^{\mathrm{T}}\right)^3; \tag{19}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{n}|\mathbf{x}_i - \mathbf{x}_j|^2\right); \tag{20}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh\left(\frac{1}{n}\mathbf{x}_i \cdot \mathbf{x}_j^{\mathrm{T}}\right). \tag{21}$$

6.3 The algorithm based on least squares method

Use this algorithm as a basic ranking algorithm. The main idea of this algorithm is in finding coefficients $\alpha_1, \ldots, \alpha_n$, which solve the optimization task:

$$\Delta = \sum_{i=1}^{m} \left\| y_i - \sum_{j=1}^{n} \alpha_j x_{ij} \right\|_2 \to \min.$$

The resulting function is given by

$$f(\mathbf{x}) = \begin{cases} \text{round}\left(\sum_{j=1}^{n} \alpha_j x_i j\right) & \text{whenever round}\left(\sum_{j=1}^{n} \alpha_j x_{ij}\right) \in \{1,2,3,4,5\}; \\ 5 & \text{whenever round}\left(\sum_{j=1}^{n} \alpha_j x_{ij}\right) > 5; \\ 1 & \text{whenever round}\left(\sum_{j=1}^{n} \alpha_j x_{ij}\right) < 1. \end{cases}$$

6.4 Transforming ordinal features into linear features

In order to use the information gathered from ordinal features, the following approach has been used [31]. Each ordinal feature with scale $\mathbb{L}_j$ is proposed to match with some latent linear feature with scale $\mathbb{L}_j^*$, which can be recovered by the following rule:

$$x_{ij} = l_{ju} \text{ whenever } l^*_{ju-1} \leqslant x^*_{ij} \leqslant l_{ju}$$

where l_{ju} is the u value of the set of values of $\mathbb{L}_j$ sorted ascendingly; x^*_{ij} is the value of latent variable; l^*_{ju} is the threshold:

$$l_{ju} = \Psi^{-1}(F_j(l_{ju})), \quad F_j(l) = \sum_{\mathbf{x}_i \in \mathbf{X}, x_{ij} \prec l} \frac{1}{m},$$

$$l_{j0} = -\infty, \quad l_{j|\mathbb{L}_j|} = \infty$$

with Ψ^{-1} being the inverse normal distribution. This transformation matches the ordinal feature with some real-valued intervals. Use the upper limit of the intervals as a representer of the latent

linear feature, i. e., $x^*_{ij} = l_{ju}$ whenever $l^*_{ju-1} \leqslant x^*_{ij} \leqslant l_{ju}$. Let the value corresponding to the largest value of the ordinal feature be $x^*_{ij} = l_{j|\mathbb{L}_j|-1} + \text{mean}(\{l_{ju} - l_{ju-1}, u \in \{1, \ldots, |\mathbb{L}_j| - 1\})$.

7 Computational experiment

In this section, the results of the experiment are presented and conclusions on the applicability of the proposed algorithm to the considered problem are drawn. The main goal of the present experiment is to confirm or deny the efficiency of the described panel matrix recovery method and recover the ranking model in the most efficient way. The dataset [20] contains a table with 284 student assessments. Each assessment contains 7 features, the class of the student, and the year of the interview. The source of the computational experiment is available at [32].

For the experiment, the following software was used:

- GNU Octave v.3.8.1;
- SVM$^{\text{light}}$ v.6.02.;
- batch high throughput multidimensional scaline for MATLAB/GNU Octave programming language; and
- Python v.2.7. with NumPy and scikit-learn packages.

7.1 Panel matrix recovery

In order to handle with missing values, k-nearest neighbors algorithm has been used for missing values imputation [33]. $k = 3$ was chosen using cross-validation. The optimal number of clusters N has been estimated by solving the optimization problem (9) and the result $N = 20$ has been got. The results of clustering for year 2007 are shown in Fig. 4. The coordinates of the objects were received by projection the data $\mathbf{X}^t$ onto two-dimensional space $\{\xi_1, \xi_2\}$ using High-Throughput Multidimensional Scaling [34] method. The colors of the plot correspond to different cluster indexes.

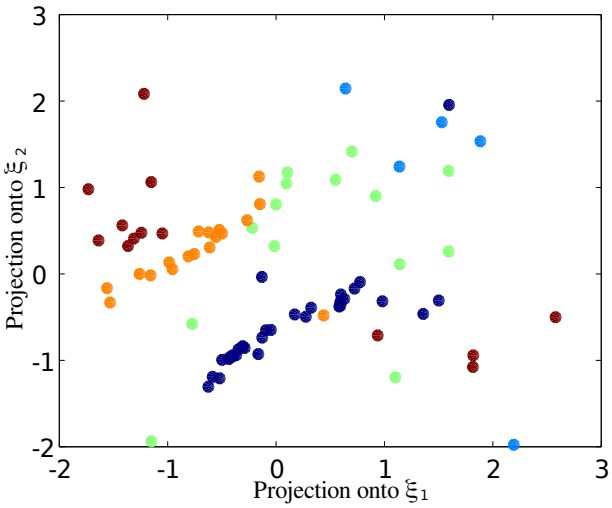


Figure 4 The result of clustering for year 2007

In order to reduce the randomnicity in the experiment, 10 tests have been proceeded and the results have been averaged. The parameter coef (15) was set to 1. The cardinality of the starting population s_1 was set to $N \cdot |\mathcal{T}|$. For each generation $\mathbf{S}^q$, $|\mathcal{T}| \cdot s^q$ mutations and $\binom{s^q}{2} = s^q(s^q - 1)/2$ crossovers have been proceeded. Such parameter values give an availability to crossover all the pairs of solutions and to mutate each column of each hypergraph matrix. The

cardinality of all the generations remained stable: $s^{q+1} = s^q$. The required average Kendall coefficient $\hat{K}_{\text{av}}$ was set to 0.85.

The results of the panel matrix recovery have been estimated by the Kendall correlation coefficient (7). The results of the Kendall correlation for the experiments with HEOM and HMOM metrics are represented in Tables 2 and 3. The computational experiment shows that the proposed algorithm of the panel matrix recovery gives good results on the considered dataset.

Table 2 Kendall coefficient for panel matrix **Z** recovery with HEOM metric

Year	2006	2007	2008	2009
2006	1	0.85629	0.80154	0.85270
2007	0.85629	1	0.84301	0.85728
2008	0.80154	0.84301	1	0.84731
2009	0.85270	0.85728	0.84731	1

Table 3 Kendall coefficient for panel matrix **Z** recovery with HMOM metric

Year	2006	2007	2008	2009
2006	1	0.87714	0.76905	0.77979
2007	0.87714	1	0.82962	0.80129
2008	0.76905	0.82962	1	0.82266
2009	0.77979	0.80129	0.82266	1

The mean of pairwise Kendall coefficients for the panel recovery with HMOM is 0.81326 while for HEOM, this value is 0.84302. Therefore, HEOM metric is quite more efficient for the considered purpose. As one can see, the panel matrix recovery gives rather stable results for all the years.

7.2 The ranking model

The difference between the real class y of an object $\mathbf{x}$ and the recovered class $\hat{y}$ has been used as the error function Q. The algorithms were tested using "Leave one out" method. Different kernel functions (17) have been used during the RankSVM algorithm testing.

The results of the experiment are shown in Table 4.

The RankSVM algorithm showed the best result and was selected as the ranking model recovery algorithm. Another good result was received from the algorithm based on pairwise-dominating matrix.

7.3 Computation for the simulated data

Investigate the performance of the proposed algorithm. Conduct two series of the experiments: the series with adding noise into object features and adding noise into object classes.

Table 4 Results of the ranking model recovery

Year	2006	2007	2008	2009	Mean value
LS-algrotihm	0.7	0.57	0.68	0.62	0.64
Pairwise-dominating matrix	1.2176	1.1412	1.2647	1.2235	1.2118
RankSVM, Eq. (18)	0.55	0.52	0.62	0.60	**0.58**
RankSVM, Eq. (19)	1,2741	0.98	1.3451	1.1667	1.1914
RankSVM, Eq. (20)	0.7511	0.5413	0.7285	0.7501	0.6927
RankSVM, Eq. (21)	1.2741	0.98	1.3451	1.1667	1.1914

During the experiment with adding noise into features, change each object feature value randomly with probability from 10% to 50%.

During the experiment with adding noise in classes, replace the class of each object by constant for each year. In these experiments, HEOM metric has been used.

The results of the experiments are shown in Figs. 5 and 6.

Figure 5 shows the mean of Kendall correlation coefficient values for each pairs of years. The genetic algorithm uses the combination of two criteria for selecting optimal solution. After adding noise into objects' features, the algorithm tries to optimize the matching of classes for different years. The experiment with adding noise into object classes shows the opposite case — the dataset lost the uniformity of classes per years and, therefore, the average error did not increase so dramatically as in the first experiment.

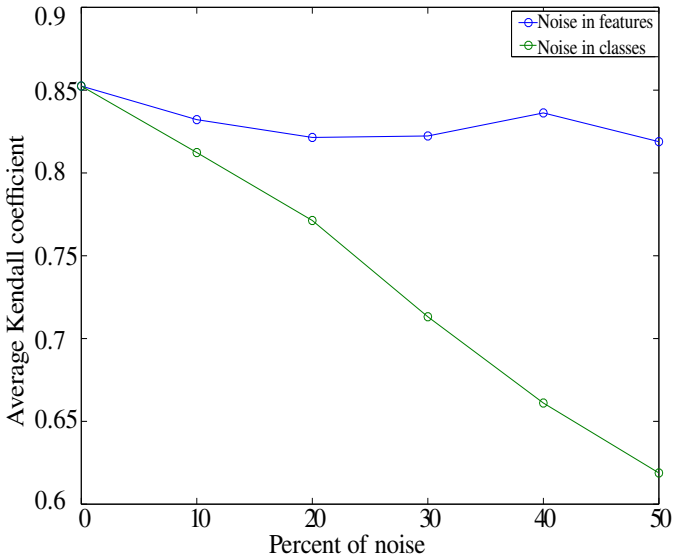


Figure 5 Average Kendall coefficient

In order to estimate the quality of the ranking model recovery in datasets with noise, the ranking model recovery has been tested on the simulated datasets generated in the first experiment series.

The result of the ranking model recovery is shown in Fig. 7. As one can see, the RankSVM algorithm and pairwise-dominating algorithm give the similar results.

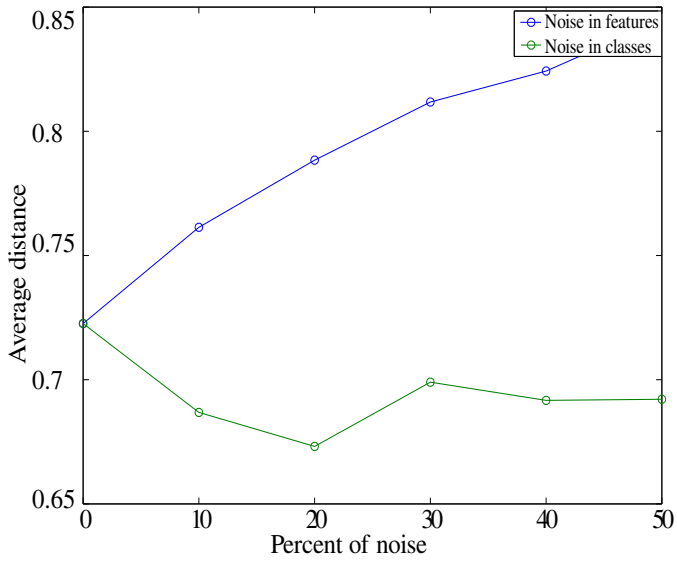


Figure 6 Average distance between cluster centroids that were mapped by the bijection

The error of the least squares-based algorithm increases dramatically; therefore, it is better not to use it if the dataset contains significant amount of noise.

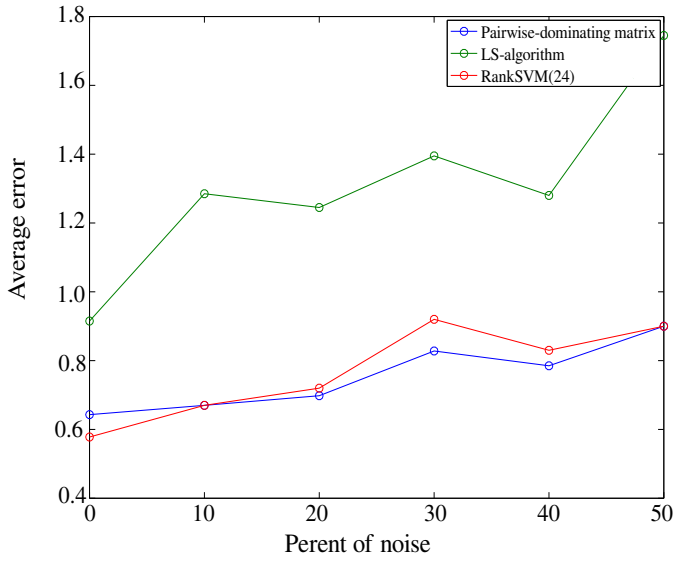


Figure 7 The results of rank model recovery on simulated data

8 Concluding remarks

In this paper, the method of the panel matrix recovery has been proposed. The heuristic method of calculating optimal number of clusters has been suggested for clustering objects per year.

Two algorithms have been considered to construct a bijection between clusters of different years based on reducing this problem to multidimensional assignment problem — the ge-

netic algorithm and the algorithm based on the reducing the problem to the transport problem.

The experiment for the panel matrix and ranking model recovery using genetic algorithm was proceeded. Two metric functions were compared. The HEOM metric showed the best result. The experiment showed that the panel matrix was stable in the sense of ranking model stability. The best result of ranking model recovery was shown by the RankSVM algorithm.

9 Acknowledgments

The author would like to thank Dr. Vadim Strijov for the formal problem statement, useful comments, suggestions, and the attention drawn to the present work.

References

[1] Davies, A., and K. Lahiri. 1995. A new framework for testing rationality and measuring aggregate shocks using panel data. *J. Econometrics* 68(1):205–227.

[2] Capponi, A., and H. de Waard. 2004. A polynomial time algorithm for the multidimensional assignment problem in multiple sensor environments. *7th Conference (International) on Information Fusion*. 1150–1157.

[3] Pardalos, P., and L. Pitsoulis. 2001. *Nonlinear assignment problems: Algorithms and applications (combinatorial optimization)*. Springer. 303 p.

[4] Aronson, J. E. 1986. The multiperiod assignment problem: A multicommodity network flow model and specialized branch and bound algorithm. *Eur. J. Oper. Res.* 23(3):367–381.

[5] Fréville, A. 2004. The multidimensional 0–1 knapsack problem: An overview. *Eur. J. Oper. Res.* 155(1):1–21. doi: http://dx.doi.org/10.1016/S0377-2217(03)00274-1

[6] Walteros, J. L., C. Vogiatzis, E. L. Pasiliao, and P. M. Pardalos. 2014. Integer programming models for the multidimensional assignment problem with star costs. *Eur. J. Oper. Res.* 235(3):553–568. doi: http://dx.doi.org/10.1016/j.ejor.2013.10.048

[7] Kuroki, Y., and T. Matsui. 2009. An approximation algorithm for multidimensional assignment problem minimizing the sum of squared errors. *Discrete Appl. Math.* 157:2124–2135.

[8] Sahu, A., and R. Tapadar. 2007. Solving the assignment problem using genetic algorithm and simulated annealing. *IAENG Int. J. Appl. Math.* 36(1). Available at: http://www.iaeng.org/IJAM/issues_v36/issue_1/IJAM_36_1_7.pdf (accessed January 9, 2016).

[9] Pistorius, J., and M. Minoux. 2003. An improved direct labeling method for the max-flow min-cut computation in large hypergraphs and applications. *Int. Trans. Oper. Res.* 10(1):1–11.

[10] Cooke, D. J., and H. E. Bez. 1984. *Computer mathematics*. 1st ed. Cambridge University Press. 408 p.

[11] Johannes, F., and H. Eyke. 2011. *Preference learning: An introduction*. Springer. 454 p.

[12] Albadvi, A. 2004. Formulating national information technology strategies: A preference ranking model using PROMETHEE method. *Eur. J. Oper. Res.* 153(2):290–296. doi: http://dx.doi.org/10.1016/S0377-2217(03)00151-6

[13] Siskos, Y., N. F. Matsatsinis, and G. Baourakis. 2001. Multicriteria analysis in agricultural marketing: The case of French olive oil market. *Eur. J. Oper. Res.* 130(2):315–331. doi: http://dx.doi.org/10.1016/S0377-2217(00)00043-6

[14] Mladineo, N., J. Margeta, J. P. Brans, and B. Mareschal. 1987. Multicriteria ranking of alternative locations for small scale hydro plants. *Eur. J. Oper. Res.* 31(2):215–222. doi: http://dx.doi.org/10.1016/0377-2217(87)90025-7

[15] Strijov, V. V. 2011. Utochnenie ekspertnykh otsenok, vystavlennykh v rangovykh shkalakh, s pomoshch'yu izmeryaemykh dannykh [Clarification of expert estimations in rank scale using measured data]. *Zavodskaya laboratoriya. Diagnostika materialov* [Factory Laboratory. Material Diagnostics] 77(7):72–78. (In Russian.)

[16] Medvednikova, M. M. 2012. Ispol'zovanie metoda glavnykh komponent pri postroenii integral'nykhx indikatorov [Using principal component analysis in construction of integral indicators]. *Machine Learning Data Anal.* 1(3):292–304. (In Russian.)

[17] Medvednikova, M. M., V. V. Strijov, and M. P. Kuznetsov. 2012. Algoritm mnogoklassovoy monotonnoy Pareto-klassifikatsii s vyborom priznakov [An algorithm of multiclass monotonic Pareto-classification with feature selection]. *Proceedings of the Tula State University, Natural Sciences* 3:132–141. (In Russian.)

[18] Kuznetsov, M. P., and V. V. Strijov. 2014. Methods of expert estimations concordance for integral quality estimation. *Expert Syst. Appl.* 41(4, Pt. 2):1551–2110. doi: http://dx.doi.org/10.1016/j.eswa.2013.08.095

[19] Vorontsov, K. V. Support vector machine lectures. Available at: http://www.ccas.ru/voron/download/SVM.pdf (accessed July 22, 2014). (In Russian.)

[20] The original dataset. Available at: http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Bakhteev014UniversityRanking/data/data.csv?format=raw (accessed August 27, 2014).

[21] Strijov, V. V. 2006. Utochnenie ekspertnykh otsenok s pomoshch'yu izmeryaemykh dannykh [Clarification of expert estimations using measured data]. *Zavodskaya laboratoriya. Diagnostika materialov* [Factory Laboratory. Material Diagnostics] 72(7):59–64. (In Russian.)

[22] Wilson, D. R., and T. R. Martinez. 1997. Improved heterogeneous distance functions. *J. Artif. Intell. Res.* 6:1–34.

[23] Batista, G. E. A. P. A., and D. F. Silva. 2009. How k-nearest neighbor parameters affect its performance. *Argentine Symposium on Artificial Intelligence Proceedings*. 1–12.

[24] Walesiak, M. 1999. Distance measure for ordinal data. *Argum. Oecon.* 2(8):167–173.

[25] Prokhorov, A. V. (originator) Kendall coefficient of rank correlation. *Encyclopedia of mathematics*. Available at: http://www.encyclopediaofmath.org/index.php?title =Kendall_coefficient_of_rank_correlation&oldid=13189 (accessed August 27, 2014).

[26] Steinhaus, H. 1956. Sur la division des corps materiels en parties. *Bull. Acad. Pol. Sci.* 4:801–804.

[27] Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein. 2009. *Introduction to algorithms*. 3rd ed. MIT Press. 1312 p.

[28] Agarwal, S., K. Branson, and S. Belongie. 2006. Higher order learning with graphs. *23rd Conference (International) on Machine Learning Proceedings*. 17–24. doi: http://dx.doi.org/10.1145/1143844.1143847

[29] Pu, L., and B. Faltings. 2012. Hypergraph learning with hyperedge expansion. *European Conference on Machine Learning and Knowledge Discovery in Databases Proceedings*. 1:410–425. doi: http://dx.doi.org/10.1007/978-3-642-33460-3_32

[30] Joachims, T. 2002. Optimizing search engines using clickthrough data. *8th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings*. 133–142. doi: http://dx.doi.org/10.1145/775047.775067

[31] Winship, C., and R. Mare. 1984. Regression models with ordinal variables. *Am. Sociol. Rev.* 49(4):512–525. doi: http://dx.doi.org/10.2307/2095465

[32] Algorithms of machine learning. Available at: http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Bakhteev014UniversityRanking/code/ (accessed August 27, 2014).

[33] Batista, G. E. A. P. A., and M. C. Monard. 2003. An analysis of four missing data treatment methods for supervised learning. *Appl. Artif. Intell.* 17(5-6):519–533. doi: http://dx.doi.org/10.1080/713827181

[34] Strickert, M., S. Teichmann, N. Sreenivasulu, and U. Seiffert. 2005. High-Throughput Multi-dimensional Scaling (HiT-MDS) for cDNA-Array expression data. *15th Conference (International) on Artificial Neural Networks: Biological Inspirations Proceedings*. 1:625–633. doi: http://dx.doi.org/10.1007/11550822_97

Received December 20, 2015

Восстановление панельной матрицы и ранжирующей модели по метризованной выборке в разнородных шкалах

О. Ю. Бахтеев

bakhteev@phystech.edu

Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9

Работа посвящена восстановлению ежегодных изменений рейтингов студентов при собеседовании в учебный центр. Рассматривается выборка, состоящая из экспертных оценок студентов, проходивших собеседование в учебный центр в течение нескольких лет и итоговых рейтингов студентов. Шкалы экспертных оценок меняются из года в год, но шкала рейтингов остается неизменной. Требуется восстановить ранжирующую модель, не зависящую от времени. Задача сводится к восстановлению панельной матрицы (т. е. матрицы объект–признак–год), ставящей во взаимное соответствие некоторого студента (или усредненный «портрет» студента) и его предполагаемую оценку на собеседованиях за каждый год, и исследованию ранжирующей модели, полученной на основе этой матрицы, а также анализу ее устойчивости на протяжении нескольких лет. Предлагается метод восстановления панельной матрицы, основанный на решении многомерной задачи о назначениях. В качестве метода восстановления ранжирующей модели используется алгоритм многоклассовой классификации с отношением полного порядка на классах.

Ключевые слова: *рейтинги; экспертные оценки; кластеризация; смешанные шкалы*

DOI: 10.21469/22233792.1.14.05

Литература

[1] *Davies A., Lahiri K.* A new framework for testing rationality and measuring aggregate shocks using panel data // J. Econometrics, 1995. Vol. 68. No. 1. P. 205–227.

[2] *Capponi A., de Waard H.* A polynomial time algorithm for the multidimensional assignment problem in multiple sensor environments // 7th Conference (International) on Information Fusion, 2004. P. 1150–1157.

[3] *Pardalos P., Pitsoulis L.* Nonlinear assignment problems: Algorithms and applications (combinatorial optimization). — Springer, 2001. 303 p.

[4] *Aronson J. E.* The multiperiod assignment problem: A multicommodity network flow model and specialized branch and bound algorithm // Eur. J. Oper. Res., 1986. Vol. 23. No. 3. P. 367–381.

[5] *Fréville A.* The multidimensional 0—1 knapsack problem: An overview // Eur. J. Oper. Res., 2004. Vol. 155. Iss. 1. P. 1—21. doi: http://dx.doi.org/10.1016/S0377-2217(03)00274-1

[6] *Walteros J. L., Vogiatzis C., Pasiliao E. L., Pardalos P. M.* Integer programming models for the multidimensional assignment problem with star costs // Eur. J. Oper. Res., 2014. Vol. 235. No. 3. P. 553–568. doi: http://dx.doi.org/10.1016/j.ejor.2013.10.048

[7] *Kuroki Y., Matsui T.* An approximation algorithm for multidimensional assignment problem minimizing the sum of squared errors // Discrete Appl. Math., 2009. Vol. 157. P. 2124–2135.

[8] *Sahu A., Tapadar R.* Solving the assignment problem using genetic algorithm and simulated annealing // IAENG Int. J. Appl. Math., 2007. Vol. 36. No. 1. http://www.iaeng.org/IJAM/issues_v36/issue_1/IJAM_36_1_7.pdf.

[9] *Pistorius J., Minoux M.* An improved direct labeling method for the max-flow min-cut computation in large hypergraphs and applications // Int. Trans. Oper. Res., 2003. Vol. 10. No. 1. P. 1–11.

[10] *Cooke D. J., Bez H. E.* Computer mathematics. — 1st ed. — Cambridge University Press, 1984. 408 p.

[11] *Johannes F., Eyke H.* Preference learning: An introduction. — Springer, 2011. 454 p.

[12] *Albadvi A.* Formulating national information technology strategies: A preference ranking model using PROMETHEE method // Eur. J. Oper. Res., 2004. Vol. 153. No. 2. P. 290–296. doi: http://dx.doi.org/10.1016/S0377-2217(03)00151-6

[13] *Siskos Y., Matsatsinis N. F., Baourakis G.* Multicriteria analysis in agricultural marketing: The case of French olive oil market // Eur. J. Oper. Res., 2001. Vol. 130. No. 2. P. 315–331. doi: http://dx.doi.org/10.1016/S0377-2217(00)00043-6

[14] *Mladineo N., Margeta J., Brans J. P., B. Mareschal.* Multicriteria ranking of alternative locations for small scale hydro plants // Eur. J. Oper. Res., 1987. Vol. 31. No. 2. P. 215–222. doi: http://dx.doi.org/10.1016/0377-2217(87)90025-7

[15] *Стрижов В. В.* Уточнение экспертных оценок, выставленных в ранговых шкалах, с помощью измеряемых данных // Заводская лаборатория. Диагностика материалов, 2011. Т. 77. № 7. С. 72–78.

[16] *Медведникова М. М.* Использование метода главных компонент при построении интегральных индикаторов // Машинное обучение и анализ данных, 2012. Т. 3. С. 292–304.

[17] *Медведникова М. М., Стрижов В. В., Кузнецов М. П.* Алгоритм многоклассовой монотонной Парето-классификации с выбором признаков // Известия Тульского гос. ун-та. Естественные науки, 2012. Т. 3. С. 132–141.

[18] *Kuznetsov M. P., Strijov V. V.* Methods of expert estimations concordance for integral quality estimation // Expert Syst. Appl., 2014. Vol. 41. Iss. 4. Pt. 2. P. 1551–2110. doi: http://dx.doi.org/10.1016/j.eswa.2013.08.095

[19] *Воронцов К. В.* Лекции по методу опорных векторов. http://www.ccas.ru/voron/download/SVM.pdf.

[20] http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Bakhteev014UniversityRanking/data/data.csv?format=raw.

[21] *Стрижов В. В.* Уточнение экспертных оценок с помощью измеряемых данных // Заводская лаборатория. Диагностика материалов, 2006. Т. 72. № 7. С. 59–64.

[22] *Wilson D. R., Martinez T. R.* Improved heterogeneous distance functions // J. Artif. Intell. Res., 1997. Vol. 6. P. 1–34.

[23] *Batista G. E. A. P. A., Silva D. F.* How k-nearest neighbor parameters affect its performance // Argentine Symposium on Artificial Intelligence Proceedings, 2009. P. 1–12.

[24] *Walesiak M.* Distance measure for ordinal data // Argum. Oecon., 1999. Vol. 2. No. 8. P. 167–173.

[25] *Prokhorov A. V.* (originator). Kendall coefficient of rank correlation // Encyclopedia of mathematics. http://www.encyclopediaofmath.org/index.php?title=Kendall_coefficient_of_rank_correlation&oldid=13189.

[26] *Steinhaus H.* Sur la division des corps materiels en parties // Bull. Acad. Polon. Sci., 1956. Vol. 4. P. 801–804.

[27] *Cormen T. H., Leiserson C. E., Rivest R. L., Stein C.* Introduction to algorithms. — 3rd. ed. — MIT Press, 2009. 1312 p.

[28] *Agarwal S., Branson K., Belongie S.* Higher order learning with graphs // 23rd Conference (International) on Machine Learning Proceedings, 2006. P. 1–24. doi: http://dx.doi.org/10.1145/1143844.1143847

[29] *Pu L., Faltings B.* Hypergraph learning with hyperedge expansion // European Conference on Machine Learning and Knowledge Discovery in Databases Proceedings, 2012. Vol. 1. P. 410–425. doi: http://dx.doi.org/10.1007/978-3-642-33460-3_32

[30] *Joachims T.* Optimizing search engines using clickthrough data // 8th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings, 2002. P. 133–142. doi: http://dx.doi.org/10.1145/775047.775067

[31] *Winship C., Mare R.* Regression models with ordinal variables // Am. Sociol. Rev., 1984. Vol. 49. No. 4. P. 512–525. doi: http://dx.doi.org/10.2307/2095465

[32] Algorithms of machine learning. http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Group074/Bakhteev014UniversityRanking/code/.

[33] *Batista G. E. A. P. A., Monard C. M.* An analysis of four missing data treatment methods for supervised learning // Appl. Artif. Intell., 2003. Vol. 17. Iss. 5-6. P. 519–533. doi: http://dx.doi.org/10.1080/713827181

[34] *Strickert M., Teichmann S., Sreenivasulu N., Seiffert U.* High-Throughput Multi-dimensional Scaling (HiT-MDS) for cDNA-Array expression data // 15th Conference (International) on Artificial Neural Networks: Biological Inspirations Proceedings, 2005. Vol. 1. P. 625–633. doi: http://dx.doi.org/10.1007/11550822_97

Поступила в редакцию 20.12.15

Методы трансформации моделей в задачах нелинейной регрессии*

Р. А. Сологуб
roman.sologub@yahoo.com
Вычислительный центр РАН им. А. А. Дородницына, Россия, г. Москва, ул. Вавилова, 40

Решается проблема автоматического построения и упрощения нелинейных регрессионных моделей. Модели предназначены для описания результатов измерений и прогнозирования экспериментов, составляющих неотъемлемую часть естественно-научных исследований. Порождаемые модели предназначены для аппроксимации, анализа и прогнозирования результатов измерений. При порождении учитываются требования, предъявляемые экспертами-специалистами в предметной области к порождаемым моделям. Это дает возможность получения экспертно-интерпретируемых моделей, адекватно описывающих результат измерения.

Ключевые слова: *анализ данных; регрессионная модель; нелинейная регрессия; порождение моделей; построение суперпозиций*

DOI: 10.21469/22233792.1.14.06

1 Введение

Для создания адекватной модели измеряемых данных используются экспертно-заданные порождающие функции и набор правил порождения. Модель задается в виде суперпозиции порождающих функций. Правила порождения определяют допустимость суперпозиции и исключают порождение изоморфных моделей.

В работе предлагается развить существующие методы автоматического порождения моделей [1, 2]. Исследуются методы и алгоритмы упрощения моделей и их свойства. Анализируется проблема возникновения различных топологически, но при этом равных функционально моделей. Предлагаются новые методы поиска изоморфных суперпозиций, основанные на поиске изоморфных подграфов и подстановке подграфов по правилам.

Использование нелинейной регрессии для решения прикладных задач описывается в работах Дж. Себера [3, 4]. В них описывается построение и оценка параметров нелинейных моделей. Для оценки параметров моделей используется алгоритм Левенберга–Марквадта [5]. Критерием качества при этом, как и в случае обычной линейной регрессии, остается среднеквадратичная ошибка. В работах [6, 7] индуктивное порождение моделей строится с помощью метода группового учета аргументов. В линейной модели предлагается порождать новые признаки с помощью операции произведения. С помощью полиномов Колмогорова–Габора алгоритм целенаправленно порождает и перебирает модели-претенденты различной сложности согласно ряду критериев. В результате находится модель оптимальной структуры в виде одного уравнения или системы уравнений [7]. Для индуктивного порождения моделей в работах Дж. Козы [8, 9], связанных с генетическим программированием [10, 11], осуществляется переход от строковой записи моделей к префиксной записи, таким образом вводится построение модели в виде графа-дерева.

*Работа выполнена при финансовой поддержке РФФИ, проект № 14-07-31326.

Работа [12], продолжающая работы Дж. Козы, связана с аналитическим программированием — дальнейшим алгебраическим развитием методов генетического программирования. Авторы используют строковое представление и цепочки логических предикатов в качестве элементов модели. В процессе построения моделей отсекаются циклические, а также имеющие комплексные или бесконечные значения.

Построение прогностической модели в виде суперпозиции заданных функций, предложенное в работе [13] позволяет получать интерпретируемые модели, а предложенный метод штрафования суперпозиций за сложность порождает менее точные, но более простые суперпозиции. Метод преобразования и упрощения суперпозиций по правилам, рассмотренный в работе [13], позволяет разделить построенные суперпозиции на классы эквивалентности и выбрать из каждого класса наиболее простую (т. е. имеющую наименьшее число структурных элементов) суперпозицию, что также позволяет обосновать возможность экспертной интерпретации. Методы построения комбинаций прогностических моделей описаны в работах [9, 12].

Для упрощения структуры моделей используются методы теории трансформации графов, предложенные в работе [14]. Для трансформации деревьев выделяются некоторые элементарные графы-шаблоны, для которых строятся оболочки изоморфных им графов более сложной структуры. Для упрощения модели производится рекурсивный поиск подграфов, изоморфных графам-шаблонам, с их заменой на более простые подграфы. Задача упрощения моделей, представленных в виде графов, рассматривается в работе [15]. Авторы рассматривают два различных метода упрощения моделей. В первом анализируется структура моделей и выделяются элементы-подграфы, которые подходят под шаблоны упрощения (например, двойное отрицание). Альтернативным методом является вычисление значений элемента модели на исходной выборке. Если значения функции совпадают со значениями более простого шаблона, осуществляется замена элемента модели шаблоном.

Цель работы — исследование проблемы построения и упрощения нелинейных регрессионных моделей как суперпозиций заданных параметрических функций. Предлагается метод трансформации суперпозиций, представленных в виде категории на множестве направленных ациклических графов без самопересечений, соответствующих суперпозициям.

2 Постановка задачи

Пусть задана выборка $D = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$, $\mathbf{x} \in \mathbb{R}^m$. Требуется построить функцию регрессии $\varphi(\mathbf{x}, \mathbf{w}) \mapsto \mathbf{y}$. Из множества функций F требуется выбрать модель f — отображение из декартова произведения множества свободных переменных $\mathbf{x} \in \mathbb{R}^n$ и множества параметров $\mathbf{w} \in \mathbb{R}^m$ в $\mathbb{R}^1$. Сужение модели есть функция регрессии φ с заданными значениями $\mathbf{w} = \mathbf{w}_0$. Требуется оценить набор параметров $\mathbf{w}_0$, доставляющих минимум внешнему критерию качества модели — квадратичной ошибке:

$$S(\mathbf{w}|D, f) = ||f(\mathbf{x}, \mathbf{w}) - y|| \,.$$

Выражение $S(\mathbf{w}|D, f)$ означает значение функции ошибки S, которое зависит от набора параметров $\mathbf{w}$ при заданной выборке D и модели f. Такая модель называется оптимальной при условии, что ее сложность $C(f)$ не превышает заданную. Сложность определяется как количество элементов во всех поддеревьях, которые можно выделить из дерева, представляющего модель. Искомую модель f будем искать среди множества суперпозиций функций $g \in G$. При этом накладываются ограничения на структуру суперпозиции.

Определение 1. Допустимой называется суперпозиция, удовлетворяющая следующим требованиям.

1. Элементами суперпозици f могут являться только порождающие функции g_j и свободные перменные $\mathbf{x}$.
2. Количество аргументов элемента суперпозиции равно арности соответствующей ему функции g_j.
3. Порядок аргументов элемента суперпозиции соотвествует порядку аргументов соответствующей функции g_j.
4. Для элемента s_i, аргументом которого является элемент s_j, область определения соответствующей порождающей функции g_i содержит область значений порождающей функции аргумента g_j: $\mathrm{dom}(g_i) \supseteq \mathrm{cod}(g_j)$.

Порождается множество моделей $f \in F$ — допустимых суперпозиций, состоящих из функций $g_i \in G$. Требуется выбрать модель, доставляющую минимум $S(f|\mathbf{w}^*_{\mathrm{ML}}, \mathfrak{D})$ при условии, накладываемом на сложность $C(f) < C^*$. Различные методы определения сложности модели будут рассмотрены в следующем разделе.

Следует заметить, что выборка вместе с суперпозициями составляет категорию $\mathfrak{F}$, так как для данной конструкции выполняются все аксиомы теории категорий:

1. $\mathfrak{F}$-объектами данной категории являются множества независимых перменных x и зависимых переменных y.
2. $\mathfrak{F}$-стрелками в данной категории являются суперпозиции f_i.
3. Функции $\mathrm{dom}(f)$ и $\mathrm{cod}(f)$ для суперпозиции f определяются естественным образом как область определения и область значений соответствующей суперпозиции.
4. Если для пары суперпозиций $\langle f_1, f_2\rangle$ выполняется условие $\mathrm{cod}(f_1) = \mathrm{dom}(f_2)$, то суперпозиция f_2 имеет область определения $\mathrm{dom}(f_2) \in \mathbb{R}^1$. Суперпозиция, в которой вместо независимых переменных из f_2 будет использоваться суперпозиция f_1, будет допустимой, т. е. композиция существует и входит в множество $\mathfrak{F}$-стрелок. Ассоциативность следует из того факта, что замена в суперпозиции одного аргумента на другой является ассоциативной операцией. Вообще все множество $\mathfrak{F}$-стрелок состоит из элементов G и их композиций.
5. Наличие единицы обеспечивается обязательным существованием в G функции $\mathrm{id}(\mathbf{x})$. Для этой функции выполняется закон тождества по определению.

2.1 Описание структуры модели

Условимся считать, что каждой суперпозиции f сопоставлено дерево Γ_f, эквивалентное этой суперпозиции и строящееся следующим образом:

– в вершинах v_i дерева Γ_f находятся соответствующие порождающие функции g_j;
– число дочерних вершин у некоторой вершины v_i равно арности соответствующей ей функции g_j;
– порядок дочерних вершин вершины v_i соотвествует порядку аргументов соответствующей функции g_j;
– листьями дерева Γ_f являются свободные переменные x_i либо числовые параметры w_i.

Таким образом, вычисление значения выражения f в некоторой точке с данным вектором параметров $\mathbf{w} = \{w_1, w_2, \ldots, w_k\}$ эквивалентно подстановке соответствующих значений свободных переменных x_i и параметров w_i в дерево Γ_f, где x_i — элементы вектора свободных переменных $\mathbf{x}$.

Заметим важное свойство таких деревьев: каждое поддерево Γ'_f дерева Γ_f, корнем которого является вершина v_i, также соответствует некоторой суперпозиции, являющейся составляющей исходной суперпозиции f.

Предложим определение сложности суперпозиции, позволяющее штрафовать суперпозиции с большим числом вложенных функций. Введем понятие сложности вершины.

Определение 2. Сложность C суперпозиции f равна сложности дерева Г, соответствующего ей, и определяется как сумма количества элементов во всех поддеревьях дерева Г.

Таким образом штрафуется суперпозиция, содержащая большое число вложенных функций. Определение позволяет вычислять сложность, производя обход дерева снизу вверх обратно обходу дерева «в глубину» — сложность родительской вершины равна удвоенной сложности вершин потомков плюс единица. Сложность корня и будет сложностью всей суперпозиции, $C(1,1) = C(f)$.

3 Трансформация моделей

При порождении моделей в общем случае одному и тому же отображению соответствуют суперпозиции различной сложности, например одно и то же отображение соответствует моделям x и $\sqrt[3]{x^3}$. Также возможны случаи порождения деревьев, некоторые ветви которых не оказывают влияния на значение функции (например, умножаются на 0). Данная проблема оказывается важной для многих классов задач, например для построения логических функций или для задачи угадывания функции [16]. Для понимания, как упрощать подобные суперпозиции, следует ввести понятие эквивалентности моделей.

Определение 3. Модель f_2 с вектором параметров $\mathbf{w}_2$ называется обобщающей для модели f_1 с вектором параметров $\mathbf{w}_1$, если для любого вектора $\mathbf{w}_1$ найдется такой вектор $\mathbf{w}_2$, что для любого $\mathbf{x} \in D$ значения функций $f_1(\mathbf{w}_1, \mathbf{x})$ и $f_2(\mathbf{w}_2, \mathbf{x})$ равны:

$$\mathbf{x} \in D \Rightarrow f_1(\mathbf{w}_1, \mathbf{x}) = f_2(\mathbf{w}_2, \mathbf{x})\,.$$

Определение 4. Модели f_1 и f_2 с векторами параметров $\mathbf{w}_1$ и $\mathbf{w}_2$ называются эквивалентными, если каждая из них является обобщающей для другой.

Для построения оптимальной модели f ограниченной сложности $C(f) < C_0$ необходимо найти способ трансформации модели f большей структурной сложности в модель меньшей сложности f' с помощью специального алгоритма упрощения. Алгоритм упрощения модели $f(\mathbf{w}, \mathbf{x})$ минимизирует сложность суперпозиции, соответствующей ее дереву, при условии, что результирующая модель $f'(\mathbf{w}', \mathbf{x})$ является обобщающей моделью для исходной модели $f(\mathbf{w}, \mathbf{x})$. При проведении данной операции какие-либо вершины и ребра из дерева, соответствующего трансформируемой модели f, будут удалены и будут построены другие вершины и ребра вместо них. Обобщим алгоритм упрощения на орграфы любого вида, а не только на деревья. Далее для каждого графа подразумевается, что это орграф.

Определение 5. Подграф L, удаляемый из графа G в алгоритме упрощения, будет называться заменяемым подграфом.

Определение 6. Создаваемый подграф R, помещаемый в граф G в алгоритме упрощения, называется замещающим подграфом.

Существуют по меньшей мере два широко используемых метода упрощения моделей: «алгебраическое упрощение», являющееся частным случаем алгебраической трансформации графов, и «упрощение эквивалентным решением» [15].

3.1 Алгебраический подход к трансформации графов

Определение трансформации графа как замены одного подграфа на другой является интуитивно понятным, однако нестрогим. Для использования математического аппарата теории категория следует строго определить трансформацию графа.

Определение 7. Трансформацией $\mathfrak{f}$ на множестве графов Γ является пара гиперсхем H_1 и H_2, функция поиска $\mathfrak{m}$, ставящая в соответствие гиперсхеме H_1 подграф Γ, соответствующий этой гиперсхеме, и взаимно однозначное отображение f, ставящее в соответствие корню и листьям H_1 корень и листья H_2. При этом порождающие функции, соответствующие этим вершинам, должны совпадать.

Каждой трансформации, таким образом, может быть поставлена в соответствие обратная трансформация $\mathfrak{f}^{-1}$.

В рамках алгебраического подхода к трансформации графов следует ввести категорию трансформаций графов $\mathfrak{G}$, объектами которой являются графы Γ, а стрелками — трансформации графов $\mathfrak{f}$. Рассмотрим аксиомы категории.

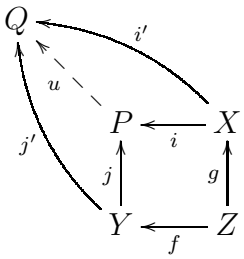
1. $\mathfrak{G}$-объектами в данной категории являются множества графов Γ.
2. $\mathfrak{G}$-стрелками в данной категории являются трансформации графов $\mathfrak{f}_i$.
3. Функции $\mathrm{dom}(\mathfrak{f})$ и $\mathrm{cod}(\mathfrak{f})$ для трансформаций графов определяются с помощью функции поиска $\mathfrak{m}$. $\mathrm{cod}(\mathfrak{f})$ может быть найден как $\mathrm{dom}(\mathfrak{f}^{-1})$.
4. Ассоциативность следует из наличия обратной функции.
5. Единицей является тривиальная трансформация с гиперсхемами $H_1 = H_2 = \#$.

Алгебраический подход к трансформации графов основывается на конструкции кодекартова квадрата морфизмов.

Определение 8. Кодекартов квадрат морфизмов $f : Z \to Y$ и $g : Z \to X$ — это объект P и два морфизма $i : X \to P$ и $j : Y \to P$, для которых следующая диаграмма коммутативна:

$$\begin{array}{ccc} P & \xleftarrow[i]{} & X \\ {\scriptstyle j}\uparrow & & \uparrow{\scriptstyle g} \\ Y & \xleftarrow[f]{} & Z \end{array} \qquad (1)$$

Кодекартов квадрат (P, i, j) является универсальным среди объектов, для которых диаграмма (1) коммутативна, т. е. для любой (Q, i', j'), такого что предыдущая диаграмма коммутирует, существует единственный морфизм $u : P \to Q$, делающий следующую диаграмму коммутативной:



(2)

Как и любой универсальный оператор, кодекартов квадрат определен с точностью до изоморфизма.

Определение 9. Кодекартов квадрат морфизмов $f : Z \to X$ и $g : Z \to Y$ — это копредел диаграммы $X \leftarrow Z \to Y$.

В контексте категории графов, используемой в данной работе, кодекартов квадрат является дизъюнктивной суммой множеств графов X и Y, при этом элементы с общим прообразом в множестве Z склеиваются, т. е. для каждого графа — элемента множества Z — образы его вершин и ребер относительно преобразований $i \cdot g$ и $j \cdot f$ будут совпадать. В рамках данной работы вместо термина «кодекартов квадрат» также будет использоваться

термин-синоним «склейка». Трансформация графа может строиться сразу как два кодекартовых квадрата. Данный подход называется двойной склейкой в противоположность к однократной склейке. Оба подхода описаны ниже. В процессе трансформации графов каждый граф Γ_1 — элемент множества X — является заменяемым и заменяющим подграфом, каждый граф Γ_2 — элемент множества Y — неизменной частью этого графа, а элементы Z — общей частью заменяемого и заменяющего подграфов. Естественным образом вводится операция соединения графов Γ_1 и Γ_2, результатом которой является объединение множеств, при этом соответствующие вершины и ребра накладываются друг на друга.

3.2 Трансформация двойной склейкой

Для рассмотрения трансформации графов необходимо ввести понятие правила, построенного в виде кодекартова квадрата морфизмов. Множества графов Λ и Φ являются в схеме кодекартова квадрата множеством X, множество граф Ψ — множеством Z, а множеству Y соответствует Δ. Множеству P для двух квадратов соответствуют начальный и конечный графы Γ и Ω.

Определение 10. Правило — это тройка $p = (\Lambda, \Psi, \Phi)$, где Λ и Φ являются заменяемым и замещающим подграфами и граф Ψ является общей частью подграфов Λ и Φ, т. е. их пересечением. Заменяемый, или начальный, подграф Λ называется условием применения правила; замещающий, или конечный, подграф Φ — итогом его применения. Подграф Ψ описывает часть графа, необходимую для применения правила, но неизменную в процессе применения. Множество $\Lambda \setminus \Psi$ является удаляемой частью графа, вместо нее создается множество $\Phi \setminus \Psi$.

Определение 11. Процедура поиска $\mathfrak{m}$ — отображение из Λ в Γ, ставящая в соответствие заменяемому графу эквивалентный ему подграф. При этом процедура $\mathfrak{m}$ сохраняет структуру графа Γ.

Определение 12. Трансформация графа — это пара, элементами которой являются правило p и процедура поиска $\mathfrak{m}$. Процедура трансформации графа Γ в граф Ω с помощью правила p и процедуры поиска $\mathfrak{m}$ будет также обозначаться как $\Gamma \overset{p,\mathfrak{m}}{\to} \Omega$.

Процедура трансформации графа правилом p и процедурой поиска $\mathfrak{m}$ состоит из двух шагов. На первом шаге все ребра и вершины, соответствующие множеству $\Lambda \setminus \Psi$, удаляются из графа Γ. Удаляемая часть может не являться графом, но оставшаяся структура $\Delta = \{\Gamma \setminus \mathfrak{m}(\Lambda)\} \bigcup \mathfrak{m}(\Psi)$ должна оставаться графом, т. е. в ней не должно быть подвешенных ребер. Таким образом, процедура поиска $\mathfrak{m}$ должна удовлетворять условию соединения графов, т. е. результатом соединения $\Lambda \setminus \Psi$ и Δ является граф Γ (см. диаграмму (3)). На втором шаге трансформации граф Δ соединяется с графом $\Phi \setminus \Psi$ для образования производного графа Ω (см. диаграмму (3)). Так как подграфы Λ и Φ могут иметь пересечение Ψ, подграф Ψ существует и в начальном графе Γ и не удаляется на первом шаге, т. е. существует и в промежуточном графе Δ. Для присоединения новых ребер и вершин к графу Δ используется граф Ψ. Таким образом определяются присоединенные вершины, с помощью которых граф Φ присоединяется к графу Δ. Для получения графа оптимальной структуры процедура одиночной трансформации графа может быть выполнена несколько раз.

Формально трансформация графа задается следующим образом. Пусть даны правило

$$p = (\Lambda \leftarrow \Psi \rightarrow \Phi)$$

и промежуточный граф Δ, который включает в себя Ψ, тогда исходный граф Γ трансформации $\Gamma \to \Omega$ с помощью правила p — это соединение Λ и Δ с помощью Ψ:

$$\Gamma = \Lambda +_{\Psi} \Delta\,,$$

а результирующий граф Ω определяется как соединение Φ и Δ с помощью Ψ:

$$\Omega = \Phi +_{\Psi} \Delta\,.$$

Более точно используются морфизмы

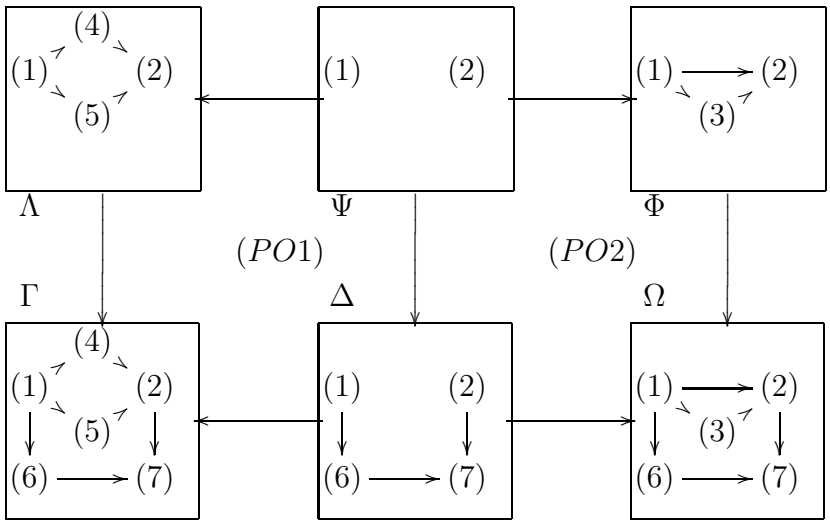
$$r : \Psi \to \Lambda\,; \qquad l : \Psi \to \Phi\,; \qquad k : \Psi \to \Delta$$

для того, чтобы показать, каким образом Ψ входит в Λ, Φ и Δ соответственно. Данный способ построения начального графа Γ и конечного графа Ω позволяет определить конструкции соединения $\Gamma = \Lambda +_{\Psi} \Delta$ и $\Omega = \Phi +_{\Psi} \Delta$ как конструкции склейки (см. диаграмму (3)). Таким образом, диаграмма (3) является двойным кодекартовым квадратом. Результирующий морфизм $\mathfrak{n} : \Phi \to \Omega$ называется ко-поиском трансформации $\Gamma \to \Omega$. Данная функция является функцией поиска в графе Ω подграфа, изоморфного заменяющему подграфу Φ. Коммутативная диаграмма для трансфомации графа строится следующим образом:

$$\begin{array}{ccccc} \Lambda & \xleftarrow{\;r\;} & \Psi & \xrightarrow{\;l\;} & \Phi \\ \downarrow{\scriptstyle \mathfrak{m}} & (1) & \downarrow{\scriptstyle k} & (2) & \downarrow{\scriptstyle \mathfrak{n}} \\ \Gamma & \xleftarrow{\;g\;} & \Delta & \xrightarrow{\;h\;} & \Omega \end{array} \tag{3}$$

Для применения правила p с процедурой поиска $\mathfrak{m}$ подграфа Λ в графе Γ, при заданном морфизме $\mathfrak{m} : \Lambda \to \Gamma$, как показано на коммутативной диаграмме (3), в первую очередь необходимо построить промежуточный граф Δ такой, что соединение $\Lambda +_{\Psi} \Delta$ даст результатом граф Γ. На следующем шаге строим соединение $\Phi +_{\Psi} \Delta$ графов Φ и Δ с помощью графа Ψ, получая граф Ω, и, таким образом, получаем процедуру двойной склейки $\Gamma \to \Omega$ с помощью правила p и процедуры поиска $\mathfrak{m}$. Для первого шага необходимо выполнение условия соединения графов, что позволяет нам построить Δ из условия $\Gamma = \Lambda +_{\Psi} \Delta$. Для процедуры $\mathfrak{m}$ условие соединения означает, что все подвешенные вершины Λ, т. е. вершины $v \in \Lambda$, такие, что $\mathfrak{m}(v)$ является начальной или конечной вершиной некоторого ребра e, принадлежащего $\Gamma \setminus \Lambda$, должны быть в Ψ.

Рассмотрим пример двойной склейки:



(4)

Данная диаграмма соответствует общей схеме (3). Следует заметить, что в диаграмме (3) граф Γ является соединением графов Λ и Δ с помощью Ψ, причем обозначения вершин показывают, как вершины размечаются при применении морфизмов.

Рассмотрим условие корректности построения структуры графа Δ. Разметка ребер может быть единственным образом выведена из разметки вершин. Условие соединения вершин выполнено на диаграмме (4), потому что подвешенные вершины (1) и (2), принадлежащие Λ, также являются соединительными вершинами. Таким образом, не остается подвешенных ребер, выходящих из вершин (1) и (2). При этом граф Ω является соединением графов Φ и Δ вместе с Ψ, что приводит к трансформации $\Gamma \to \Omega$ с помощью правила p. Фактически диаграммы (3) и (4) являются кодекартовыми квадратами в категории графов, состоящей из графов и морфизмов на них.

Сформулируем точное условие соединения графов при трансформации графа. Для этого вводим следующие определения.

Определение 13. Точки соединения — вершины и ребра в Λ, которые не удаляются при применении правила p.

Определение 14. Точки обнаружения — вершины и ребра в Λ, образы которых относительно $\mathfrak{m}$ имеют более одного прообраза.

Определение 15. Подвешенные вершины — вершины в Λ, образы которых относительно $\mathfrak{m}$ в Γ имеют входящие или выходящие ребра, не содержащиеся в Λ.

В данных определениях условие соединения графа выглядит следующим образом.

Теорема 1. **[14]** Пусть даны правило $p = (\Lambda \leftarrow \Psi \rightarrow \Phi)$, граф Γ и процедура поиска $\mathfrak{m} : \Lambda \to \Gamma$. Вершины графов обозначаются буквой V, ребра — E. Тогда правило p с процедурой поиска $\mathfrak{m}$ удовлетворяет условию соединения, если все точки обнаружения и подвешенные вершины также являются точками соединения.

Докажем данную теорему от противного. Пусть существует подвешенная вершина v_0, не являющаяся точкой соединения. Данная вершина удаляется из Γ при применении правила p. Однако в Γ существуют ребра, не содержащиеся в Λ и присоединенные к v_0. Таким образом, полученный граф будет недопустимым, потому что у некоторых ребер не будет

начала или конца. Точки обнаружения являются точками соединения, так как иначе правило будет внутренне противоречивым.□

Ограничения, накладываемые естестенным образом на трансформации двойной склейкой, не позволяют удобно производить многие операции с графами, используемые на практике. Так, операция замены вершины поддерева v_i не может быть описана в виде заменяемого и замещающего графов, состоящих из одной вершины, так как если в заменяемом графе всего одна вершина v_i, эта вершина не будет являться подвешенной, только если весь граф состоит из одной вершины. Таким образом, для применения трансформаций предлагается метод, сопоставляющий неудовлетворяющей условиям трансформации набор допустимых трансформаций.

Теорема 2. Любой трансформации $t = (\Lambda_t, \Psi_t, \Phi_t)$ графа соответствует набор правил $p_t = (\Lambda_{p_t}, \Psi_{p_t}, \Phi_{p_t})$, удовлетворяющих условию соединения, такой, что любое применение трансформации t аналогично применению одного из правил p_t.

Данная теорема доказывается конструктивно — рассматриваются все возможные наборы количеств ребер, которые могут иметь точки соединения v_c, и для каждого набора создаются заменяемый и замещающий подграфы, в который добавляются вершины типа # на концах всех ребер, выходящих из v_c и не содержавшихся ранее в заменяемом подграфе Λ.□

Морфизмы $\Psi \to \Lambda$ и $\Psi \to \Phi$ в произведениях могут быть ограничены как инъективные морфизмы — каждому образу в Λ и Φ соответствует только один проообраз из Ψ. Тем не менее возможны неинъективные варианты процедур поиска $\mathfrak{m} : \Lambda \to \Gamma$ и ко-поиска $\mathfrak{n} : \Phi \to$ $\to \Omega$. Это может быть особенно важным, когда рассматривается параллельное применение правил:

$$p_1 \bigoplus p_2 : \Lambda_1 \bigoplus \Lambda_2 \leftarrow \Psi_1 \bigoplus \Psi_2 \to \Phi_1 \bigoplus \Phi_2 \,,$$

где $\bigoplus$ означает дизъюнктивное объединение. Даже для инъективных вариантов $\mathfrak{m}_1 : \Lambda_1 \to$ $\to \Gamma$ с помощью p_1 и $\mathfrak{m}_2 : \Lambda_2 \to \Gamma$ с помощью p_2, итоговая операция $\mathfrak{m} : \Lambda_1 + \Lambda_2 \to \Gamma$ не является инъективной, если образы процедур поиска $\mathfrak{m}_1(\Lambda_1)$ и $\mathfrak{m}_2(\Lambda_2)$ имеют непустое пересечение в Γ.

Теорема 3. Существует набор трансформаций $(p_1, \mathfrak{m}_1)$ и $(p_2, \mathfrak{m}_2)$, такой, что их параллельное применение имеет неинъективную функцию поиска $\mathfrak{m} = \mathfrak{m}_1 \bigoplus \mathfrak{m}_2$.

Построим пример таких трансформаций. Пусть трансформация преобразует дерево Γ_0, соответствующее функции $f_0 = (x+1) * (x - 1 + x^2 - x + 1)$, и есть два правила

$$p_1 = \{(x+1)(x-1), x^2 - 1\}\,; \quad p_2 = \{(x+1)(x^2 - x + 1), x^3 + 1\}\,.$$

В обоих этих правилах подграф Ψ является пустым. Обе процедуры поиска $\mathfrak{m}_1$ и $\mathfrak{m}_2$ будут находить часть суперпозиции $(x+1)$. Из этого следует, что при объединении Λ_1 и Λ_2 у одного образа из Γ будет более одного прообраза, т. е. объединенное правило p_{12} дважды найдет в графе Γ_0 подграф, соотвествующий суперпозиции $(x+1)\cdot$.□

Для рассмотрения случаев применения нескольких трансформаций необходимо определить условие, при котором трансформации могут применяться последовательно и параллельно. Введем понятия параллельно и последовательно независимых трансформаций.

Определение 16. Две трансформации графов $\Gamma \overset{p_1,m_1}{\to} \Omega_1$ и $\Gamma \overset{p_2,m_2}{\to} \Omega_2$ являются параллельно независимыми, если все вершины и ребра, попадающие в образ обоих морфизмов поиска, являются соединительными:

$$\mathfrak{m}_1(\Lambda_1) \bigcap \mathfrak{m}_2(\Lambda_2) \subseteq \mathfrak{m}_1(l_1(\Psi_1)) \bigcap \mathfrak{m}_2(l_1(\Psi_2)).$$

Две трансформации графов $\Gamma \overset{p_1,m_1}{\to} \Omega_1$ и $\Omega_1 \overset{p_2,m_2}{\to} \Omega_2$ являются последовательно независимыми, если все вершины и ребра, попадающие в пересечение морфизмов $\mathfrak{n}_1$ и m_2, являются соединительными:

$$\mathfrak{n}_1(\Phi_1) \bigcap \mathfrak{m}_2(\Lambda_2) \subseteq \mathfrak{n}_1(r_1(\Psi_1)) \bigcap \mathfrak{m}_2(l_2(\Psi_2)).$$

Следует заметить, что для графов-деревьев возникает простой достаточный критерий пареллельной и последовательной независимости трансформаций, если их замещаемые графы Λ являются односвязными.

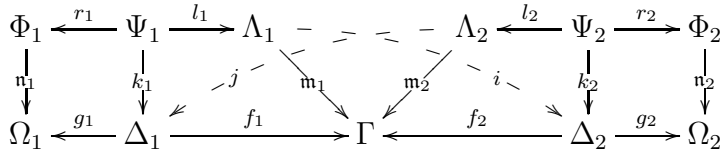
Теорема 4. Две трансформации графов-деревьев $\Gamma \overset{p_1,m_1}{\to} \Omega_1$ и $\Gamma \overset{p_2,m_2}{\to} \Omega_2$ являются параллельно и последовательно независимыми, если образы корней v_1 и v_2 деревьев $\mathfrak{m}_1(\Lambda_1)$ и $\mathfrak{m}_2(\Lambda_1)$ не принадлежат друг другу:

$$v_1 \notin \mathfrak{m}_2(\Lambda_2)\,; \quad v_2 \notin \mathfrak{m}_1(\Lambda_1)\,. \tag{5}$$

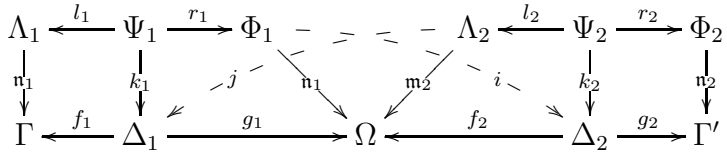
Данная теорема простым образом доказывается от противного. Пусть условие (5) выполняется и существует вершина v_0, принадлежащая пересечению множеств $\mathfrak{m}_1(\Lambda_1)$ и $\mathfrak{m}_2(\Lambda_1)$. Тогда в графе будет цикл, проходящий через вершины v_1, v_0, v_2 и корень дерева. Но в дереве не может быть циклов.□

Определение независимости оказывается неудобным для применения, так как оно слабо формализовано. Определим необходимое и достаточное условие, при котором графы являются параллельно или последовательно независимыми через существование соответствующих морфизмов.

Теорема 5. **[14]** Две трансформации графов $\Gamma \overset{p_1,m_1}{\to} \Omega_1$ и $\Gamma \overset{p_2,m_2}{\to} \Omega_2$ являются параллельно независимыми, если существуют морфизмы $i : \Lambda_1 \to \Delta_2$ и $j : \Lambda_2 \to \Delta_1$, такие, что $f_2 \circ i = \mathfrak{m}_1$ и $f_1 \circ j = \mathfrak{m}_2$:



Теорема 6. Две трансформации графов $\Gamma \overset{p_1,m_1}{\to} \Omega$ и $\Omega \overset{p_2,m_2}{\to} \Gamma'$ являются последовательно независимыми, если существуют морфизмы $i : \Phi_1 \to \Delta_2$ и $j : \Lambda_2 \to \Delta_1$, такие, что $f_2 \circ i = \mathfrak{n}_1$ и $g_1 \circ j = \mathfrak{m}_2$:



Доказательство. Рассмотрим необходимость и достаточность критерия для параллельной независимости. Для последовательной независимости доказательство будет строиться аналогичным образом. Вершина $v \in \Lambda_1$ или принадлежит множеству $\mathfrak{m}_2(\Lambda_2)$, или лежит вне его. Рассмотрим оба случая.

1. Множество $\mathfrak{m}_1(v) \notin \mathfrak{m}_2(\Lambda_2)$. Все вершины графа Γ являются образами при применении отображений $\mathfrak{m}_2$ или f_2. Отсюда $\mathfrak{m}_1(v) \in f_2(\Delta_2)$.
2. Множество $\mathfrak{m}_1(v) \in \mathfrak{m}_2(\Lambda_2)$. Тогда $\mathfrak{m}_1(v) \in \mathfrak{m}_1(\Lambda_1) \cap \mathfrak{m}_2(\Lambda_2) \subseteq \mathfrak{m}_1(l_1(\Psi_1)) \cap \mathfrak{m}_2(l_2(\Psi_2))$. При этом из коммутативной диаграммы следует, что $\mathfrak{m}_2(l_2(\Psi_2)) = f_2(k_2(\Psi_2))$. Отсюда $\mathfrak{m}_1(v) \in f_2(\Delta_2)$.

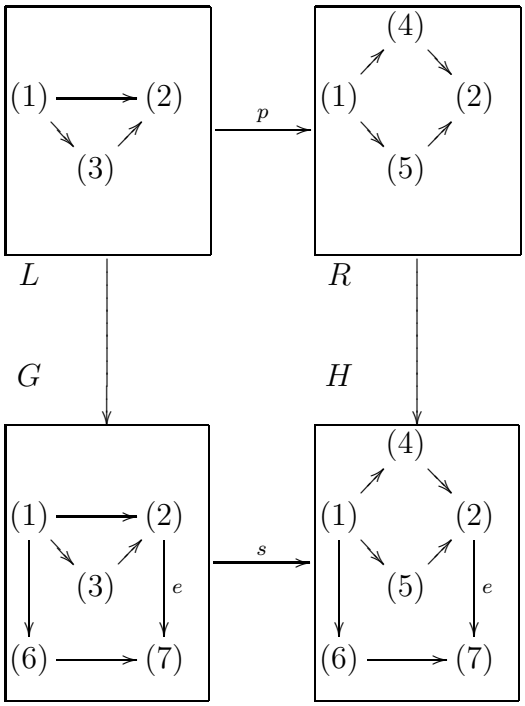
В обоих случаях оказывается, что $\mathfrak{m}_1(x) \in f_2(\Delta_2)$, так что инъективность f_2 позволяет нам определить $i(x) = f_2^{-1} \circ \mathfrak{m}_1(x)$. Аналогично, j определяется из условия $f_1 \circ j = \mathfrak{m}_2$.

При данных i, j с $f_2 \circ i = m_1$ и $f_1 \circ j = m_2$ пусть $y \in \mathfrak{m}_1(\Lambda_1) \bigcap \mathfrak{m}_2(\Lambda_2)$. Тогда $y \in \in \mathfrak{m}_1(L_1) \bigcap f_1(j(\Lambda_2))$. Из условия кодекартова квадрата следует, что существует $z_1 \in \Psi_1$, такое, что $y = \mathfrak{m}_1(l_1(z_1)) = f_1(k_1(z_1))$. Значит, $y \in \mathfrak{m}_1(l_1(\Psi_1))$, аналогично $y \in \mathfrak{m}_2(l_2(\Psi_2))$, откуда следует условие независимости $\mathfrak{m}_1(\Lambda_1) \bigcap \mathfrak{m}_2(\Lambda_2) \subseteq \mathfrak{m}_1(l_1(\Psi_1)) \bigcap \mathfrak{m}_2(l_1(\Psi_2))$.□

С использованием данных критериев можно определить, как связаны друг с другом независимые параллельно и последовательно трансформации. Данная теорема является частным случаем теоремы Черча–Россера.

3.3 Трансформация одиночной склейкой

Как было отмечено, конструкции соединения в алгебраическом подходе являются кодекартовыми квадратами в смысле морфизмов категории графов. С другой стороны, правило $p = (\Lambda \leftarrow \Psi \rightarrow \Phi)$ может быть также рассмотрено как частичный морфизм графов $p : \Lambda \rightarrow \Phi$, доменом которого является множество $\mathrm{dom}(p) = \Psi$. Более того, диаграмма $(\Gamma \leftarrow \Delta \rightarrow \Omega)$ может быть рассмотрена как частичный морфизм графов $s : \Gamma \rightarrow \Omega$ с доменом $\mathrm{dom}(s) = \Delta$. Таким образом, получается следующая диаграмма:



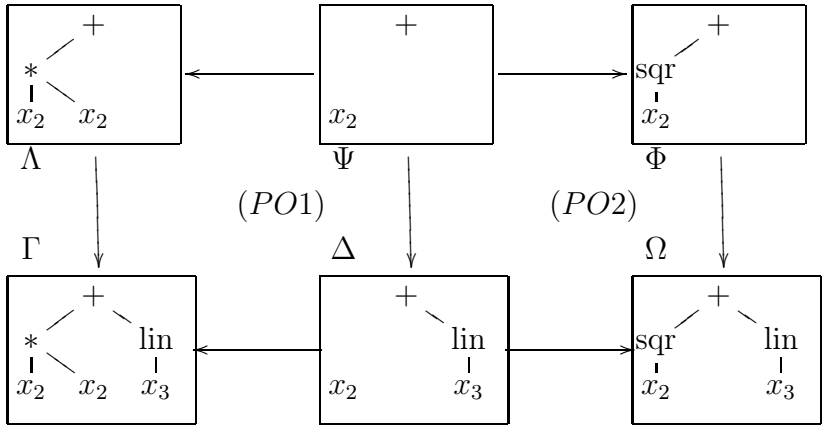
(6)

На данной диаграмме горизонтальные морфизмы являются частичными, а вертикальные — полными морфизмами графов. По сути, диаграмма (6) является кодекартовым квадратом в расширенной категории графов, которая состоит из графов и частичных морфизмах на графах и показывает, что трансформации графов могут быть выражены как одиночные кодекартовы квадраты в расширенной категории графов. Данный подход развивался Раулем [17] и был полностью разработан Лёве [18], итогом их работы является подход однократного вытеснения.

С точки зрения прикладного использования подход с одиночным вытеснением отличается от подхода с двойным вытеснением в одном главном отношении, которое касается удаления вспомогательных элементов графа в процессе трансформации графа. Процедура поиска $m : \Lambda \rightarrow \Gamma$ не удовлетворяет условию соединения по отношению к правилу $p = \\ = (\Lambda \leftarrow \Psi \rightarrow \Phi)$, поэтому данное правило не применимо в подходе с двойным вытеснением.

Но оно может быть применимо в подходе с однократным вытеснением, которое позволяет появляться подвешенным ребрам после удаления подграфа $\Lambda \setminus \Psi$ из Γ. Следует заметить, что подвешенные ребра из Γ также удаляются для создания допустимого графа Ω.

Если на диаграмме (4) вершина (2) была бы удалена из Ψ, то конструкция соединения не удовлетворяла бы подходу с двойным вытеснением. В подходе с однократным вытеснением это значило бы, что вершина (2) не находится в домене p, что ведет к подвешенному ребру в Γ после удаления $\Lambda \setminus \mathrm{dom}(p)$ на диаграмме:



В результате ребро e удялется из Ω.

Более подобное описание и сравнение данных подходов разобрано в [19].

3.4 Прикладная задача упрощения суперпозиций

При последовательном порождении моделей зачастую оказывается так, что некоторые части модели становятся рудиментарными. Упрощение Соула [20] является вариантом алгебраического упрощения, в котором объектами упрощения являются элементы моделей, параметры которых не влияют на значение функции. Область применения подобных методов ограничена [20], однако они показывают хороший результат на некоторых задачах, например при обнаружении функции. В данном типе задачи восстановления регрессии дисперсия случайной ошибки равна нулю и выборка генерируется в соответствии с какой-либо эталонной функцией f_0, которая должна быть обнаружена алгоритмом.

Упрощение эквивалентным решением заключается в сравнении значений моделей, а не структур. Эквивалентность моделей проверяется не по структуре деревьев, соответствующих им, а численно. В таком случае два выражения, дающие равные значения на области определения независимых переменных модели, считаются равными.

Определение 17. Шаблон θ — гиперсхема, обладающая наименьшей сложностью среди всех гиперсхем, таких, что при их взаимном замещении получаемые модели оказываются эквивалентными. Сложность гиперсхемы определяется как сложность суперпозиции при замещении всех символов $\{=\}$ и $\{\#\}$, означающих соответственно произвольную независимую переменную и произвольное поддерево, на константы.

Экспертно выбирается некоторый набор шаблонов Θ. Процедура упрощения состоит из двух шагов.

1. Все поддеревья Γ_j в выбранном дереве Γ проверяются на эквивалентность шаблонам из Θ согласно заданным правилам.
2. Если какое-либо поддерево Γ_j в дереве эквивалентно дереву из Θ, данное поддерево заменяется соответствующим элементом из Θ.

Процедура повторяется до тех пор, пока после вышеперечисленных итераций дерево Г не останется неизменным. При наличии в множестве порождающих функций коммутативных функций вводится алфавитное упорядочение для ветвей, выходящих из вершины γ_i дерева Г, соответствующей коммутативной порождающей функции g_i.

Эквивалентное упрощение является альтернативой алгебраическому упрощению, позволяя упрощать некоторые модели за меньшее количество операций.

Рассмотрим сложность алгоритма, упрощающего поддерево высоты l с вершинами арности не более m. Количество вершин в таком дереве ограничивается сверху как m^l. Рассмотрим дерево с максимальным количеством вершин — для такого дерева все вершины, кроме листьев, будут иметь арность m. Для сравнения всех возможных поддеревьев с шаблонами из Θ необходимо рассмотреть поддеревья любой высоты с корнем в каждой из вершин дерева. Подсчитаем количество поддеревьев всех возможных высот в таком дереве. Обозначим высоту данного дерева $l = \log_m k + 1$. Тогда для вершины, находящейся на расстоянии x от корня, количество поддеревьев с корнем в этой вершине составляет не менее чем $l - x$. Тогда искомое количество поддеревьев:

$$\sum_{x=0}^{l-1}(l-x)m^x = \frac{m(m^l-1)-lm+l}{(m-1)^2}\,.$$

Данное выражение пропорционально m^l, т. е. количеству элементов в дереве, все вершины которого (кроме листьев) имеют максимальное число потомков. Сложность алгоритма, упрощающего дерево, состоящее из k вершин, оказывается порядка не менее чем k. В случае если алгоритм проверки правил эквивалентности имеет значительную сложность, подсчет значений оценок зависимых переменных $\hat{y}$ на множестве независимых переменных $x \in D$ и сравнение этих значений с получаемыми при использовании шаблонов Θ имеет значительно меньшую сложность. Данный метод может применяться в случае, если независимые и зависимые переменные принимают ограниченное число значений. Для такого поддерева, вне зависимости от количества элементов k в нем, область определения соответствующей функции содержит 2^t точек, где t — количество независимых переменных, являющихся листьями данного поддерева. При небольших t число 2^t не превосходит k, и в таком случае алгоритм сравнения по значениям оказывается менее сложным, чем алгоритм сравнения структур поддеревьев с шаблонами.

Важным частным случаем использования алгоритма упрощения по значениям является случай равенства функций на области определения независимых переменных при необязательном равенстве вне этой области. Для решения прикладной задачи функции, дающие равные значения на области определения, будут равны, и подобная замена будет корректна.

4 Заключение

В работе предложены методы направленного порождения, модификации и упрощения нелинейных регрессионных моделей. Описаны условия существования решений, получаемых в результате порождения, доказаны необходимые теоремы. Разработан метод последовательного направленного порождения суперпозиций, введено понятие изоморфных суперпозиций, исследованы свойства порождаемых суперпозиций. Предлагаемые в работе методы упрощения моделей предназначены непосредственно для применения на практике. Создана базовая библиотека правил порождения экспертно-интерпретируемых моделей.

Литература

[1] *Стрижов В. В., Сологуб Р. А.* Индуктивное порождение поверхности волатильности опционных торгов // Вычислительные технологии, 2009. № 5. С. 102–113.

[2] *Сологуб Р. А.* Алгоритмы порождения нелинейных регрессионных моделей // Информационные технологии, 2013. № 5. С. 8–12.

[3] *Seber G., Wild C. J.* Nonlinear regression. — Wiley ser. in probability and statistics. — John Wiley & Sons, 2005. 2907 p.

[4] *Seber G.* The collected works of George A. F. Seber. — Wiley ser. in probability and statistics. — Wiley, 2009. 792 p.

[5] *Levenberg K.* A method for the solution of certain non-linear problems in least squares // Quart. J. Appl. Math., 1944. Vol. II. No. 2. P. 164–168.

[6] *Ивахненко А. Г.* Индуктивный метод самоорганизации моделей сложных систем. — Киев: Наукова думка, 1982. 296 с.

[7] *Madala H. R., Ivakhnenko A. G.* Inductive learning algorithms for complex systems modeling. — CRC Press, 1994. 384 p.

[8] *Koza J. R.* Genetic programming: On the programming of computers by means of natural selection. — Complex adaptive systems ser. — 1st ed. — Cambridge, MA – London: The MIT Press, 1992. 840 p.

[9] *Koza J. R., Keane M. A., Streeter M. J., et al.* Genetic programming IV: Routine human-competitive machine intelligence. — Norwell, MA, USA: Kluwer Academic Publs., 2003. 590 p.

[10] *Banzhaf W., Francone F. D., Keller R. E., Nordin P.* Genetic programming — an introduction: On the automatic evolution of computer programs and its applications. — San Francisco, CA, USA: Morgan Kaufmann Publs. Inc., 1998. 490 p.

[11] *Michell M.* An introduction to genetic algorithms. — Cambridge, MA – London: The MIT Press, 1998. 164 p.

[12] *Kominkova Oplatkova Z., Senkerik R., Zelinka I., Pluhacek M.* Analytic programming in the task of evolutionary synthesis of a controller for high order oscillations stabilization of discrete chaotic systems // Comput. Math. Appl., 2013. Vol. 66. No. 2. P. 177–189.

[13] *Рудой Г. И., Стрижов В. В.* Алгоритмы индуктивного порождения суперпозиций для аппроксимации измеряемых данных // Информатика и её применения, 2013. Т. 7. Вып. 1. С. 17–26.

[14] *Ehrig H., Ehrig K., Prange U., Taentzer G.* Fundamentals of algebraic graph transformation. — Berlin: Springer, 2006. 390 p.

[15] *Naoki M., McKay B., Xuan N., et al.* A new method for simplifying algebraic expressions in genetic programming called equivalent decision simplification // Distributed computing, artificial intelligence, bioinformatics, soft computing, and ambient assisted living / Eds. S. Omatu, M. P. Pocha, J. Bravo, *et al.* — Lecture notes in computer science ser. — Salamanca, Spain: Springer, 2009. Vol. 5518. P. 171–178. doi: `http://dx.doi.org/10.1007/978-3-642-02481-8_24`

[16] *D'haeseleer P.* Context preserving crossover in genetic programming // 1st IEEE Conference on Computational Intelligence Proceedings: Evolutionary Computation, 1994. Vol. 1. P. 256–261.

[17] *Raoult J. C.* On graph rewritings // Theor. Comput. Sci., 1984. Vol. 32. No. 1. P. 1–24. doi: `http://dx.doi.org/10.1016/0304-3975(84)90021-5`

[18] *Löwe M., Ehrig H.* Algebraic approach to graph transformation based on single pushout derivations // Graph-theoretic concepts in computer science / Ed. R. Möhring. — Lecture notes in computer science ser. — Berlin–Heidelberg: Springer, 1991. Vol. 484. P. 338–353.

[19] Handbook of graph grammars and computing by graph transformation. Vol. 3: Concurrency, parallelism, and distribution / Eds. H. Ehrig, H.-J. Kreowski, U. Montanari, G. Rozenberg. — River Edge, NJ, USA: World Scientific Publishing Co., Inc., 1999. 472 p. doi: http://dx.doi.org/10.1142/9789812814951

[20] *Soule T., Heckendorn R. B.* An analysis of the causes of code growth in genetic programming // Genet. Program. Evol. M., 2002. Vol. 3. No. 3. P. 283–309. doi: http://dx.doi.org/10.1023/A:1020115409250

Поступила в редакцию 08.12.15

Methods of the nonlinear regression model transformation*

R. A. Sologub

roman.sologub@yahoo.com

Dorodnicyn Computing Centre of the Russian Academy of Sciences, 40 Vavilova st., Moscow, Russia

The problem of the nonlinear regression models automatic construction and simplification has been addressed. The models describe the results of measurements and forecasting experiments. The generated models are designed for the approximation, analysis, and forecasting of the experimental results. To generate the models, the expert requirements in the subject field have been considered. This approach allows to get the interpretable models, adequately describing the given measurements. The goal of the paper is to investigate the problem of generation and simplification of the nonlinear regression models. The models are supposed to be the superpositions of the given parametric functions. A method of the function superposition transformation has been suggested. The superpositions category defined over the set of directed acyclic graphs corresponding to the superpositions have been considered. The isomorphic superpositions notion have been introduced and a method of their detection has been developed. An algorithm of finding the isomorphic subgraphs corresponding to the generated superpositions has been developed.

Keywords: *data analysis; regression model; nonlinear regression; model generation; superposition construction*

DOI: 10.21469/22233792.1.14.06

References

[1] Strijov, V. V., and R. A. Sologub. 2009. Inductive generation of the regression models for option volatility. *Russ. J. Computational Technologies* 5:102–113. (In Russian.)

[2] Sologub, R. A. 2013. Algorithms for the nonlinear regression models generation. *Russ. J. Information Technologies* 5:8–12. (In Russian.)

[3] Seber, G., and C. J. Wild. 2005. *Nonlinear regression*. Wiley ser. in probability and statistics. John Wiley & Sons. 2907 p.

[4] Seber, G. 2009. *The collected works of George A. F. Seber*. Wiley ser. in probability and statistics. Wiley. 792 p.

[5] Levenberg, K. 1944. A method for the solution of certain non-linear problems in least squares. *Quart. J. Appl. Math.* II(2):164–168.

*This work was done under financial support of the Russian Foundation for Basic Research (grant 14-07-31326)

[6] Ivakhnenko, A. G. 1982. *Inductive method of self-organized complex models system*. Kiev: Naukova Dumka. 296 p. (In Russian.)

[7] Madala, H. R., and A. G. Ivakhnenko. 1994. *Inductive learning algorithms for complex systems modeling*. CRC Press. 384 p.

[8] Koza, J. R. 1992. *Genetic programming: On the programming of computers by means of natural selection*. 1st ed. Complex adaptive systems ser. Cambridge, MA – London: The MIT Press. 840 p.

[9] Koza, J. R., M. A. Keane, M. J. Streeter, *et al.* 2003. *Genetic programming IV: Routine human-competitive machine intelligence*. Norwell, MA: Kluwer Academic Publs. 590 p.

[10] Banzhaf, W., F. D. Francone, R. E. Keller, and P. Nordin. 1998. *Genetic programming — an introduction: On the automatic evolution of computer programs and its applications*. San Francisco, CA: Morgan Kaufmann Publs. Inc. 490 p.

[11] Michell, M. 1998. *An introduction to genetic algorithms*. Cambridge, MA – London: The MIT Press. 164 p.

[12] Kominkova Oplatkova, Z., R. Senkerik, I. Zelinka, and M. Pluhacek. 2013. Analytic programming in the task of evolutionary synthesis of a controller for high order oscillations stabilization of discrete chaotic systems. *Comput. Math. Appl.* 66(2):177–189.

[13] Rudoy, G. I., and V. V. Strijov. 2013. Algorithms for inductive generation of superpositions for approximation of experimental data. *Informatika i ee Primeneniya — Inform. Appl.* 7(1):17–26.

[14] Ehrig, H., K. Ehrig, U. Prange, and G. Taentzer. 2006. *Fundamentals of algebraic graph transformation*. Berlin: Springer. 390 p.

[15] Naoki, M., B. McKay, N. Xuan, *et al.* 2009. A new method for simplifying algebraic expressions in genetic programmingc alled equivalent decision simplification. *Distributed computing, artificial intelligence, bioinformatics, soft computing, and ambient assisted living*. Eds. S. Omatu, M. P. Rocha, J. Bravo, *et al.* Lecture notes in computer science ser. Salamanca, Spain: Springer. 5518:171–178. doi: http://dx.doi.org/10.1007/978-3-642-02481-8_24

[16] D'haeseleer, P. 1994. Context preserving crossover in genetic programming. *1st IEEE Conference on Computational Intelligence Proceedings: Evolutionary Computation*. 1:256–261.

[17] Raoult, J. C. 1984. On graph rewritings. *Theor. Comput. Sci.* 32(1):1–24. doi: http://dx.doi.org/10.1016/0304-3975(84)90021-5

[18] Löwe, M., and H. Ehrig. 1991. *Algebraic approach to graph transformation based on single pushout derivations*. Ed. R. Möhring. Lecture notes in computer science ser. Berlin–Heidelberg: Springer. 484:338–353.

[19] Ehrig, H., H.-J. Kreowski, U. Montanari, and G. Rozenberg, eds. 1999. *Handbook of graph grammars and computing by graph transformation. Vol. 3: Concurrency, parallelism, and distribution*. River Edge, NJ: World Scientific Publishing Co., Inc. 472 p. doi: http://dx.doi.org/10.1142/9789812814951

[20] Soule, T., and R. B. Heckendorn. 2002. An analysis of the causes of code growth in genetic programming. *Genet. Program. Evol. M.* 3(3):283–309. doi: http://dx.doi.org/10.1023/A:1020115409250

Received December 8, 2015

Estimation of radio impulse parameters using the maximum likelihood method*

***K. V. Vlasova*[1], *V. A. Pachotin*[2], *D. M. Klionskiy*[3], *and D. I. Kaplun*[3]**
p_ksenia@mail.ru, VPakhotin@kantiana.ru, klio2003@list.ru, dikaplun@etu.ru
[1]Baltic Fishing Fleet State Academy, 6 Molodezhnaya st., Kaliningrad, Russia
[2]Immanuel Kant Baltic Federal University, 14 A. Nevskogo st., Kaliningrad, Russia
[3]Saint-Petersburg Electrotechnical University "LETI," 5 Prof. Popova st., St. Petersburg, Russia

The paper is devoted to the development of an algorithm for resolution and parameter estimation of radio impulses with partially overlapping spectra in the area of their nonorthogonality (the correlation coefficient varies from 0 to 0.9). The implementation of the suggested algorithm makes it possible to design filters for resolving frequency-dependent signals and, therefore, to increase the capacity of a communication channel. The maximum likelihood method has been used to obtain analytical expressions and to perform model investigations for frequency resolution of nonorthogonal signals. The dynamic range of signal parameter estimates has been found as a function of signal-to-noise ratio and correlation coefficient. It has been shown that the likelihood functional value in its global minimum allows one to estimate noise variance and the number of radio impulses in a received signal.

Keywords: *maximum-likelihood method; radio impulse; resolution; frequency-dependent signal; communication channel; nonorthogonal signals; signal-to-noise ratio*

DOI: 10.21469/22233792.1.14.07

1 Introduction

The paper discusses the problem of estimating parameters of a set of radio impulses with close frequencies. Their spectra partially overlap so that it is difficult to determine the precise number of radio impulses and estimate their parameters using the Rayleigh criterion. In this case, radio impulses are nonorthogonal in the frequency domain so that one can consider several radio impulses orthogonal if their spectra satisfy the Rayleigh criterion. It is necessary to determine amplitudes and frequencies of a set of nonorthogonal radio impulses (when the reception time is known) and their duration.

Modern radio systems function according to the analysis of orthogonal signals. Active pulse radar systems can be used for estimating the parameters of two or even greater number of targets, if their correlation functions do not overlap, i. e., when they are orthogonal relative to the reception moment. It is possible in spectroscopy to separate two spectral lines, if the Rayleigh criterion is satisfied (i. e., orthogonality in the frequency domain is observed). At present, if the orthogonality is not satisfied, this might result in gross errors in signal parameter estimation or a total failure in signal detection.

Nowadays, there are several techniques applied in practice for improving the quality of signal parameter estimation due to a better resolution: Prony technique, MUSIC, etc. [1]. The most widely used techniques are based on the optimal reception theory [2, 3]. However, the implementation of the aforementioned techniques is connected with a number of difficulties [4–6]. The problem of signal parameter estimation and resolution within the scope of the optimal reception theory can be handled only when the signals under study are orthogonal [7–12].

*The work is supported by Contract No. 02.G25.31.0149 dated 01.12.2015 (Ministry of Education and Science of Russian Federation).

However, when one works in the area of signal nonorthogonality, the logarithm of the likelihood function has a complex structure with a lot of local minima and its minimization is ambiguous.

Let us employ the results of the optimal reception theory, which were developed and extended for the area of signal nonorthogonality. The authors suggest the solution of this problem for nonorthogonal signals using the logarithm of the likelihood function transformed by the system of likelihood equations obtained for energetic signal parameters [13]. As a result of this transform, minimization of the logarithm of the likelihood function proves to be ambiguous. The logarithm of the likelihood function (after transformation) presents a surface in a multidimensional space of nonenergetic signal parameters. The minimum of this surface is the base for estimating both nonenergetic and energetic signal parameters. The minimum of the surface of the logarithm of the likelihood function for a set of signals is the same and, therefore, the term "resolution of similar signals" is not used and one can solve the problem of signal parameter estimation in the area of their nonorthogonality.

The solution of the signal parameter estimation problem depends on the determinant of a correlation matrix when the Rao–Kramer variance estimate is obtained.

2 Theoretical background

This section contains the main analytical expressions for an algorithm for processing a set of nonorthogonal (in the frequency domain) radio impulses based on the transformed likelihood functional. Two radio impulses are used to obtain certain expressions for calculating amplitude variances. These variances were used for estimating the application area of the algorithm depending on the correlation coefficient between the signals.

Consider a received signal consisting of N radio impulses in the complex form:

$$\hat{y}(t) = \sum_{n=1}^{N} \hat{U}_n e^{i\omega_n t} + U_n(t) \tag{1}$$

where $\hat{U}_n$ is the complex amplitude of the nth radio impulse; ω_n is the circular frequency of the nth radio impulse; and $U_n(t)$ is the additive Gaussian noise with the mean value equal to zero, variance equal to σ^2 and the correlation interval equal to τ_k.

Using (1), one can obtain the logarithm of the likelihood function:

$$\ln L(\acute{\boldsymbol{\lambda}}) = -\frac{1}{2\sigma^2\tau_k}\int_0^T \left|\hat{y}(t) - \sum_{n=1}^{N} \widehat{\acute{U}}_n e^{i\acute{\omega}_n t}\right|^2 dt = -\frac{1}{2\sigma^2\tau_k}\Delta(\widehat{\acute{U}}_n, \acute{\omega}_n) \tag{2}$$

where $\acute{\boldsymbol{\lambda}}$ is the vector of the estimated parameters of radio impulses; $\Delta(\widehat{\acute{U}}_n, \acute{\omega}_n)$ is the likelihood functional; T is the duration of the received signal; and the dashes on top of the variables correspond to the parameters that have to be estimated. The received message $\hat{y}(t)$ corresponds to the mathematical model of a signal with N copies differing by $\acute{\omega}_1, \ldots, \acute{\omega}_N$. The energy of the message depends on complex amplitudes of radio impulses $\widehat{\acute{U}}_n$ and does not depend on other parameters $(\acute{\omega}_1, \ldots, \acute{\omega}_N)$. Therefore, one can consider complex amplitudes to be energetic parameters and frequencies — nonenergetic parameters.

Minimization of (2) is ambiguous due to variability of amplitudes and frequencies of radio impulses, which leads to ambiguity in the solution of the system of likelihood equations in the area of nonorthogonality. The surface of functional (2) has many local minima, which makes it difficult to estimate the parameters of the received message. If it is necessary to obtain one

minimum of the likelihood surface, Eq. (2) should be transformed in order to exclude energetic parameters from the minimization process.

Amplitude exclusion requires one to differentiate (2) and then solve the corresponding equation. As a result, one will have a system of likelihood equations:

$$\left.\begin{array}{l} \int\limits_0^T \left(\hat{y}(t) - \sum_{n=1}^N \widehat{\acute{U}}_n e^{i\acute{\omega}_n t}\right) e^{-i\acute{\omega}_1 t}\, dt = 0\,; \\ \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\,; \\ \int\limits_0^T \left(\hat{y}(t) - \sum_{n=1}^N \widehat{\acute{U}}_n e^{i\acute{\omega}_n t}\right) e^{-i\acute{\omega}_N t}\, dt = 0\,. \end{array}\right\} \tag{3}$$

These equations can be used to find the amplitudes of N radio impulses, which depend on the estimated frequencies $\acute{\omega}_1, \ldots, \acute{\omega}_N$. Now, designate the amplitudes as $\widehat{\acute{U}}1_n$.

The likelihood functional can also be written in the following form:

$$\Delta(\widehat{\acute{U}}_n, \acute{\omega}_n) = \int\limits_0^T \left|\hat{y}(t) - \sum_{n=1}^N \widehat{\acute{U}}_n e^{i\acute{\omega}_n t}\right|^2 dt =$$

$$= \int\limits_0^T \left(\hat{y}(t) - \sum_{n=1}^N \widehat{\acute{U}}_n e^{i\acute{\omega}_n t}\right)\left(\hat{y}^*(t) - \widehat{\acute{U}}_n^* e^{i\acute{\omega}_1 t} - \cdots - \widehat{\acute{U}}_n^* e^{i\acute{\omega}_N t}\right) dt\,.$$

After certain transformations, one arrives at

$$\Delta(\widehat{\acute{U}}_n, \acute{\omega}_n) = \int\limits_0^T |\hat{y}(t)|^2 - \int\limits_0^T \left(\hat{y}^*(t) \sum_{n=1}^N \widehat{\acute{U}}_n e^{i\acute{\omega}_n t}\right) dt$$

where the asterisk means complex conjugate.

However, system (3) can be used for finding the amplitudes of radio impulses $\widehat{\acute{U}}_n$ depending on the frequency estimates $\acute{\omega}_1, \ldots, \acute{\omega}_N$. These amplitudes can be designated as $\widehat{\acute{U}}1_n$. Now, one can write the transformed likelihood functional

$$\Delta(\acute{\omega}_n) = \int\limits_0^T |\hat{y}(t)|^2 - \int\limits_0^T \left(\hat{y}^*(t) \sum_{n=1}^N \widehat{\acute{U}}1_n e^{i\acute{\omega}_n t}\right) dt\,, \tag{4}$$

depending on only nonenergetic radio impulse parameters, i. e., the estimated frequencies $\acute{\omega}_1, \ldots, \acute{\omega}_N$. Therefore, the surface will have only one global minimum indicating the estimated frequency values $\acute{\omega}_1, \ldots, \acute{\omega}_N$. Substituting them in the likelihood system (3), one can estimate the amplitude values $\widehat{\acute{U}}_n$.

If the number of radio impulses in the received signal and in the signal model is the same, the likelihood functional (4) determines the noise variance in the received message. However, in practice, the apriori information on the expected number of radio impulses is not always available. This fact can be illustrated using the following example. In communications applications, one can receive two or a greater number of radio impulses with unknown frequencies. The number of radio impulses is also unknown and they have to be resolved. Another example

is related to radio spectroscopy, where the number of spectral lines is unknown. The spectral lines can be orthogonal or nonorthogonal. These examples confirm the necessity of estimating the number of radio impulses in the received message. The authors suggest an algorithm based on changing the number of signal copies N in the mathematical model and finding the values of the likelihood functional (4) in its minimum.

If the number of signal copies N in the mathematical model is greater than the number of radio impulses in the received message, the amplitudes of extra components are close to zero since they are determined by separate noise maxima. The minimal value of likelihood (4) will be determined by the noise variance.

If the number of signal copies N is smaller than in the received message, the value of the likelihood functional rises dramatically since it determines the noise variance in the received message and the radio impulses which are not included in the model.

Thus, processing the received message requires one to increase the number of signal copies N till the minimum value of the likelihood functional stops decreasing.

Evaluation of the solution quality in the optimal reception theory is based on estimating variances using Rao–Kramer estimates. Consider two radio impulses in a data set starting at the same time point. One can estimate the variance of radio impulse amplitudes in the minimal point of the likelihood functional for $\acute{\omega}_1 = \omega_1$ and $\acute{\omega}_2 = \omega_2$. By computing the elements of the Fisher information matrix according to the expression

$$J_{ij} = -M\left(\frac{d^2 \ln L(\acute{\boldsymbol{\lambda}})}{d\lambda_i \lambda_j}\right),$$

one can find a new matrix consisting of the elements

$$J_{ij} = \int_0^T e^{i(\acute{\omega}_i - \acute{\omega}_j)t}\, dt\,.$$

These are complex correlation coefficients between the radio impulses. The diagonal elements of a matrix, which is inverse to the Fisher information matrix, define the variances of radio impulse amplitudes. The inverse matrix has the following form:

$$\hat{D}_U = \frac{A_{ij}}{\det \hat{J}}$$

where A_{ij} is the algebraic adjunct for the element of the Fisher matrix with the indices i, j and $\det \hat{J}$ is the determinant of the Fisher information matrix.

Thus, the variance of radio impulse amplitudes is determined by $\det \hat{J}$ of the Fisher information matrix. In the case of two radio impulses, the amplitude variance will have the following form:

$$D_{U_1} = D_{U_2} = \frac{\sigma^2}{K(1 - |\hat{r}|^2)}$$

where K is the number of noncorrelated noise samples on the processing interval T and $\hat{r}$ is the normalized correlation coefficient between radio impulses.

If $|\hat{r}| = 0$, the amplitude variance will have the following form:

$$D_{U_0} = \frac{\sigma^2}{K}.$$

Now, let us introduce the normalized variance:

$$D = \frac{D_U}{D_{U_0}} = \frac{1}{1 - |\hat{r}|^2}.$$

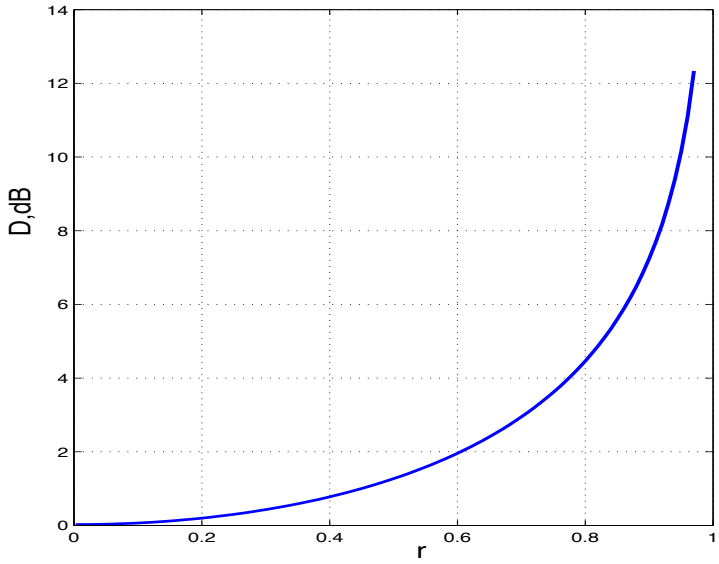


Figure 1 Relative variance vs. normalized correlation coefficient. Relative variance increases by ~ 7 dB in the range $|\hat{r}| = 0$–0.9; when $|\hat{r}| > 0.9$, the relative variance rises dramatically

This expression shows that for $|\hat{r}| = 0$–0.9 the changes of the relative variance are approximately 7 dB. In the area where $|\hat{r}| > 0.9$, the relative variance grows abruptly as shown in Fig. 1. Thus, if a signal contains two radio impulses, the problem of radio impulse parameter estimation will be solved for the following range of the normalized correlation coefficient: $|\hat{r}| = 0$–0.9.

3 Model studies

This section is devoted to studying the problem of estimating the parameters of two radio impulses in the area of their nonorthogonality using the maximum likelihood technique and spectral analysis. It will be shown that the maximum likelihood technique can be used for solving this problem in the area of nonorthogonality. The results of the experiments are provided to estimate the quality of the solution depending on the signal-to-noise ratio and correlation coefficient between two radio impulses. The estimation of the noise variance in the received message is illustrated using the dataset (the functional value in its minimum).

Model studies have been carried out for a signal containing two radio impulses combined in the time domain. The radio impulses have the following parameters: amplitudes $U_1 = = 2$ and $U_2 = 1$; initial phases $\varphi_1 = 10°$ and $\varphi_2 = 170°$; frequencies $f_1 = 2$ kHz and $f_2 = 2.08$ kHz; signal-to-noise ratio is 20 dB; and radio impulse duration is 25 ms. The Rayleigh restriction for radio impulse resolution arises when the frequencies become close by $\Delta f = 40$ Hz.

Figure 2 shows the normalized correlation coefficient vs. the frequency difference Δf between two signals. When $\Delta f = 40$ and 80 Hz, the correlation coefficient is equal to zero and the signals

are orthogonal. The values $\Delta f < 40$ Hz correspond to the signal nonorthogonality area. The correlation coefficient values different from zero for $\Delta f > 40$ Hz are caused by side lobes of the signal spectrum.

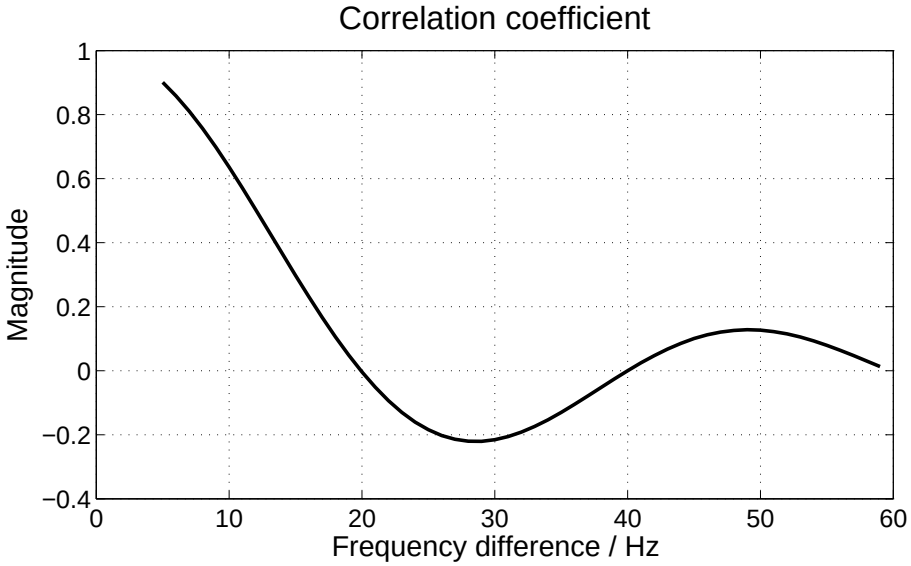


Figure 2 Correlation coefficient between radio impulses. Radio impulse duration is 25 ms. The radio impulses are orthogonal for 40, 80, and 120 Hz. When the radio impulse frequency difference is 10 Hz and the correlation coefficient is equal to 0.9, the theoretical estimate of the working area is confirmed

Figure 3 shows the received signal containing two radio impulses with the frequencies $f_1 = 2$ kHz and $f_2 = 2.08$ kHz. According to Fig. 2, these signals are orthogonal since the frequency difference $\Delta f = 80$ Hz leads to the correlation coefficient $\hat{r} = 0$.

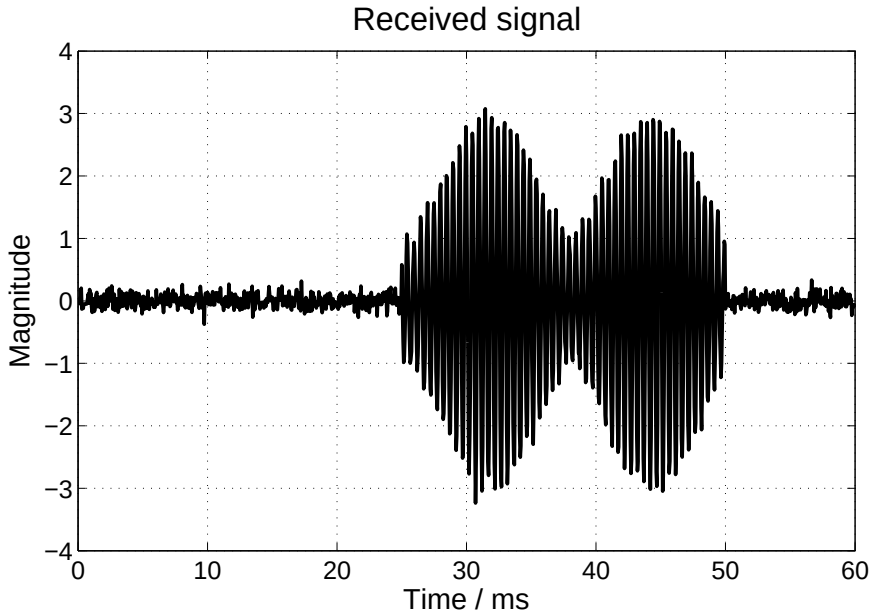


Figure 3 Received signal containing two orthogonal radio impulses combined in the time domain

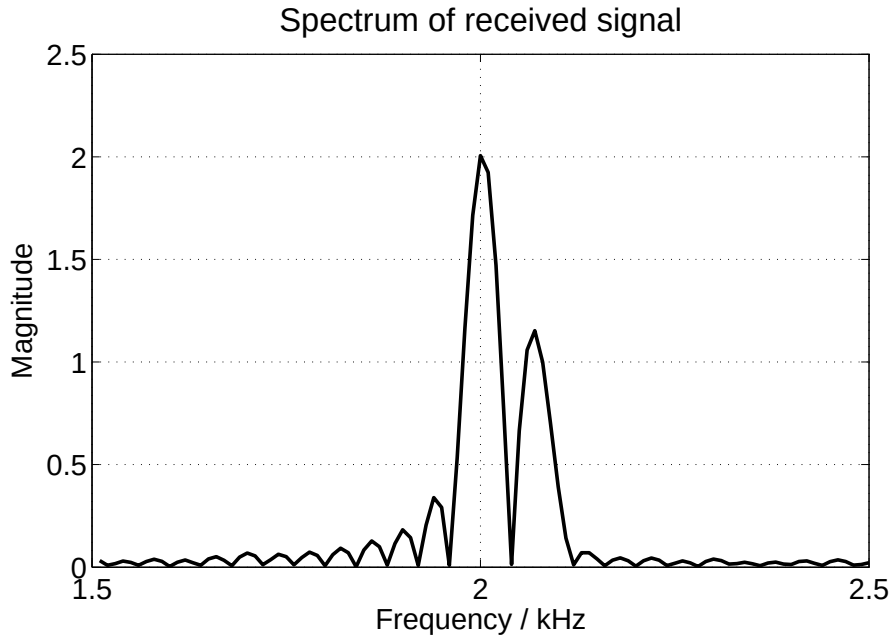


Figure 4 Spectrum of two orthogonal radio impulses combined in the time domain. The maxima of spectral lines determine amplitude and frequency estimates of the radio impulses

The spectra of these two radio impulses is shown in Fig. 4. The side lobes of the first radio impulse ($U_1 = 2$) distort the amplitude of the second radio impulse ($U_2 = 1$).

Figure 5 shows the surface of the transformed inverse likelihood functional:

$$\Delta 1 = \frac{1}{\Delta(\acute{\omega}_1, \acute{\omega}_2)}\,.$$

The maximum of the functional surface is the only one which determines the unambiguity of the problem solution. The location of the maximum makes it possible to find the values of estimation frequencies $\acute{\omega}_1$ and $\acute{\omega}_2$. Their substitution into likelihood equations makes it possible to find the estimates of complex amplitudes of radio impulses. The value of maximum of the likelihood functional allows one to find the noise variance in a received signal. Figures 3–5 illustrate the possibility of obtaining the solution in the area of radio impulse orthogonality.

Figures 6–8 show the possibility of obtaining the solution in the area of radio impulse nonorthogonality. In this case, the radio impulse frequencies are $f_1 = 2$ kHz and $f_2 = 2.02$ kHz. The difference is 20 Hz. Figure 6 shows the received signal.

Figure 7 shows the spectrum of two radio impulses, which makes it difficult to obtain the information on radio impulse parameters.

Figure 8 shows the transformed likelihood functional, which allows one to estimate the radio impulse parameters.

Figure 9 shows the variations of radio impulse magnitude estimates depending on the changes in radio impulse frequencies. The best amplitude estimates can be found when $\Delta f = 10$ Hz. This corresponds to the correlation coefficient between radio impulses $\hat{r} \sim 0.9$, which fully confirms the theoretical estimate of the working area ($\hat{r} = 0$–0.9) for parameter estimation of two radio impulses.

Figure 10 shows radio impulse frequency estimates vs. their frequency differences. When the signal-to-noise ratio is ~ 20 dB and radio impulse duration is 25 ms, the best estimates are obtained for $\Delta f = 10$ Hz. Thus, one can point out that the suggested technology of radio

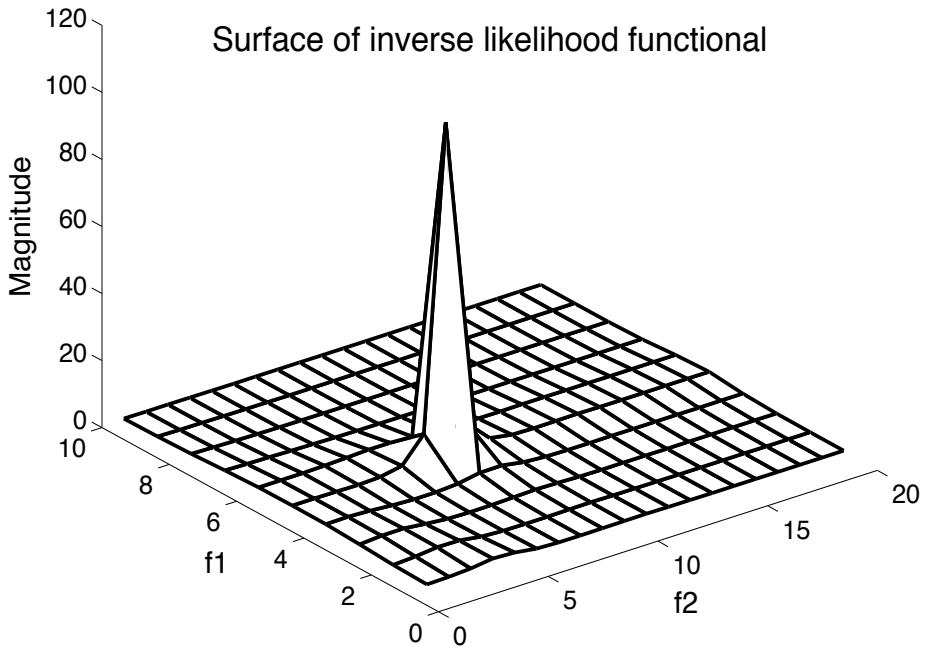


Figure 5 Surface of the inverse likelihood functional. The single maximum determines the parameter estimates of two radio impulses and the term "resolution" is not required. The base width of the maximum of the likelihood functional depends on a signal-to-noise ratio and its maximal value (95.45) determines the estimate of the noise variance in a received signal (−19.79 dB relative to 1)

impulse parameter estimation is characterized by high resolution. When the signal-to-noise ratio is 20 dB, the resolution of two radio impulses has increased 4 times in comparison with the Rayleigh resolution.

Figure 11 shows the statistics of radio impulse magnitudes depending on the signal-to-noise ratio when the radio impulse frequency is $\Delta f = 20$ Hz. As can be seen from the figure, the amplitude estimates are quite satisfactory when the signal-to-noise ratio exceeds 0 dB.

Figure 12 shows the statistics of radio impulse frequencies depending on the signal-to-noise ratio for $\Delta f = 20$ Hz. As can be seen from the figure, radio impulse frequency estimates are quite satisfactory for signal-to-noise ratios exceeding −5 dB.

Figure 13 illustrates the changes of maxima of the inverse transformed likelihood functional depending on the signal-to-noise ratio in dB. The linear dependence confirms the theoretical suggestion that the noise variance can be estimated on the basis of the maximum of the inverse transformed likelihood functional.

4 Concluding remarks

An algorithm for solving the problem of parameter estimation of a set of nonorthogonal radio impulses has been introduced using the maximum likelihood method on the basis of the transformed likelihood functional. Theoretical and practical studies lead to the following conclusions:

– the transformed likelihood functional allows one to solve the problem of resolving and parameter estimation of two or several radio impulses in the area of their nonorthogonality. Application of this technique is efficient for different radio engineering problems and can be used for improving the capacity of communication channels using frequency division multiplexing;

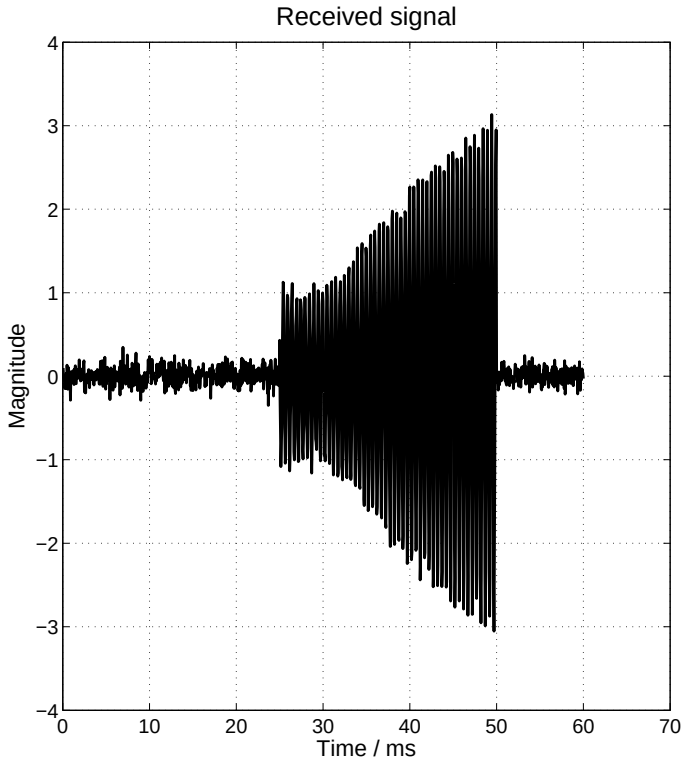


Figure 6 Received signal containing two nonorthogonal impulses combined in the time domain

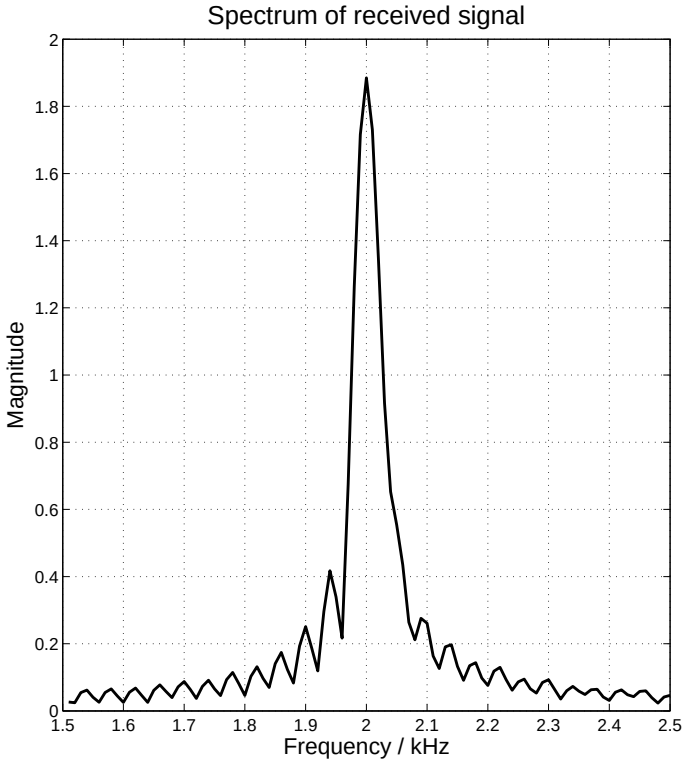


Figure 7 Spectrum of two nonorthogonal radio impulses combined in the time domain

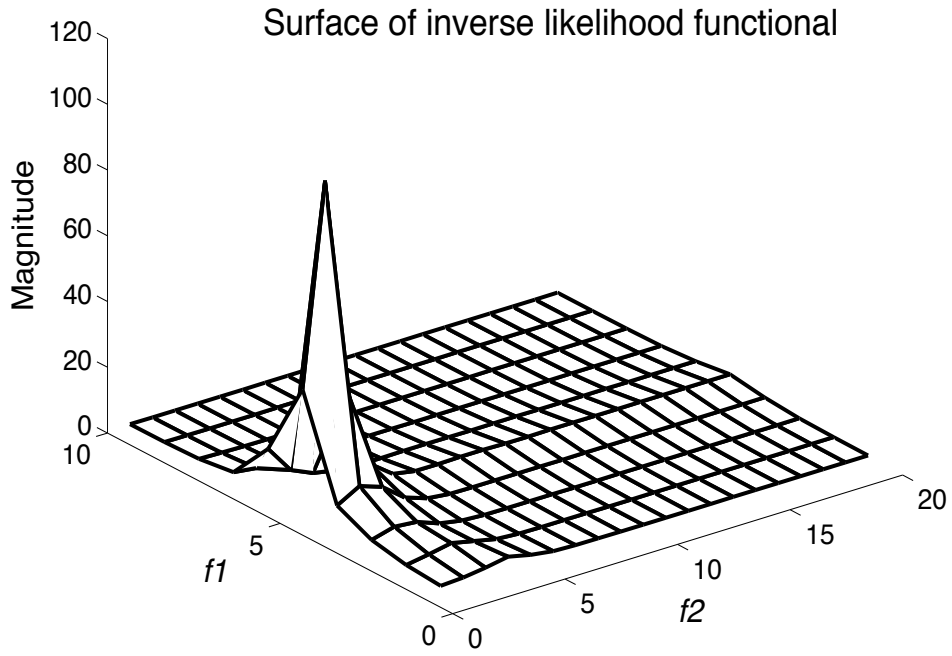


Figure 8 Surface of the inverse likelihood functional obtained in the area of radio impulse nonorthogonality

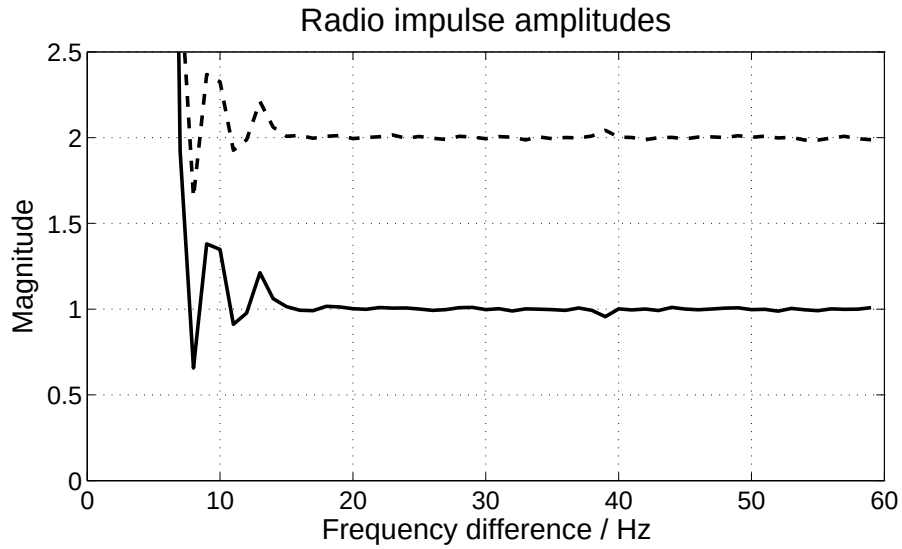


Figure 9 Variation of the amplitude estimates of radio impulses depending on the changes of the frequency differences. The nonorthogonality area of the radio impulses for $\Delta f \leqslant 40$ Hz. The best amplitude estimates for radio impulse frequency differences of 10 Hz. Radio impulse resolution has increased 4 times in comparison with the Rayleigh resolution

– the working area for parameter estimation of a set of radio impulses is determined by determinant variations of a correlation matrix of a set of signals. Conditionality of the correlation matrix influences the possibility of resolving a set of signals. In the case of two radio impulses the working area depends on the normalized correlation coefficient in the range 0–0.9;

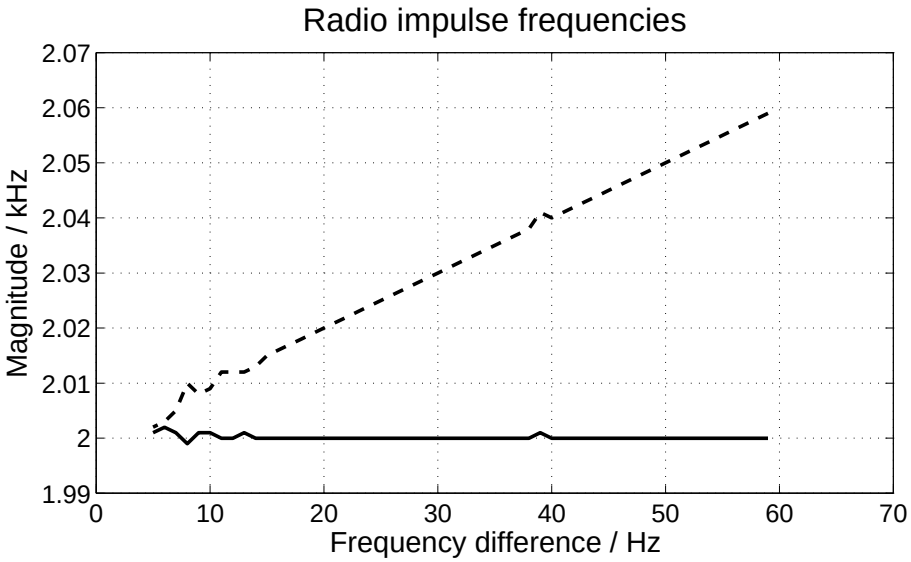


Figure 10 Frequency estimates vs. frequency differences. The area of radio impulse nonorthogonality for $\Delta f \leqslant 40$ Hz. The best frequency estimates for radio impulse frequency difference of 10 Hz. Radio impulse resolution has increased 4 times in comparison with the Rayleigh resolution

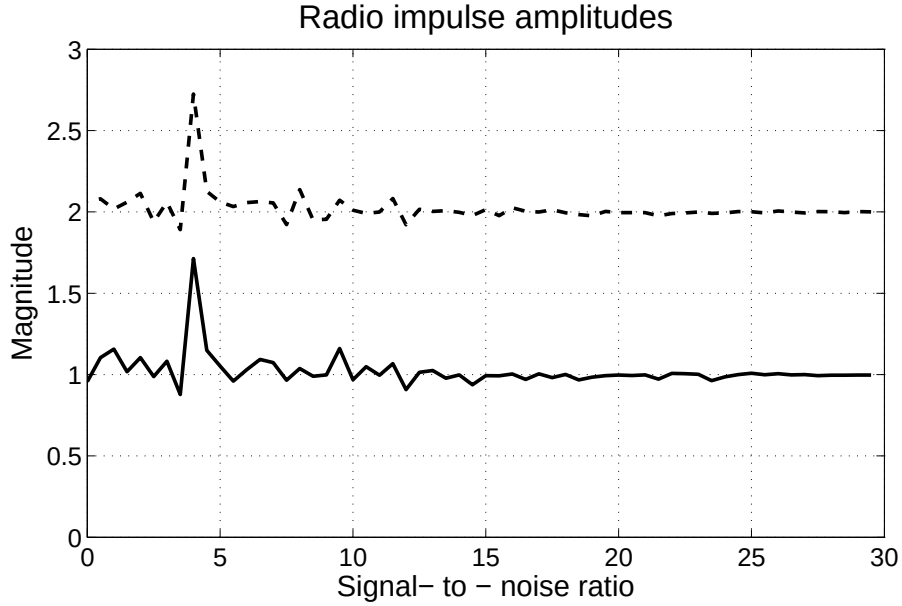


Figure 11 Statistics of radio impulse amplitude estimates. The difference of radio impulse frequencies is 20 Hz (nonorthogonality area). The estimates are satisfactory for signal-to-noise ratios exceeding 0 dB

– the value of the transformed likelihood functional in its global minimum allows one to estimate the noise variance in a dataset and can be used as a criterion for estimating the number of radio impulses.

References

[1] Marple, S. L., J. 1986. *Digital spectral analysis: With applications*. Upper Saddle River, NJ: Prentice-Hall, Inc. 492 p.

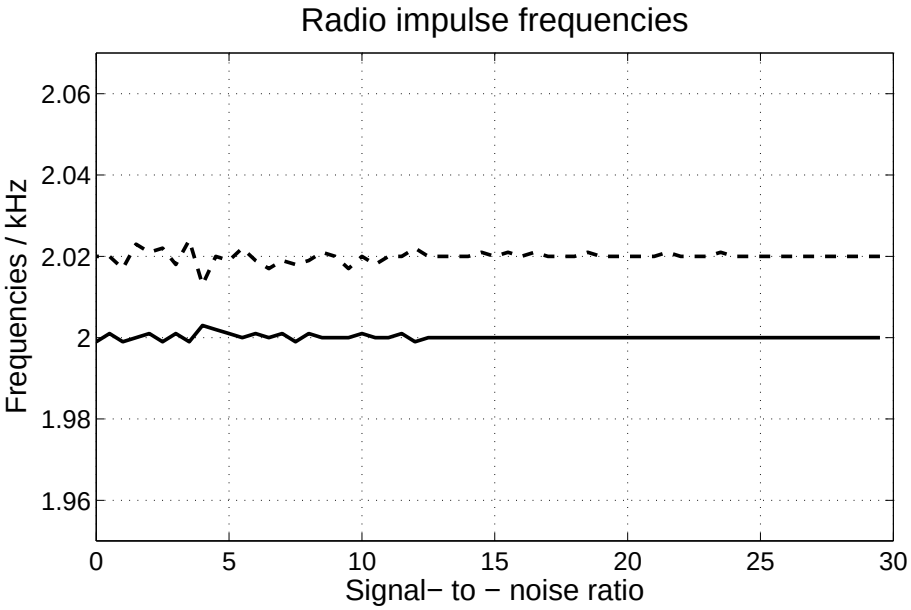


Figure 12 Statistics of radio impulse frequency estimates. The difference of radio impulse frequencies is 20 Hz (nonorthogonality area). The estimates are satisfactory for signal-to-noise ratios exceeding −5 dB

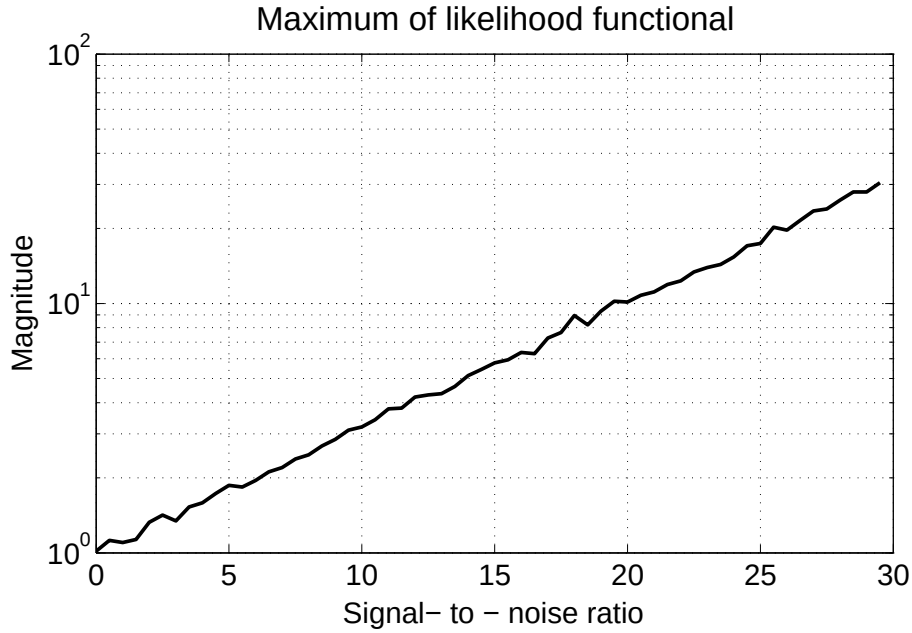


Figure 13 Maximum of the likelihood functional vs. signal-to-noise ratio. The linearity of the dependence allows us to estimate the noise variance in the received message using the maximum of the inverse likelihood functional

[2] Helstrom, C. W. 1960. *Statistical theory of signal detection*. International ser. of monographs on electronics and instrumentation. Macmillan. Vol. 9. 364 p.

[3] Tikhonov, V. I. 1983. *Optimal signal reception*. Moscow: Radio and Communications. 320 p.

[4] Chigov, A. A. 2010. *Superrayleigh resolution. Classical approach to the problem*. Moscow: Krasand. 104 p. (In Russian.)

[5] Volkov, V. Y., L. S. Turnetskyi, and A. V. Oneshko. 2011. Straight line edge extraction in noisy images. *Informatsionno-upravliaiushchie sistemy — Information and Control Systems* 4(53):13–17.

[6] Volkov, V. Y. 2013. *Discrete filtering techniques and problems of image processing in radio engineering observation systems*. St. Petersburg: Saint Petersburg University of Telecommunications. 144 p. (In Russian.)

[7] Trifonov, A. P., and Yu. S. Shinakov. 1986. *Joint signal distinction and their parameter estimation in the presence of noise*. Moscow: Radio and Communication. 246 p. (In Russian.)

[8] Stoica, P., B. Ottersten, M. Viberg, and R. L. Moses. 1996. Maximum likelihood array processing for stochastic coherent sources. *IEEE Trans. Signal Proces.* 44(1):96–105. doi: `http://dx.doi.org/10.1109/78.482015`

[9] Troyan, V. N., and Yu. V. Kiselev. 2000. *Statistical techniques of processing and interpretation of geophysical data*. St. Petersburg: St. Petersburg Polytechnical University. 578 p. (In Russian.)

[10] Perov, A. I. 2003. *Statistical theory of radio engineering systems*. Moscow: Radio Engineering. 400 p.

[11] Zelniker, E. E., and I. V. L. Clarkson. 2003. Maximum-likelihood circle-parameter estimation via convolution. *7th Conference on Digital Image Computing: Techniques and Applications Proceedings*. Eds. C. Sun, H. Talbot, S. Ourselin, and T. Adriaansen. Sydney. 509–518.

[12] Babu, P. 2012. Spectral analysis of nonuniformly sampled data and applications. Uppsala, Sweden: Uppsala University. Dissertation.

[13] Pachotin, V. A., V. A. Bessonov, S. V. Molostova, and K. V. Vlasova. 2008. *Theoretical fundamentals of optimal signal processing*. Kaliningrad. 186 p.

Received June 15, 2015

Оценивание параметров радиоимпульса с использованием метода максимального правдоподобия*

***К. В. Власова*[1], *В. А. Пахотин*[2], *Д. М. Клионский*[3], *Д. И. Каплун*[3]**

`p_ksenia@mail.ru, VPakhotin@kantiana.ru, klio2003@list.ru, dikaplun@etu.ru`

[1]Балтийская государственная академия рыбопромыслового флота, Россия, г. Калининград, ул. Молодежная, 6

[2]Балтийский федеральный университет имени Иммануила Канта, Россия, г. Калининград, ул. А. Невского, 14

[3]Санкт-Петербургский государственный электротехнический университет «ЛЭТИ», Россия, г. Санкт-Петербург, ул. Профессора Попова, 5

Статья посвящена разработке алгоритма для разрешения и оценивания параметров радиоимпульсов с частично перекрывающимися спектрами в области их неортогональности (коэффициент корреляции изменяется в пределах от 0 до 0,9). Предложенный алгоритм позволяет проектировать фильтры для разрешения частотно-зависимых сигналов и,

*Работа поддержана контрактом № 02.G25.31.0149 от 01.12.2015 (Минобрнауки России)

как следствие, появляется возможность повышения пропускной способности канала связи. В статье использован метод максимального правдоподобия для получения аналитических выражений и проведения модельных исследований для частотного разрешения неортогональных сигналов. Динамический диапазон оценок параметров сигналов был определен как функция отношения сигнал/шум и коэффициента корреляции. Показано, что значения функционала правдоподобия в точке глобального минимума позволяют оценить дисперсию шума и количество радиоимпульсов в принятом сигнале.

Ключевые слова: *метод максимального правдоподобия; радиоимпульс; разрешение; частотно-зависимый сигнал; канал связи; неортогональные сигналы; отношение сигнал/шум*

DOI: 10.21469/22233792.1.14.07

Литература

[1] *Марпл-мл. С. Л.* Цифровой спектральный анализ и его приложения / Пер. с англ. — М.: Мир, 1990. 584 с. (*Marple S. L.* Digital spectral analysis: With applications. — Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1986. 492 p.)

[2] *Хелстром К.* Статистическая теория обнаружения сигналов / Пер. с англ. — М.: ИЛ, 1963. 432 с. (*Helstrom C. W.* Statistical theory of signal detection. — International ser. of monographs on electronics and instrumentation. — Macmillan, 1960. Vol. 9. 364 p.)

[3] *Тихонов В. И.* Оптимальный прием сигналов. — М.: Радио и связь, 1983. 320 с.

[4] *Чижов А. А.* Сверхрэлеевское разрешение. Т. 1: Классический взгляд на проблему. — М.: КРАСАНД, 2010. 96 с.; Т. 2: Преодоление фактора некорректности обратной задачи рассеяния и проекционная радиолокация. — М.: КРАСАНД, 2010. 104 с.

[5] *Волков В. Ю., Турнецкий Л. С., Онешко А. В.* Выделение прямолинейных кромок на зашумленных изображениях // Информационно-управляющие системы, 2011. Вып. 4(53). С. 13–17.

[6] *Волков В. Ю.* Методы дискретной фильтрации и задачи обработки изображений в радиотехнических системах наблюдения. — СПб.: СПбГУТ, 2013. 144 с.

[7] *Трифонов А. П., Шинаков Ю. С.* Совместное различение сигналов и оценка их параметров на фоне помех. — М.: Радио и связь, 1986. 246 с.

[8] *Stoica P., Ottersten B., Viberg M., Moses R. L.* Maximum likelihood array processing for stochastic coherent sources // IEEE Trans. Signal Proces., 1996. Vol. 44. No. 1. P. 96–105. doi: `http://dx.doi.org/10.1109/78.482015`

[9] *Троян В. Н., Киселев Ю. В.* Статистические методы обработки и интерпретации геофизических данных. — СПб.: Изд-во С.-Петерб. ун-та, 2000. 578 с.

[10] *Перов А. И.* Статистическая теория радиотехнических систем: Учебное пособие для вузов. — М.: Радиотехника, 2003. 400 с.

[11] *Zelniker E. E., Clarkson I. V. L.* Maximum-likelihood circle-parameter estimation via convolution // 7th Conference on Digital Image Computing: Techniques and Applications Proceedings / Eds. C. Sun, H. Talbot, S. Ourselin, T. Adriaansen. — Sydney, 2003. P. 509–518.

[12] *Babu P.* Spectral analysis of nonuniformly sampled data and applications. — Uppsala, Sweden: Uppsala University, 2012. Dissertation.

[13] *Пахотин В. А., Бессонов В. А., Молостова С. В., Власова К. В.* Теоретические основы оптимальной обработки сигналов. — Калининград: Изд-во РГУ им. И. Канта, 2008. 186 с.

Поступила в редакцию 15.06.15

Поиск внешней и внутренней границ радужной оболочки на изображении глаза методом парных градиентов*

Ю. С. Ефимов, И. А. Матвеев
yuri.efimov@phystech.edu, matveev@ccas.ru
Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9
Вычислительный центр РАН им. А.А.Дородницына, Россия, г. Москва, ул. Вавилова, 40

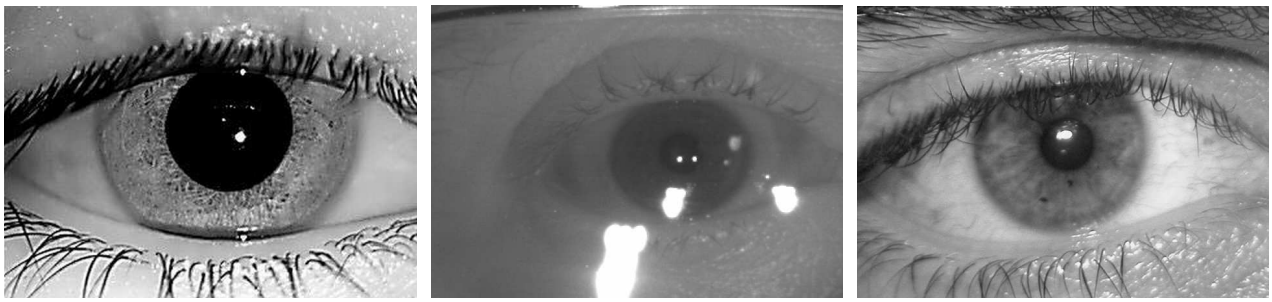
Рассматривается задача выделения области радужной оболочки на изображении глаза. Внешняя и внутренняя границы радужной оболочки аппроксимируются окружностями. Для отбора точек, принадлежащих предполагаемым окружностям, предлагается использовать модификацию преобразования Хафа, использующую пары градиентов яркости. Вводятся вероятностные коэффициенты подобия для построения изображения-аккумулятора. Для анализа эффективности алгоритма и демонстрации его работы используются материалы открытой базы изображений радужки.

Ключевые слова: *метод парных градиентов; компьютерное зрение; поиск окружностей; преобразование Хафа*

DOI: 10.21469/22233792.1.14.08

1 Введение

В современной биометрии существует проблема выделения радужной оболочки на изображении глаза человека. Требуется найти две приближенно концентрические окружности, соответствующие внутренней и внешней границам радужки. Внутри радужка ограничена зрачком — темной круглой областью, снаружи — белком глаза — наоборот, светлым фоном. В процессе поиска граничных окружностей возникают сложности, связанные с шумами изображения, искажениями формы радужной оболочки, бликами при съемке, а также возможным наличием посторонних объектов, таких как ресницы или части оправы очков. Примеры таких дефектов изображений приведены на рис. 1.



(*а*) (*б*) (*в*)

Рис. 1 Изображения глаз

Распространенным алгоритмом выделения объектов заданного класса на изображениях является преобразование Хафа [1]. Оно находит на изображении плоские кривые,

*Работа выполнена при финансовой поддержке РФФИ (проект 16-07-01171).

заданные параметрически, в том числе и окружности. Идея преобразования Хафа заключается в поиске локальных максимумов в пространстве параметров. Для окружностей пространство параметров является трехмерным, что увеличивает сложность поиска максимумов до $O(N^3)$. Существуют разновидности преобразования Хафа, позволяющие уменьшить вычислительную сложность. В работах [2–4], например, предлагаются подходы к детектированию окружностей, использующие информацию о градиенте яркости в каждой точке изображения для отбора точек интереса. В [2, 4] описывается случай выделения концентрических окружностей, использующий двухмерный массив-аккумулятор, в [3] предлагается использовать кривые равной освещенности и сложную систему голосования в пространстве параметров. На точность алгоритмов также влияет наличие шума на изображении и искажение формы искомых объектов. В таких случаях часто применяют стохастические алгоритмы, например метод случайного выделения точек интереса на каждой итерации [5], значительно уменьшающий время работы алгоритма по сравнению с классическим преобразованием Хафа, и основанный на нем метод случайного выделения окружностей [6], не использующий массив-аккумулятор и позволяющий тем самым уменьшить требования к ресурсам системы, а также сократить вычислительную сложность. Для поиска окружностей также применяется метод парного вероятностного голосования [7], оперирующий прямыми в трехмерном пространстве параметров. Вероятностная весовая схема с быстрым алгоритмом поиска моды повышает устойчивость метода [7] к шумам и искажениям исходного изображения, а также сокращает вычислительное время.

Суть предлагаемого метода состоит в сочетании нескольких подходов. Точки интереса, т. е. точки, предположительно лежащие на окружностях, отбираются при помощи анализа вектора градиента яркости. Отбираются точки с достаточно большим его значением. Ищутся такие пары точек, векторы градиентов яркости которых приблизительно противоположно направлены, причем данная пара точек должна лежать на прямой, приблизительно коллинеарной этим векторам. Далее определяются центры предполагаемых окружностей как центры отрезков, соединяющих точки. Для каждой пары точек определяются весовые коэффициенты, используемые при голосовании в двухмерном пространстве параметров. Радиусы искомых окружностей определяются как локальные максимумы в одномерном пространстве предполагаемых радиусов.

2 Постановка задачи

Входные данные метода — растровое изображение $\mathbf{I}$ размера $W \times H$. Каждый пиксель входного изображения описывается одним байтом, что соответствует одной из двухсот пятидесяти шести градаций серого. Требуется аппроксимировать внешнюю и внутреннюю границы радужной оболочки двумя приближенно концентрическими окружностями. Решение можно записать как

$$\omega = \{x_{\mathrm{P}}, y_{\mathrm{P}}, r_{\mathrm{P}}, x_{\mathrm{I}}, y_{\mathrm{I}}, r_{\mathrm{I}}\},$$

где x_{P} и y_{P} — координаты центра; r_{P} — радиус окружности, аппроксимирующей зрачок; x_{I}, y_{I} и r_{I} — координаты центра и радиус окружности, аппроксимирующей внешнюю границу радужки. Качество работы метода оценивается на основе сравнения с экспертной разметкой. Для изображения известны «истинные» параметры границ радужки, $\tilde{\omega} = \{\tilde{x}_{\mathrm{P}}, \tilde{y}_{\mathrm{P}}, \tilde{r}_{\mathrm{P}}, \tilde{x}_{\mathrm{I}}, \tilde{y}_{\mathrm{I}}, \tilde{r}_{\mathrm{I}}\}$, определенные экспертом. Функционал качества определения центров рассчитывается как сумма абсолютных значений отклонений вычисленных абсциссы и ординаты центров от истинных значений:

$$S_c(\omega) = |x_{\mathrm{P}} - \tilde{x}_{\mathrm{P}}| + |y_{\mathrm{P}} - \tilde{y}_{\mathrm{P}}| + |x_{\mathrm{I}} - \tilde{x}_{\mathrm{I}}| + |y_{\mathrm{I}} - \tilde{y}_{\mathrm{I}}|.$$

Функционал качества определения радиусов — аналогично, как сумма абсолютных отклонений вычисленных радиусов от их истинных значений:

$$S_r(\omega) = |r_{\mathrm{P}} - \tilde{r}_{\mathrm{P}}| + |r_{\mathrm{I}} - \tilde{r}_{\mathrm{I}}|.$$

Итоговый функционал качества определяется как сумма вышеописанных:

$$S(\omega) = S_c(\omega) + S_r(\omega)\,.$$

Для оценки качества решения использовалась относительная ошибка, определяемая как отношение величины функционала качества к истинному радиусу внешней границы радужной оболочки:

$$e = \frac{S(\omega)}{r_{\mathrm{I}}}\,.$$

Аналогичным образом определялась относительная ошибка определения центра глаза:

$$e_c = \frac{S_c(\omega)}{r_{\mathrm{I}}}\,.$$

Качество решения определяется гистограммой величины относительной ошибки. Численный критерий качества решения — доля изображений, на которых ошибка не превышает 20% истинного радиуса радужки.

3 Некоторые обозначения, используемые в методологии Хафа

$\mathbf{I}$ — исходное растровое изображение указанных размеров $W \times H$.

$\boldsymbol{q} = (x, y)^{\mathsf{T}}$ — точка исходного растрового изображения.

$\boldsymbol{g}(\boldsymbol{q}) = (g_x, g_y)^{\mathsf{T}}$ — вектор градиента яркости в точке $\boldsymbol{q}$ исходного изображения.

$\boldsymbol{p} = (x_c, y_c, r)^{\mathsf{T}}$ — вектор параметров для окружности.

$\mathbf{Q}$ — двухмерное пространство параметров. В данном случае — изображение-аккумулятор, соразмерное $\mathbf{I}$. Каждой точке аккумулятора $(x, y)^{\mathsf{T}}$ соответствует центр некоторой гипотетической окружности $(x_c, y_c, r)^{\mathsf{T}} : (x_c, y_c) = (x, y)^{\mathsf{T}}$. Чем больше значение аккумулятора $\mathbf{Q}(x^*, y^*)$ в точке (x^*, y^*), тем вероятнее присутствие на исходном изображении окружности с центром $(x_c, y_c) = (x^*, y^*)^{\mathsf{T}}$.

$\mathbf{G} = \{x, y, |\mathbf{g}(x, y)|, \varphi\}$ — множество точек, принадлежащих границам объектов на изображении. Под границей объекта в данном исследовании следует понимать множество точек с большим значением модуля градиента яркости. $\mathbf{G}$ содержит координаты x и y граничной точки, модуль $|\boldsymbol{g}(x, y)|$ и угол φ направления градиента (угол отсчитывается от направления оси абсцисс).

$\mathbf{P} = \{\boldsymbol{q}_i, \boldsymbol{p}_i\}_{i=1}^{m}$ — множество точек интереса вместе с соответствующими им параметрами гипотетических окружностей $\boldsymbol{p}_i = \{x_i, y_i, r_i\}$, $\boldsymbol{p}_i \in \mathbb{R}^3$. Данные точки предположительно принадлежат границам радужки, а значит, их признаки представляют интерес для анализа.

4 Метод парных градиентов

Данный метод предполагает использование градиентов яркости в качестве критерия отбора точек интереса, т. е. точек, предположительно лежащих на одной окружности.

Как известно, окружность на плоскости можно задать вектором параметров: $\boldsymbol{p} = (x_c, y_c, r)^{\mathsf{T}} \in \mathbb{R}^3$, где пара $(x_c, y_c)^{\mathsf{T}} \in \mathbb{R}^2$ задает центр окружности и $r \in \mathbb{R}$ определяет ее

радиус. Предположим, что на изображении присутствует единственная окружность с параметрами $O = O(x_c, y_c, r)$. Тогда в ее точках модуль градиента яркости $|\boldsymbol{g}(x, y)|$ будет превосходить его значение в остальных точках изображения. В идеальном случае для пары точек $\boldsymbol{q}_1 = (x_1, y_1)^\intercal \in O$ и $\boldsymbol{q}_2 = (x_2, y_2)^\intercal \in O$, принадлежащих диаметру окружности O, векторы $\boldsymbol{g}(\boldsymbol{q}_1)$ и $\boldsymbol{g}(\boldsymbol{q}_2)$ будут противоположно направлены. В данном предположении эти векторы также будут лежать на одной прямой $d(O)$ — диаметре окружности (рис. 2).

Таким образом, можно сформулировать условия отбора для пары точек $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$:

$$||\boldsymbol{g}(\boldsymbol{q_1})|| > T_g; \quad ||\boldsymbol{g}(\boldsymbol{q_2})|| > T_g; \tag{1}$$

$$\frac{|(\boldsymbol{g}(\boldsymbol{q_1}) \cdot \boldsymbol{q_1} - \boldsymbol{q_2})|}{||\boldsymbol{g}(\boldsymbol{q_1})|| \cdot ||\boldsymbol{q_1} - \boldsymbol{q_2}||} > T_\varphi; \quad \frac{|(\boldsymbol{g}(\boldsymbol{q_2}) \cdot \boldsymbol{q_1} - \boldsymbol{q_2})|}{||\boldsymbol{g}(\boldsymbol{q_1})|| \cdot ||\boldsymbol{q_1} - \boldsymbol{q_2}||} > T_\varphi; \tag{2}$$

$$(\boldsymbol{g}(\boldsymbol{q_1}) \cdot \boldsymbol{g}(\boldsymbol{q_2})) < 0, \tag{3}$$

где условия (1) определяют отбор точек с градиентом яркости с нижним порогом T_g, а (2) задают приблизительную коллинеарность градиентов предполагаемому диаметру с точностью до T_φ, и (3) задает антиколлинеарность векторов градиентов.

Если пара точек $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}$ удовлетворяет условиям отбора, то она лежит на диаметре некоторой предполагаемой окружности $\tilde{O}$. Тогда координаты центра $\tilde{O}$ определяются как

$$\tilde{x}_c = \frac{x_1 + x_2}{2}; \quad \tilde{y}_c = \frac{y_1 + y_2}{2},$$

а радиус, соответственно, как

$$\tilde{r} = \sqrt{(x_1 - \tilde{x}_c)^2 + (y_1 - \tilde{y}_c)^2}.$$

5 Применение метода к сегментации радужки

Поиск границ радужной оболочки на изображении осуществляется в несколько шагов. На предварительном шаге входное изображение подвергается первичной обработке с целью повышения его качества. Для выделения граничных точек к изображению применяется оператор выделения границ Кэнни [8], что снимает условия (1) на величину градиента яркости. На первом шаге из числа граничных точек отбираются точки интереса,

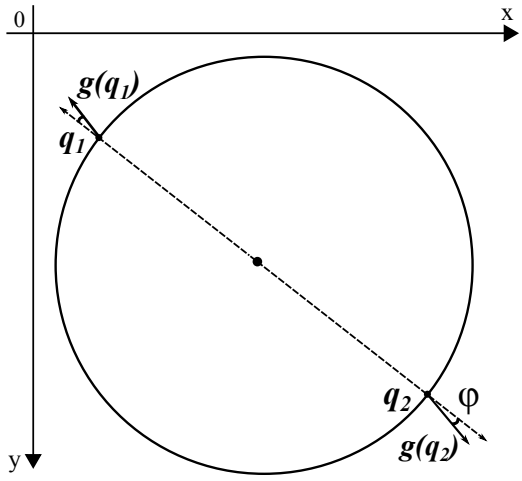
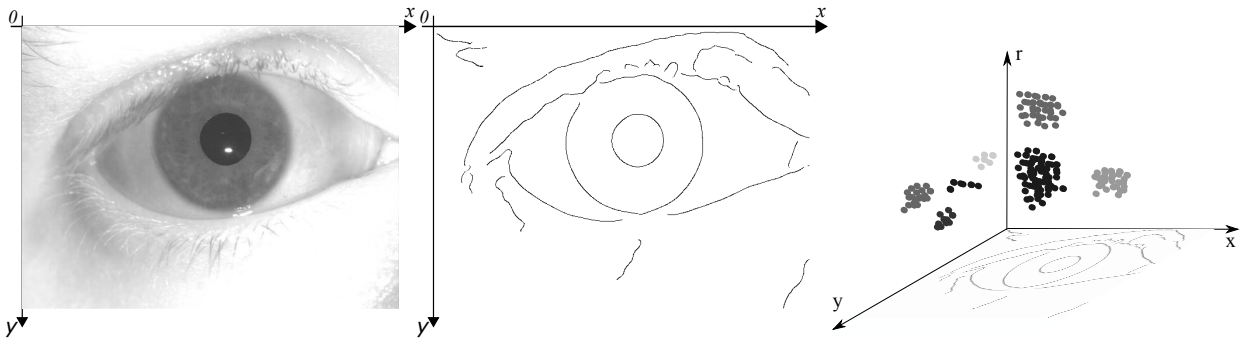


Рис. 2 Иллюстрация идеи метода парных градиентов



(*а*) Входное изображение глаза (*б*) Множество граничных точек (*в*) Пространство параметров

Рис. 3 Модификация преобразования Хафа с коэффициентами подобия

предположительно лежащие на окружности. На втором шаге в двумерном пространстве параметров $\mathbf{Q}$ методом Хафа осуществляется голосование с весовыми коэффициентами для определения центра наиболее выраженной предполагаемой окружности, затем при помощи одномерного аккумулятора определяется ее радиус. На последнем шаге в зависимости от того, какой из границ радужки соответствует найденная окружность, осуществляется поиск второй границы либо внутри данной окружности, либо вне ее.

5.1 Шаг 1: Применение метода парных градиентов

После применения оператора выделения границ из точек изображения формируется множество граничных точек (рис. 3, *б*), которое можно также представить в виде списка. Список граничных точек хранится в массиве $\mathbf{G} = \{x, y, |\mathbf{g}(x, y)|, \varphi\}$. Методом парных градиентов среди граничных точек осуществляется поиск точек интереса.

В отсортированном по значению φ массиве поиск точек с противоположно направленными градиентами можно осуществить за $O(N)$. Воспользуемся предположением о том, что направления градиентов яркости на изображениях глаза распределены практически равномерно на интервале $[-\pi; \pi]$. Для уменьшения сложности поиска пар определяется номер j первого элемента с углом $\varphi_j > 0$, поскольку первому элементу отсортированного массива соответствует значение $\varphi_0 \approx -\pi$. При однократном проходе по массиву для i-й точки с вектором градиента $\boldsymbol{g}_i$ точки с приблизительно антиколлинеарным градиентом будут лежать в некоторой δ-окрестности $(i+j)$-го элемента, т. е. среди элементов с индексами от $i + j - \delta$ до $i + j + \delta$, δ определяется погрешностью для угла φ.

Следуя данным предположениям, из числа граничных точек отбираются пары $\{\boldsymbol{q}_1, \boldsymbol{q}_2\}_i$, где $\boldsymbol{q}_1 = (x_1, y_1)^\intercal$ и $\boldsymbol{q}_2 = (x_2, y_2)^\intercal$. Далее, для каждой i-й пары точек, предположительно лежащей на некоторой окружности $\tilde{O}_i$, определяются параметры этой окружности. При отборе пар согласно условиям (2) и (3) используется порог $T_\varphi = 0{,}984$, что соответствует погрешности в 10°.

В итоге из точек, принадлежащих отобранным парам, формируется множество точек интереса $\mathbf{P} = \{\boldsymbol{q}_i, \boldsymbol{p}_i\}_{i=1}^m$.

5.2 Шаг 2: Поиск наиболее выраженной границы

Если на изображении присутствует окружность O, то в пространстве параметров множеству точек ее границы будет соответствовать единственная точка

$$\boldsymbol{p}_1^o = (x_1^*, y_1^*, r_1^*)^\intercal,$$

где (x_1^*, y_1^*) — координаты ее центра, а r_1^* — радиус. Второй окружности будет соответствовать вторая четко выделенная точка $\boldsymbol{p}_2^o = (x_2^*, y_2^*, r_2^*)^\intercal$ в пространстве параметров. При

искажениях формы окружностей соответствующие им точки в пространстве параметров будут «размыты», т. е. каждой окружности O_i будет соответствовать не четко выделенная точка $\boldsymbol{p}_i^o = (x_i^*, y_i^*, r_i^*)^\intercal, i \in \{1, 2\}$, а множество точек небольшого диаметра по сравнению с расстояниями до другой группы, отвечающей, соответственно, другой окружности. Если на изображении присутствуют шумы, в пространстве параметров появятся «побочные» точки (см. рис. 3, *в*), в том числе искажающие сгруппированные точки, отвечающие параметрам детектируемых окружностей. Таким образом, чем меньше расстояние между точками данного подмножества элементов пространства параметров, тем больше вероятность того, что данное подмножество отвечает некоторой окружности на изображении.

Рассмотрим две граничные точки $\{\boldsymbol{q}_i, \boldsymbol{p}_i\}$ и $\{\boldsymbol{q}_j, \boldsymbol{p}_j\}$. Введем для рассматриваемой пары точек весовой коэффициент качества как

$$w_{ij} = \begin{cases} \dfrac{1}{C}\exp\left(\dfrac{-||\boldsymbol{p_i} - \boldsymbol{p_j}||_2^2}{r_i^2}\right), & \text{если } \frac{||\boldsymbol{p_i} - \boldsymbol{p_j}||_2}{r_i} < t_{\text{lc}}; \\ 0 & \text{иначе,} \end{cases}$$

где C — нормировочный коэффициент, а t_{lc} — постоянная, отвечающие за допустимые искажения формы окружностей. При расстоянии $\|\boldsymbol{p_i} - \boldsymbol{p_j}\| > t_{\text{lc}}$ вклад j-й точки в ячейку аккумулятора, отвечающую параметрам $\boldsymbol{p_i}$, будет нулевым. Это соответствует предположению, что при больших расстояниях между параметрами граничных точек вероятность их принадлежности одной и той же окружности минимальна.

При данном фиксированном i и при $j \in \{1, .., m\}, j \neq i$, рассчитываются весовые коэффициенты w_{ij}, суммируются со значением аккумулятора $\mathbf{Q}$ в точке $(x_i, y_i)^\intercal$

$$Q(x_i, y_i) = \sum_{j=1, j\neq i}^{m} w_{ij},$$

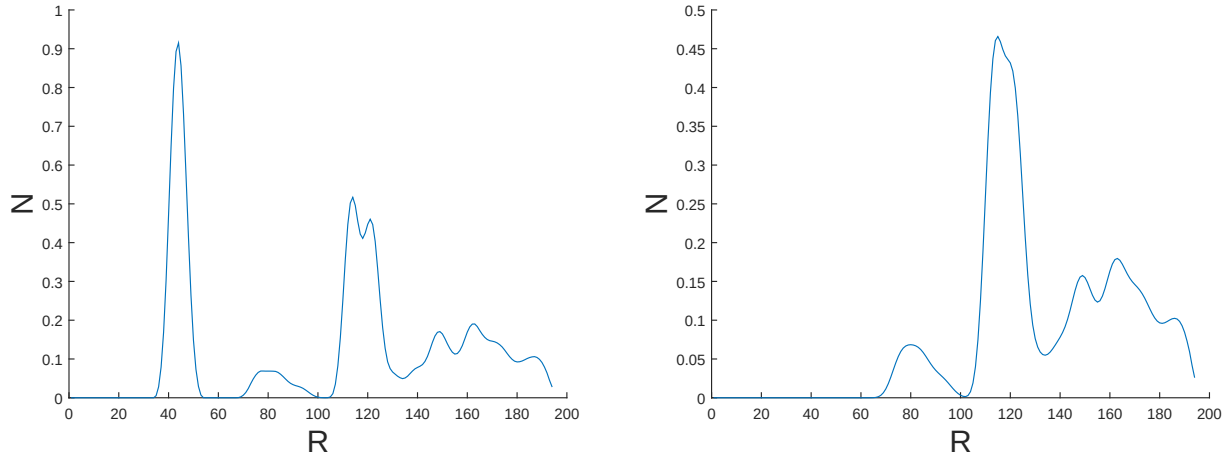
и значение i увеличивается на единицу, т. е. алгоритм переходит к рассмотрению $(i+1)$-й точки. Таким образом, все точки интереса, кроме i-й вносят свою поправку в значение элемента аккумулятора, соответствующего i-й точке, т. е. происходит парное голосование. Два локальных максимума (чаще один, но более размытый) двумерного массива-аккумулятора соответствуют наиболее вероятным положениям центров зрачка и радужки. Наиболее выраженный максимум $(x_1^*, y_1^*) = \underset{(x,y)}{\text{argmax}}\mathbf{Q}(x, y)$ соответствует центру наиболее выраженной границы радужки. Для определения радиуса строится гистограмма $H(r)$ расстояний от $q_1^* = (x_1^*, y_1^*)$ до граничных точек из множества $\mathbf{G}$:

$$H(r) = \left|\{q : q = (x, y) \in \mathbf{G}, ||q - q_1^*|| \in (r - 0{,}5, r + 0{,}5)\}\right|.$$

Для устранения шума каждый столбец гистограммы нормируется на его номер, т. е. на радиус

$$\forall r \to H(r) = \frac{H(r)}{r},$$

и для полученной гистограммы применяется сглаживание при помощи одномерного приближения гауссиана $\exp(-x^2/(2\sigma^2))$ со среднеквадратичным отклонением $\sigma = 10{,}0$. Максимум гистограммы соответствует искомому радиусу r_1^* (рис. 4).



(*а*) Исходная гистограмма расстояний

(*б*) Гистограмма расстояний с исключенными столбцами

Рис. 4 Вид гистограмм расстояний от найденных центров до граничных точек

5.3 Шаг 3: Поиск второй границы радужки

Для поиска второй границы используются предельные соотношения между радиусами радужной оболочки и зрачка, полученные на основании статистических данных [4]:

$$r_{\mathrm{P}} > \frac{1}{7}\, r_{\mathrm{I}}\,; \tag{4}$$

$$r_{\mathrm{P}} < \frac{3}{4}\, r_{\mathrm{I}}\,; \tag{5}$$

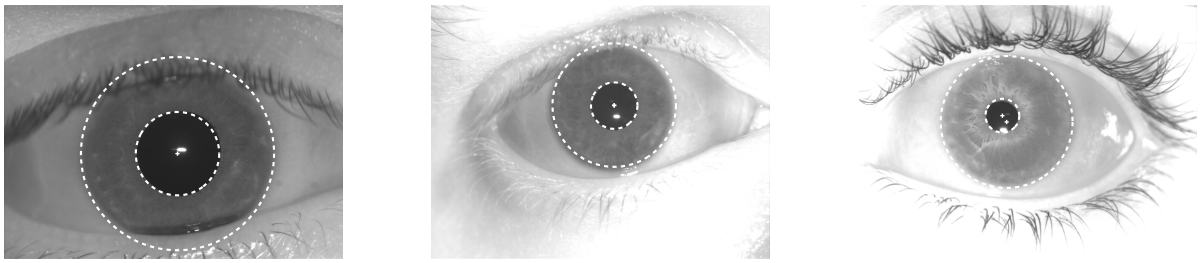
$$r_{\mathrm{P}} > \sqrt{(x_{\mathrm{I}} - x_{\mathrm{P}})^2 + (y_{\mathrm{I}} - y_{\mathrm{P}})^2}\,. \tag{6}$$

где $(x_{\mathrm{I}}, y_{\mathrm{I}})$ — координаты центра внешней границы радужки; r_{I} — ее радиус; $(x_{\mathrm{I}}, y_{\mathrm{I}})$ и r_{I} — аналогичные параметры внешней границы зрачка. Неравенство (4) означает, что радиус радужной оболочки не может превосходить радиус зрачка более чем в 7 раз. Неравенство (5) вводит ограничение с другой стороны: радиус зрачка не может достигать 75% радиуса радужки. Неравенство (6) утверждает, что центр радужки лежит внутри зрачка.

Таким образом, для учета случая неконцентрических границ из массива $\mathbf{P}$ исключаются все точки, соответствующие гипотетическим окружностям с центрами вне найденной, и процедура парного голосования повторяется снова для полученного массива. Для найденного центра (x_2^*, y_2^*) строится гистограмма расстояний до граничных точек $\tilde{H}(r)$, значения которой для столбцов $r \in [0; (1/7)r_1^*] \cup [(3/4)r_1^*; (4/3)r_1^*]$ зануляются в соответствии с условиями (4)–(6), чтобы исключить возможность повторного детектирования уже найденной границы. Полученная гистограмма нормируется на r и сглаживается гауссианом. Аналогично максимум $\tilde{H}(r)$ соответствует искомому радиусу r_2^* (рис. 5).

6 Вычислительный эксперимент

Целью вычислительного эксперимента является проверка работы алгоритма на реальных изображениях глаз. Для тестирования предлагаемого алгоритма использовались изображения глаз разрешением 640×480 точек из базы изображений радужки CASIA-2 [9]



(*а*) (*б*) (*в*)

Рис. 5 Примеры работы алгоритма

в количестве 2335 шт. Для каждого входного изображения экспертом были определены точные координаты центров зрачка и радужной оболочки, а также их соответствующие радиусы, и помещены в файл разметки. В эксперименте использовался персональный компьютер с процессором Inter Core i5-2450M с частотой 2,5 ГГц, 4 ГБ оперативной памяти.

Для каждого изображения по данным, размеченным экспертом, были рассчитаны величины абсолютной и относительной ошибки. Результаты эксперимента для различных величин константы t_{lc} представлены в таблице.

Распределение относительной ошибки определения границ радужки в зависимости от параметра t_{lc}

Суммарная относительная ошибка				
t_{lc}	$e < 5\%$	$e < 10\%$	$e < 15\%$	$e < 20\%$
0,01	34,1%	58,4%	70,8%	75,2%
0,02	35,0%	59,5%	72,9%	75,8%
0,03	35,3%	59,7%	73,1%	76,3%
0,04	37,5%	66,1%	73,4%	76,8%
0,05	38,1%	67,3%	70,3%	74,6%
Суммарная относительная ошибка определения центров				
t_{lc}	$e_{\text{c}} < 5\%$	$e_{\text{c}} < 10\%$	$e_{\text{c}} < 15\%$	$e_{\text{c}} < 20\%$
0,01	69,7%	77,3%	85,1%	89,9%
0,02	72,5%	79,9%	86,1%	90,3%
0,03	74,1%	82,0%	86,7%	90,8%
0,04	74,3%	82,9%	87,0%	91,0%
0,05	73,3%	80,7%	84,1%	90,3%

На основании полученных результатов можно сделать вывод, что оптимальным значением t_{lc} является 0,04. При данном значении было проведено сравнение данного алгоритма с наиболее близким к нему методом [4], также основанным на преобразовании Хафа, в котором поиск центра глаза осуществлялся при помощи голосования вдоль направления антиградиентов яркости в граничных точках (рис. 6).

При проведении вычислительного эксперимента были выявлены некоторые недостатки предлагаемого метода. Во-первых, при отборе пар точек, удовлетворяющих условиям (1)–(3), возникают пары, одна из точек которых принадлежит внешней границе радужки, а вторая — внутренней. Таким образом, предполагаемый центр смещается от действи-

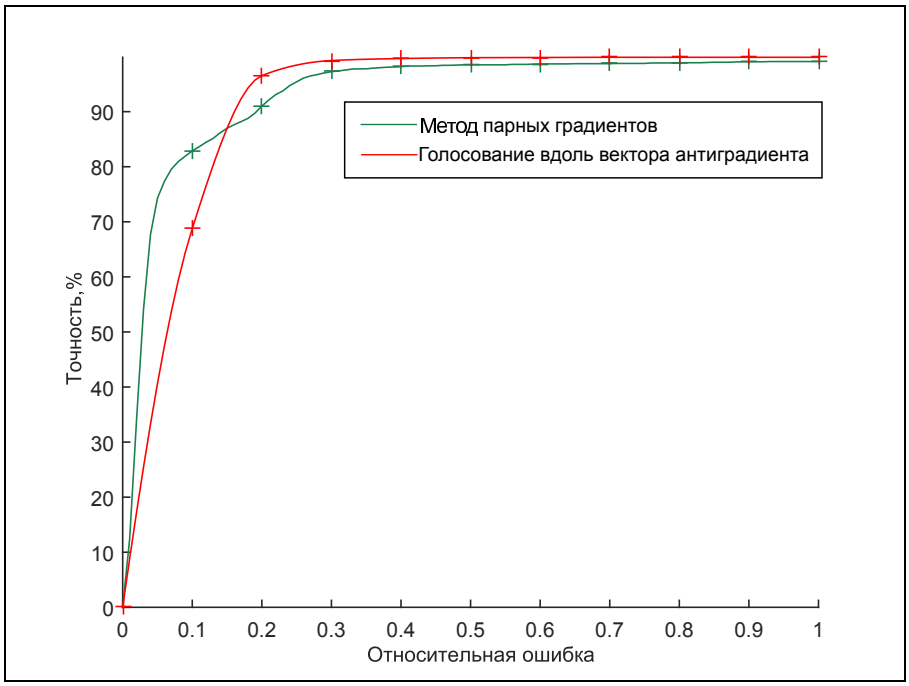
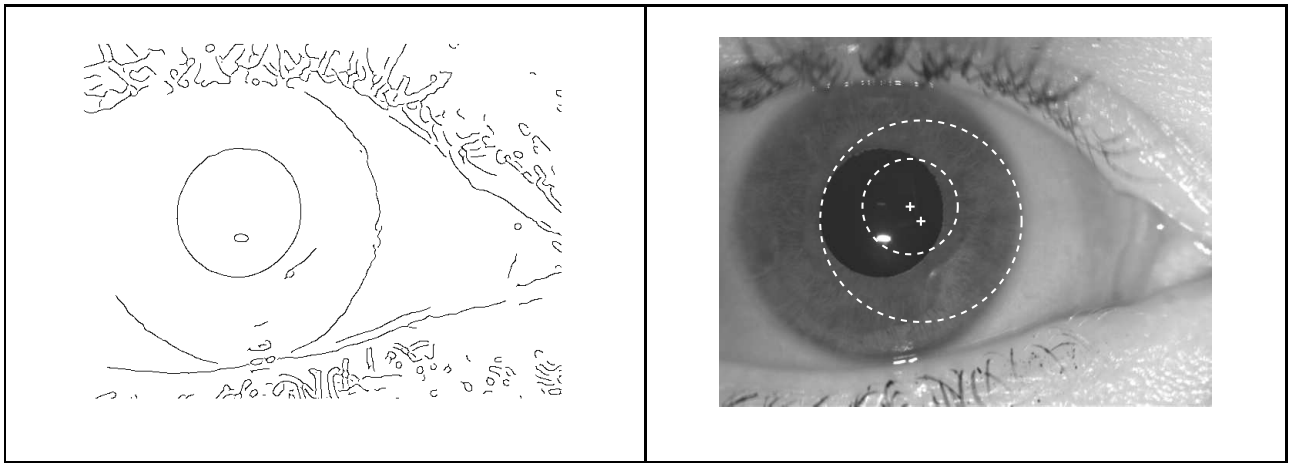


Рис. 6 Распределение относительной ошибки определения центров

тельного центра глаза, и в процедуре голосования такие точки порой порождают «побочный» максимум, в частности, если границы радужной оболочки слабо выделены. Пример порождения побочного максимума приведен на рис. 7. Во-вторых, некорректная работа алгоритма возможна в случае низкоконтрастных изображений и/или сильно прикрытых веками глаз (рис. 8). Наконец, сам метод выделения границ Кэнни не всегда точно выделяет границы радужной оболочки, особенно в случае отсутствия контраста между радужкой и белком глазного яблока. Уменьшение же величины данного порога приводит к возрастанию времени парного голосования и к снижению точности, так как на изображениях помимо ресниц и бровей проявляются более мелкие побочные детали и соответствующие им точки интереса, влияющие на распределение голосов в аккумуляторе.

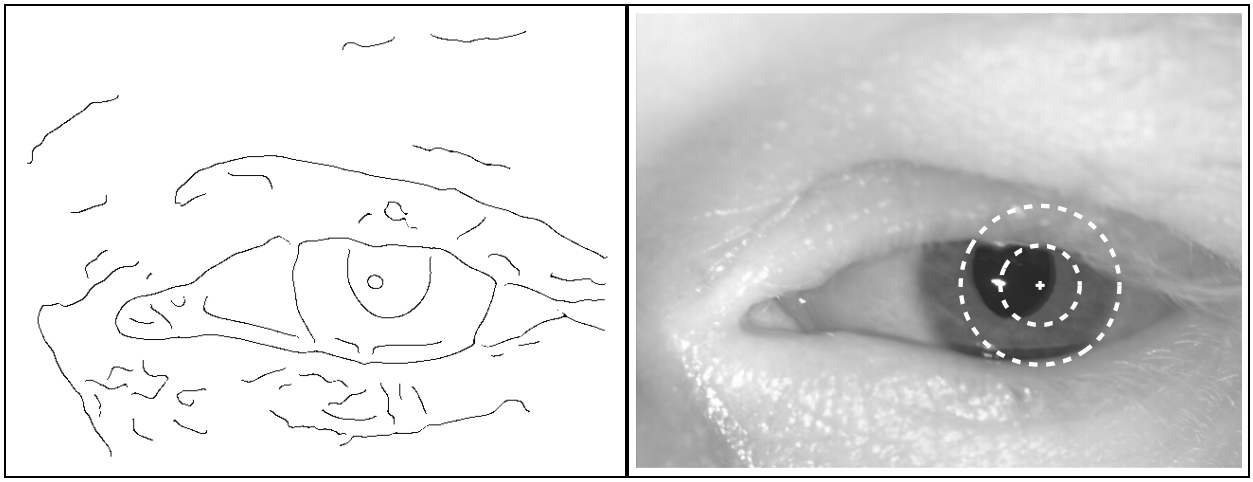
7 Заключение

Предложен алгоритм поиска аппроксимирующих окружностей для границ радужной оболочки. Проведен вычислительный эксперимент, проверяющий работоспособность алгоритма. Результаты представлены в виде таблицы. Данный метод позволяет значительно сократить перебор граничных точек при определении центра методом Хафа и учесть качество параметров с помощью весовых коэффициентов при голосовании в аккумуляторе. Однако данный метод не всегда корректно работает, поскольку условию антиколлинеарности векторов градиентов яркости, используемому при отборе точек и определению параметров гипотетических окружностей, удовлетворяют также и пары, принадлежащие разным границам радужной оболочки, что вносит существенную ошибку в работу алгоритма в случае нечетких границ. Данный метод может быть использован при первичной сегментации радужки для последующего уточнения ее границ, однако применение метода требует контроля и коррекции результатов.



(*а*) Множество граничных точек (*б*) Результат работы алгоритма

Рис. 7 Нечеткая внешняя граница радужки



(*а*) Множество граничных точек (*б*) Результат работы алгоритма

Рис. 8 Недостаточно раскрытые веки

Литература

[1] Duda R. O., Hart P. E. Use of the hough transformation to detect lines and curves in pictures // Comm. ACM, 1972. Vol. 15. No. 1. P. 11–15. doi: http://dx.doi.org/10.1145/361237.361242

[2] Cauchie J., Fiolet V., Villers D. Optimization of an hough transform algorithm for the search of a center // Pattern Recogn., 2008. Vol. 41. No. 2. P. 567–574. doi: http://dx.doi.org/10.1016/j.patcog.2007.07.001

[3] Valenti R., Gevers T. Accurate eye center location through invariant isocentric patterns // IEEE Trans. Pattern Anal. Machine Intelligence Arch., 2012. Vol. 34. No. 9. P. 1785–1798. doi: http://dx.doi.org/10.1109/TPAMI.2011.251

[4] *Ганькин К. А., Гнеушев А. Н., Матвеев И. А.* Сегментация изображения радужки глаза, основанная на приближенных методах с последующими уточнениями // Известия РАН.

Теория и системы управления, 2014. № 2. С. 80–94. doi: http://dx.doi.org/10.1134/S1064230714020099

[5] Xu L., Oja E., Kultanen P. A new curve detection method: Randomized hough transform // Pattern Recogn. Lett., 1990. Vol. 11. No. 5. P. 331–338. doi: http://dx.doi.org/10.1016/0167-8655(90)90042-Z

[6] Chen T.-C., Chung K.-L. An efficient randomized algorithm for detecting circles // Computer Vision Image Understanding, 2001. Vol. 83. No. 2. P. 172–191. doi: http://dx.doi.org/10.1006/cviu.2001.0923

[7] Pan L., Chu W.-S., Saragih J. M. Fast and robust circular object detection with probabilistic pairwise voting // IEEE Signal Proc. Lett., 2011. Vol. 18. No. 11.

[8] Canny J. A computational approach to edge detection // IEEE Trans. Pattern Anal. Machine Intelligence, 1986. Vol. 8. No. 6. P. 679–698. doi: http://dx.doi.org/10.1109/TPAMI.1986.4767851

[9] Chinese Academy of Sciences Institute of Automation Iris Image Database. Ver. 2.0.

Поступила в редакцию 15.06.15

Iris border detection using a method of paired gradients*

Y. S. Efimov and I. A. Matveev

yuri.efimov@phystech.edu, matveev@ccas.ru

Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Russia

Dorodnicyn Computer Centre of the Russian Academy of Sciences, 40 Vavilova st., Moscow, Russia

Circular object detection is one of the challenging problems of modern computer vision systems. In this study, to search for circular representations of inner and outer boundaries of the iris, a method of paired gradients is used which is a modification of the Hough methodology. Image is processed with Canny filter and from the resulting boundaries, pairs of pixels are selected which have high probability to belong to one circle. Selection criteria and probability coefficients of likelihood are introduced for reduction of number of these pairs. The Hough transform uses two accumulators: the two-dimensional isomorphic to the original image in which voting is done by centers of segments defined by pixel pairs and one-dimensional histogram of the diameters where lengths of these segments are collected. Computational experiment is performed to check the efficiency of the algorithm on data from the public iris image databases and to compare the proposed method of paired gradients to the resembling antigradient voting method, which is also based on the Hough methodology and used for the eye center search. Drawbacks of the algorithm that may cause incorrect handling of some of the input images are identified. Further analysis of the proposed algorithm and increasing its stability are required.

Keywords: *paired gradients method; computer vision; circular object detection; circular Hough transform*

DOI: 10.21469/22233792.1.14.08

References

[1] Duda, R. O., and P. E. Hart. 1972. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM* 15(1):11–15. doi: http://dx.doi.org/10.1145/361237.361242

*This work was done under financial support of the Russian Foundation for Basic Research (grant 16-07-01171)

[2] Cauchie, J., V. Fiolet, and D. Villers. 2008. Optimization of an hough transform algorithm for the search of a center. *Pattern Recogn.* 41(2):567–574. doi: http://dx.doi.org/10.1016/j.patcog.2007.07.001

[3] Valenti, R., and T. Gevers. 2012. Accurate eye center location through invariant isocentric patterns. *IEEE Trans. Pattern Anal. Machine Intelligence Arch.* 34(9):1785–1798. doi: http://dx.doi.org/10.1109/TPAMI.2011.251

[4] Gankin, K., A. Gneushev, and I. Matveev. 2014. Iris image segmentation based on approximate methods with subsequent refinements. *J. Computer Syst. Sci. Int.* 53(2):224–238. doi: http://dx.doi.org/10.1134/S1064230714020099

[5] Xu, L., E. Oja, and P. Kultanen. 1990. A new curve detection method: Randomized hough transform. *Pattern Recogn. Lett.* 11(5): 331–338. doi: http://dx.doi.org/10.1016/0167-8655(90)90042-Z

[6] Chen, T.-C., and K.-L. Chung. 2001. An efficient randomized algorithm for detecting circles. *Computer Vision Image Understanding* 83(2):172–191. doi: http://dx.doi.org/10.1006/cviu.2001.0923

[7] Pan, L., W.-S. Chu, and J. M. Saragih. 2011. Fast and robust circular object detection with probabilistic pairwise voting. *IEEE Signal Proc. Lett.* 18(11).

[8] Canny, J. 1986. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Machine Intelligence* 8(6):679–698. doi: http://dx.doi.org/10.1109/TPAMI.1986.4767851

[9] Chinese Academy of Sciences Institute of Automation Iris Image Database. Ver. 2.0.

Received June 15, 2015