О некоторых вопросах анализа пучков временных рядов*

***Н. В. Филипенков*[1], *М. А. Петрова*[2]**
[1]n.filipenkov@mail.ru, [2]marina_petrova@mail.ru
[1]САС институт, Москва, ул. Станиславского, 21-1; [2]НИЯУ МИФИ, Москва, Каширское ш., 31

В настоящей работе рассматривается разрабатываемый авторами подход к поиску закономерностей в пучках нестационарных k-значных временных рядов. Этот подход позволяет выявлять закономерности, которые подвергаются «плавным» структурным изменениям с течением времени.

Настоящая работа посвящена описанию результатов апробации разрабатываемого подхода на модельных и реальных задачах. Испытания на модельных задачах показали, что подход позволяет эффективно находить заложенные закономерности при достаточно высоком уровне шума. Эксперименты на модельных пучках временных рядов показали, что использование меры сходства закономерностей в функционале качества существенно повышает точность прогнозирования. В рамках экспериментов был получен диапазон весов, при котором достигается максимальное качество распознавания. Анализ реальных временных рядов с применением разрабатываемого алгоритма свидетельствовал об эффективности алгоритма при краткосрочном прогнозировании. Вместе с тем алгоритм решает и задачу интеллектуального анализа данных, предлагая закономерности, описывающие взаимосвязь одномерных временных рядов.

Таким образом, апробация разрабатываемого подхода к прогнозированию процессов с плавно меняющимися закономерностями на модельных и реальных данных позволяет судить о достаточной эффективности разрабатываемых авторами алгоритмов при анализе пучков временных рядов с плавно меняющимися закономерностями.

Ключевые слова: временные ряды; интеллектуальный анализ данных; мера сходства закономерностей; вычислительный алгоритм

On the analysis of multidimensional time series*

***N. V. Filipenkov*[1], *M. A. Petrova*[2]**
[1]SAS Institute, 21-1 Stanislavskogo Str., Moscow; [2]MEPhI, 31 Kashirskoye Sh., Moscow

In this paper, an approach for discovering rules in nonstationary finite-valued multidimensional time series is discussed. It allows one to discover rules that slightly change their structure over time. A measure of rule similarity is introduced and studied as a weight on the graph of rules.

This paper focuses on the results of the application of the discussed algorithm to the modeled and real problems. The experiments on the model problems show that the approach allows to mine the hidden rules efficiently even under high noise conditions. The experiments on the modeled multidimensional time series show that using the rules similarity measure in the quality function significantly increases the forecast accuracy. During the experiments, the

*Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-00293.

weight range for maximum data mining quality was identified. The analysis of real time series based on the discussed approach show the algorithm's efficiency for short-term forecasting. In addition to that, the algorithm solves the data mining problem while finding the rules describing the interconnection of the univariate time series.

The application of the discussed approach for forecasting the processes with slightly changing rules on modeled and real data shows the efficiency of the developed algorithms for the analysis of multidimensional time series with slightly changing rules.

Keywords: time series; data mining; similarity measure; algorithm

Введение

В настоящее время анализ временных рядов является крайне актуальной задачей в различных сферах деятельности человека: медицине, экономике, физике, кибернетике. При этом часто возникает необходимость исследования сразу нескольких процессов или показателей одного процесса в их взаимосвязи и взаимовлиянии — изучения пучков временных рядов. Пучки временных рядов могут, например, описывать процессы жизнедеятельности человека, стоимость акций на бирже, курсы валют и т. д. Пучок временных рядов, учитывая множество характеристик явления, позволяет описать процесс или систему процессов наиболее полно, что, в свою очередь, позволяет сделать более точный прогноз. Возможность системного анализа процессов, их более точного описания определила высокий интерес исследователей к изучению пучков временных рядов [1 — 12].

Изучение пучков временных рядов подразумевает не только прогнозирование значений рядов, но предполагает решение задачи в рамках области интеллектуального анализа данных. Это означает, что необходимо выявить и описать закономерности, определяющие поведение временных рядов. Найденные закономерности могут быть представлены в виде уравнений [2, 4, 5], ассоциативных [6, 7], «эпизодических» [8] или прочих правил [10, 11, 9].

В большинстве реальных задач измерения проводятся в дискретные моменты времени, поэтому во многих работах рассматриваются дискретные временные ряды. При этом в работах, посвященных поиску правил [6, 10, 11, 9], рассматриваются пучки, где значениями временных рядов являются элементы некоторого конечного алфавита. Для поиска правил в «непрерывных» временных рядах используются методы дискретизации или символьного представления [10, 12].

Пучок временных рядов отражает характеристики явления во времени, но само явление может меняться с течением времени. Нестационарными в общем смысле называются временные ряды, свойства которых непостоянны во времени. Во многих областях такие ряды составляют большинство, так как почти все явления под воздействием различных факторов претерпевают изменения. Для анализа нестационарных временных рядов был предложен целый ряд адаптивных методов: экспоненциальное сглаживание и его модификации [13, 14], модели семейства ARIMA [2], модели семейства ARCH [4, 15], множественная регрессия [5], модели, основанные на использовании спектральных характеристик рядов [16, 17].

Основные определения *Пучком временных рядов* $\mathfrak{S}$ называется совокупность взаимосвязанных временных рядов S_i, $i \in \{1, 2 \ldots, N\}$. Каждый ряд S_i представляет собой последовательность значений конечнозначной логики E_{k_i}. Каждому элементу ряда соответствует некоторый момент времени, и эти моменты времени для всех рядов одинаковы. Поэтому одинакова и длина всех рядов, которая обозначается через T. Таким образом,

пучок временных рядов $\mathfrak{S}$ есть матрица размера $N \times T$, где элемент i-й строки принадлежит множеству E_{k_i}. Значения ряда $S_i, i \in \{1, 2 \ldots, N\}$ в момент времени $t \in \{1, 2 \ldots, T\}$ обозначим через $a(i,t)$ или $a_{i,t}$.

Маской ω на прямоугольнике $N \times \Delta$ назовем булеву матрицу размера $N \times \Delta$ (здесь параметр Δ определяет максимальный отступ по времени). Число единиц в маске ω будем называть *мощностью* маски и обозначать через $||\omega||$. Элемент маски, находящийся в i-й строке и j-м столбце, будем обозначать через $\omega(i,j)$ или $\omega_{i,j}$. *Закономерностью* R назовем набор (p, ω, f) с такими особенностями:

1) число $p \in \{1, 2, \ldots, N\}$ указывает на целевой ряд (ряд, значения которого определяются закономерностью R);
2) маска ω указывает на значения рядов, являющиеся аргументами функции f;
3) частично определенная функция f задает зависимость значений целевого ряда от переменных, на которые указывает маска ω.

$$f\colon E_{k_{i_1}} \times \ldots \times E_{k_{i_{||\omega||}}} \to E_{k_p} \cup \{\lambda\},$$

где $\omega(i_1, j_1), \ldots, \omega(i_{||\omega||}, j_{||\omega||})$ — единичные элементы матрицы ω; p — номер целевого ряда; символ λ обозначает, что f не определена на соответствующем наборе значений переменных.

Если значения всех рядов представляют собой числа k-значной логики ($E_{k_1} = \cdots = E_{k_N} = E_k$), то функция f принадлежит множеству P_k^* всех частично определенных функций k-значной логики.

Задача состоит в поиске закономерностей и прогнозировании. Найденные закономерности позволяют прогнозировать значения целевого ряда, делать выводы о характере зависимостей между рядами, моделировать целевой ряд или весь пучок временных рядов.

В предыдущих работах авторов [18, 19] рассматриваются алгоритм поиска постоянных закономерностей, по аналогии с [6, 7, 20] вводятся понятия достоверности $\mathrm{Conf}\,(R, \mathfrak{S})$ и поддержки $\mathrm{Supp}\,(R, \mathfrak{S})$ закономерности R на пучке $\mathfrak{S}$, оценивается необходимая длина пучка временных рядов, вводится понятие системы закономерностей, исследуется понятие её полноты, объясняется, каким образом подход позволяет применять к построенным закономерностям конструкции алгебраического подхода, предложенные в работах Ю. И. Журавлева [21] и К. В. Рудакова [22]. В работах производится построение меры сходства закономерностей, определяется понятие изменяющейся закономерности.

Изменяющейся закономерностью $\tilde{R}$ для последовательности отрезков $\mathfrak{S}^1, \ldots, \mathfrak{S}^m$ на пучке временных рядов $\mathfrak{S}$ называется система закономерностей $R^1, \ldots, R^m$, где каждая закономерность взаимно однозначно соответствует некоторому отрезку $\mathfrak{S}^i$, $i = 1, 2, \ldots, m$. Вообще говоря, отрезки могут пересекаться между собой. Будем называть стационарные закономерности $R^1, \ldots, R^m$ *шагами*, которые *составляют* изменяющуюся закономерность $\tilde{R}$.

В работах [18, 19] описывается поиск плавно меняющихся закономерностей на графе (рис. 1).

Вершинами графа являются стационарные закономерности, найденные на каждом из отрезков, а также две дополнительные вершины: beg и end. С каждой вершиной ассоциированы показатели качества закономерности. Дугами на графе связаны закономерности соседних отрезков, что отражает факт возможного «превращения» одной закономерности в другую. С каждой дугой ассоциирован вес — мера сходства соответствующих закономерностей. Веса дуг, соединяющие закономерности крайних отрезков с вершинами *beg* и *end*, полагаются равными нулю.

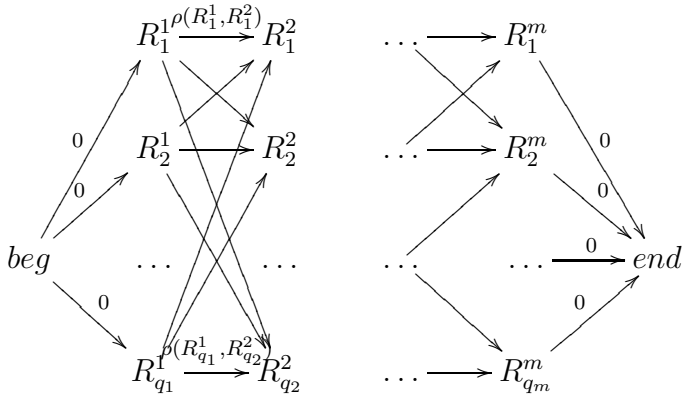

Рис. 1. Граф закономерностей

Задача выделения наилучшей изменяющейся закономерности состоит в поиске пути между вершинами beg и end на ориентированном графе, который максимизирует показатели качества закономерностей вершин, входящих в него, и минимизирует суммарный вес ребер.

Эта задача сводится к стандартной задаче поиска кратчайшего пути на графе, если использовать в качестве веса вершины величину $(1-Q_{\text{step}})$, где Q_{step} — *функционал качества шага изменяющейся закономерности* $\tilde{R}$, который задается следующим образом:

$$
\begin{aligned}
&Q_{\text{step}}(R_i^j, R_l^{j+1}) = w_{\text{conf}}\text{Conf}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right) + w_{\text{supp}}\text{Supp}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right) + w_{\text{similarity}}(1-\rho(R_i^j, R_l^{j+1}));\\
&Q_{\text{step}}(\text{beg}, R_i^j) = 0;\\
&Q_{\text{step}}(R_i^j, \text{end}) = w_{\text{conf}}\text{Conf}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right) + w_{\text{supp}}\text{Supp}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right),\\
&\qquad\qquad j = 1, 2, \ldots, m-1;\ \ i = 1, 2, \ldots, q_j;\ \ l = 1, 2, \ldots, q_{j+1}.
\end{aligned}
$$

Здесь $\text{Conf}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right)$ и $\text{Supp}\left(R_i^j, \mathfrak{S}_{\text{valid}}^j\right)$ — показатели качества закономерности; $\rho(R_i^j, R_l^{j+1})$ — мера сходства закономерностей, детально описанная в работе [19]. Веса w_{conf}, w_{supp} и $w_{\text{similarity}}$ функционала качества шага удовлетворяют следующим условиям:

$$
\begin{aligned}
0 \leqslant w_{\text{conf}} &\leqslant 1;\\
0 \leqslant w_{\text{supp}} &\leqslant 1;\\
0 \leqslant w_{\text{similarity}} &\leqslant 1;\\
w_{\text{conf}} + w_{\text{supp}} + w_{\text{similarity}} &= 1.
\end{aligned}
$$

В работе [19] доказывается, что для произвольных закономерностей R_1, R_2 верно неравенство:

$$0 \leqslant Q_{\text{step}}(R_1, R_2) \leqslant 1.$$

Таким образом, вес вершины $(1 - Q_{\text{step}})$ является неотрицательным и для решения задачи поиска кратчайшего пути на графе удобно использовать стандартные алгоритмы:

1) алгоритм Дейкстры [23] со сложностью $\underline{\underline{O}}(n^2)$ или его реализацию с фибоначчиевой кучей со сложностью $\underline{\underline{O}}(n\log n)$, где n — число вершин графа;
2) алгоритм поиска кратчайшего расстояния в топологически отсортированном графе [23] со сложностью $\underline{\underline{O}}(n^2)$, где n — число вершин графа.

Изменяющуюся закономерность $\tilde{R}$ будем называть *плавно меняющейся*, если она составлена из закономерностей, лежащих на кратчайшем пути из вершины beg в вершину end и выполнено неравенство $w_{similarity} > 0$ для веса меры сходства закономерностей функционала качества шага.

Примеры решения модельных задач. С целью испытания предложенного подхода для решения практических задач был подготовлен экспериментальный стенд. Стенд позволяет импортировать и генерировать временные ряды, проводить поиск стационарных и изменяющихся закономерностей, а также решать задачи прогнозирования. С использованием стенда было проведено несколько серий экспериментов. Обозначения для параметров экспериментов представлены в табл. 1 и 2.

Таблица 1. Обозначения параметров

Обозначение	**Параметр**
Параметры генерации	
K	Значность пучка временных рядов
N	Количество рядов в пучке
T	Длина рядов
Δ_{gen}	Максимальный отступ по времени
$\|\|\omega_1\|\|$	Мощность маски первой закономерности
p_{gen}	Индекс целевого ряда
m_{gen}	Количество сегментов
ξ_{mask}	Количество изменений маски при переходе к следующему отрезку
π_{mask}	Вероятность каждого изменения маски при переходе к следующему отрезку
ξ_{func}	Доля изменяемых значений функции при переходе к новому отрезку
π_{func}	Вероятность каждого изменения функции при переходе к следующему отрезку
ε	Уровень шума (доля значений целевого ряда, определяемых случайно)
Параметры поиска стационарных закономерностей	
p_{mine}	Индекс целевого ряда
Δ_{mine}	Максимальный отступ по времени
μ	Максимальный вес маски
$\min \text{supp}_{\text{set}}$	Минимальная поддержка набора
Valid	Доля отрезка, которая используется для валидации закономерностей

Проводились две серии экспериментов на модельных рядах с целью выявить условия для наиболее эффективного применения предложенных в [19] алгоритмов интеллектуального анализа временных рядов.

Таблица 2. Обозначения параметров

Обозначение	Параметр
Фильтры базы знаний	
$\mathrm{conf}_{\min}$	Минимальная достоверность на обучении закономерности для включения в базу знаний
$err^{\mathrm{valid}}_{\max}$	Максимальная ошибка на валидации
$\mathrm{supp}_{\min}$	Минимальная поддержка закономерности для включения в базу знаний
Параметры поиска меняющихся закономерностей	
m_{mine}	Количество сегментов
v	Стоимость перемещения аргумента по вертикали (используется в мере сходства масок)
h	Стоимость перемещения аргумента по горизонтали (используется в мере сходства масок)
w_λ	Расстояние до значения λ (используется в мере сходства функций)
$\varkappa_{\mathrm{mask}}$	Вес меры сходства масок (используется в мере сходства закономерностей)
$\varkappa_{\mathrm{func}}$	Вес меры сходства функций (используется в мере сходства закономерностей)
w_{conf}	Вес меры, характеризующей точность закономерности (используется в функционале качества закономерностей)
w_{supp}	Вес достоверности (используется в функционале качества закономерностей)
$w_{\mathrm{similarity}}$	Вес меры сходства закономерностей (используется в функционале качества закономерностей)

В первой серии модельных экспериментов было проведено исследование влияния уровня шума в моделируемых пучках временных рядов на качество распознавания. Для каждого значения уровня шума проводилась серия из 100 экспериментов.

В каждом эксперименте в соответствии с параметрами K (значность), N (количество рядов) и T (длина рядов) генерировался пучок временных рядов. Все ряды, за исключением целевого, генерировались случайным образом при равномерном распределении. Пучок разбивался по времени на m_{gen} равных сегментов.

Затем генерировалась плавно меняющаяся закономерность, состоящая из m_{gen} шагов — стационарных закономерностей. Стационарная закономерность первого отрезка создавалась случайным образом в соответствии с условиями на ширину маски (Δ_{gen}) и количество аргументов ($||\omega_1||$). Каждый следующий шаг плавно меняющейся закономерности был получен из предыдущего путем сдвига ξ_{mask} элементов маски, где каждый сдвиг происходил с вероятностью π_{mask}. Вместе с тем при генерации следующего шага плавно меняющейся закономерности изменялась часть значений функции, определяемая долей ξ_{func} от общего количества наборов, на которых определена функция. Вероятность каждого изменения задавалась долей π_{func}. Значения параметров генерации представлены в табл. 3. Таким

образом, за счет плавных изменений стационарных закономерностей получалась плавно меняющаяся закономерность.

Целевой ряд заполнялся с использованием сгенерированной плавно меняющейся закономерности на основе значений других рядов. При этом каждое значение целевого ряда с вероятностью ε генерировалось случайным образом, а не в соответствии с закономерностью. Таким образом, в целевом временном ряде учитывался заданный уровень шума ε.

После этапа генерации пучка временных рядов в каждом эксперименте производился поиск плавно меняющихся закономерностей в сгенерированном пучке временных рядов. Значения параметров алгоритмов поиска закономерностей представлены в табл. 3 и 4. В соответствии с параметрами функционала качества шага изменяющейся закономерности (w_{conf}, w_{supp} и $w_{\text{similarity}}$) определялась наилучшая изменяющаяся закономерность, которая затем сравнивалась со сгенерированной закономерностью.

Таблица 3. Значения параметров в экспериментах

Параметр	**Серия 1**	**Серия 2**
Параметры генерации		
K	4	4
N	10	10
T	1000	1000
Δ_{gen}	20	20
$\|\|\omega_1\|\|$	3	2
p_{gen}	0	0
m_{gen}	3	3
ξ_{mask}	1	1
π_{mask}	1	1
ξ_{func}	0,03	0,03
π_{func}	1	1
ε	изменяется	изменяется
Параметры поиска стационарных закономерностей		
p_{mine}	0	0
Δ_{mine}	20	20
μ	5	4
$\min \text{supp}_{\text{set}}$	0	0
Valid	20%	20%

Критерии успешного эксперимента были определены следующим образом. Генерируемая стационарная закономерность и найденная стационарная закономерность называются *совпадающими*, если полностью совпадают их маски и доля различных значений функции не превышает 5% от общего числа наборов, на которых определены функции. *Изменяющиеся* закономерности называются *совпадающими*, если совпадают все их соответствующие шаги — стационарные закономерности. Эксперимент признается *успешным*, если найденная изменяющаяся закономерность оказывается совпадающей с генерируемой.

Для каждого значения уровня шума рассчитывалась доля успешных экспериментов по отношению к общему числу экспериментов при данном уровне шума.

Таблица 4. Значения параметров в экспериментах

Параметр	Серия 1	Серия 2
Фильтры базы знаний		
$\text{conf}_{\min}$	0,3	0,3
$\text{err}^{\text{valid}}_{\max}$	1,0	1,0
$\text{supp}_{\min}$	0	0
Параметры поиска меняющихся закономерностей		
m_{mine}	3	3
v	1	1
h	1	1
w_λ	4	4
$\varkappa_{\text{mask}}$	0,5	0,5
$\varkappa_{\text{func}}$	0,5	0,5
w_{conf}	0,5	изменяется
w_{supp}	0	0
$w_{\text{similarity}}$	0,5	изменяется

Результаты моделирования в первой серии экспериментов представлены на рис. 2.

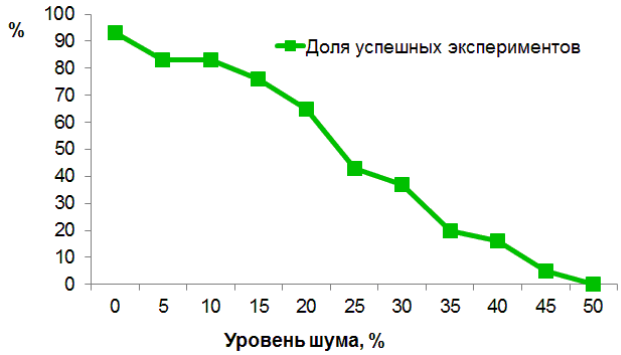

Рис. 2. Качество распознавания при различных весах функционала качества изменяющейся закономерности

Результаты показывают, что качество распознавания линейно убывает при увеличении уровня шума. При этом алгоритм поиска плавно меняющихся закономерностей является достаточно стабильным и проводит вполне эффективный интеллектуальный анализ данных даже для зашумленных пучков временных рядов.

Во второй серии экспериментов исследовалось влияние меры сходства закономерностей на качество распознавания. С этой целью проводился поиск изменяющихся закономерностей для разных комбинаций весов функционала качества Q_{step} шага изменяющейся закономерности. При этом исследование влияния меры сходства закономерностей проводилось для нескольких значений уровня шума.

Эксперименты проводились следующим образом. Был задан набор значений уровня шума ε: 0,2, 0,3 и 0,5. Для каждого значения уровня шума генерировался пучок времен-

ных рядов способом, аналогичным примененному в первой серии экспериментов. Значения параметров генерации представлены в табл. 3.

В сгенерированном пучке временных рядов происходил поиск изменяющихся закономерностей при различных весах функционала качества Q_{step} шага изменяющейся закономерности. Вес поддержки w_{supp} полагался равным нулю, что исключило влияние уровня поддержки на выбор оптимальной изменяющейся закономерности. Вес меры сходства закономерностей $w_{\text{similarity}}$ увеличивался от 0 до 1 с шагом 0,05. Вес достоверности w_{conf} соответственно уменьшался от 1 до 0 с шагом 0,05.

Для каждого значения уровня шума было сгенерировано 100 пучков временных рядов. В каждом пучке временных рядов происходил поиск плавно меняющихся закономерностей при 21-й различной комбинации весов функционала качества Q_{step}. Значения параметров алгоритмов поиска закономерностей представлены в табл. 3 и 4.

Для каждого значения уровня шума и для каждой комбинации весов была рассчитана доля успешных экспериментов. Успешный эксперимент определялся аналогично первой серии экспериментов. Доля определялась по отношению к общему количеству экспериментов для данного значения уровня шума и комбинации весов.

Результаты второй серии экспериментов представлены в табл. 5 и на рис. 3.

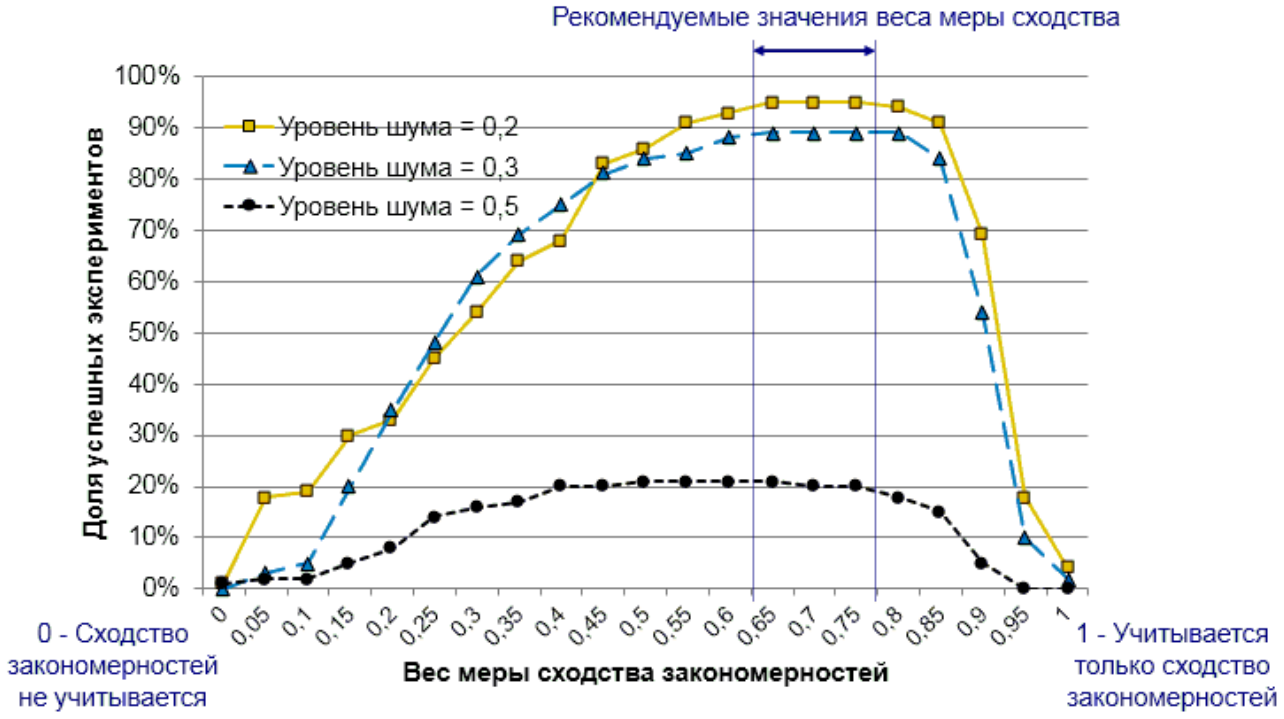

Рис. 3. Качество распознавания при различных весах функционала качества изменяющейся закономерности

Результаты второй серии экспериментов указывают на необходимость использования как меры сходства закономерностей, так и достоверности в функционале качества шага изменяющейся закономерности. При граничных значениях весов доля успешных экспериментов снижается, и, наоборот, она достигает максимального уровня при значениях веса меры сходства закономерностей $w_{\text{similarity}}$ в интервале от 0,65 до 0,8 (соответственно значениях веса достоверности w_{conf} от 0,35 до 0,2).

Комбинация значений весов $w_{\text{conf}} = 1$ и $w_{\text{similarity}} = 0$ соответствует алгоритму, при котором на каждом отрезке выбирается закономерность, обладающая максимальной достоверностью. Объединенные вместе данные закономерности составляют изменяющуюся закономерность. Мера сходства закономерностей в этом случае не используется.

При указанных значениях весов и высоком уровне шума алгоритм поиска изменяющихся закономерностей склонен к «переобучению». Так как мера сходства закономерностей не используется в функционале качества, в изменяющуюся закономерность объединяются «локальные оптимумы» каждого из отрезков. В итоге при любом уровне шума поиск плавно меняющихся закономерностей без использования меры сходства закономерностей приводит к низкому качеству распознавания: не удается правильно определить плавно меняющуюся закономерность.

Теперь рассмотрим случай, когда веса принимают другое пограничное значение: $w_{\text{conf}} = 0$ и $w_{\text{similarity}} = 1$. Тогда единственным критерием для выбора закономерностей является их близость. Если нет ограничения на количество закономерностей, записываемых в базу знаний алгоритмом поиска стационарных закономерностей, то такой выбор весов исключает возможность изменения закономерности. То есть при данном выборе весов оптимальной является любая изменяющаяся закономерность составленная из одинаковых стационарных закономерностей (т. е. фактически стационарная закономерность).

При комбинации весов $w_{\text{conf}} = 0$ и $w_{\text{similarity}} = 1$ бывает удобно упорядочить закономерности по убыванию достоверности и поставить ограничение на количество закономерностей, записываемых в базу знаний алгоритмом поиска стационарных закономерностей. Например, в базу знаний на каждом из отрезков могут записываться только 50 закономерностей, обладающих наилучшей достоверностью. Тогда достоверность будет неявно учитываться при выборе плавно меняющейся закономерности.

Таким образом, результаты второй серии экспериментов показывают, что добавление меры сходства закономерностей в функционал качества позволяет существенно повысить точность распознавания в пучках с плавно меняющимися закономерностями. При этом рекомендуемыми значениями веса меры сходства закономерностей $w_{\text{similarity}}$ являются числа в диапазоне от 0,65 до 0,8.

Примеры решения реальных задач

С целью сравнить предложенный в настоящей работе подход с другими методами была проведена серия экспериментов по краткосрочному прогнозированию временных рядов. Данными послужили курсы акций компаний Adobe, BMC, Business Objects, Cognos, Computer Associate, Novell, Oracle, Peoplesoft, Rational. Рассматривался средний почасовой курс акций в долларах за период с 13 мая 2002 г. по 10 декабря 2004 г. Средний почасовой курс получался как среднее арифметическое из четырех чисел: цены открытия (цены акции в начале часа), верхней цены (максимальной цены акции за час), нижней цены (минимальной цены акции за час), цены закрытия (цены акции в конце часа). Все девять временных рядов рассматривались как единый пучок, так как перечисленные выше компании работают в одной сфере разработки программного обеспечения для предприятий и не исключены взаимосвязи между поведением акций этих компаний.

Для прогнозирования цены акции помимо предложенного в настоящей работе метода применялось экспоненциальное сглаживание с параметром α, принимающим значения 0,1 и 0,3.

В связи с тем, что предложенный в настоящей работе подход предложен для пучков конечнозначных временных рядов, исходные действительные временные ряды были пре-

Таблица 5. Результаты экспериментов при различных весах функционала качества

w_{conf}	$w_{similarity}$	Доля успешных экспериментов, %		
		$\varepsilon = 0,2$	$\varepsilon = 0,3$	$\varepsilon = 0,5$
1,00	0,00	1	0	1
0,95	0,05	18	3	2
0,90	0,10	19	5	2
0,85	0,15	30	20	5
0,80	0,20	33	35	8
0,75	0,25	45	48	14
0,70	0,30	54	61	16
0,65	0,35	64	69	17
0,60	0,40	68	75	20
0,55	0,45	83	81	20
0,50	0,50	86	84	21
0,45	0,55	91	85	21
0,40	0,60	93	88	21
0,35	0,65	95	89	21
0,30	0,70	95	89	20
0,25	0,75	95	89	20
0,20	0,80	94	89	18
0,15	0,85	91	84	15
0,10	0,90	69	54	5
0,05	0,95	18	10	0
0,00	1,00	0	0	0

образованы в четырехзначные. Для каждого из рядов в пучке было произведено два преобразования. Первое преобразование состояло в переходе от исходных значений к разностям. Второе сопоставило каждой разности элемент алфавита $E_4 = \{0, 1, 2, 3\}$. Группировка действительных значений осуществлялась разбиением на квартили: 25 процентам разностей ставится в соответствие 0, следующим 25 процентам разностей ставится в соответствие 1 и т. д. Разбиение на квартили для каждого из рядов происходило независимо.

При прогнозировании действительных временных рядов с использованием предложенного метода производились обратные преобразования. Выбирался последний шаг плавно меняющейся закономерности — стационарная закономерность последнего отрезка и применялась к известной части пучка временных рядов. На основе прогнозируемого значения из E_4 определялось прогнозируемое изменение цены акции, которое добавлялось к последнему известному значению исходного ряда. Таким образом получался прогноз на 1 шаг вперед для целевого ряда в исходном пучке временных рядов.

Соответствие, полученное в результате дискретизации, приводится в табл. 6 и 7. Для каждого временного ряда в таблицах представлены диапазоны разностей, которым ставится в соответствие значение из E_4. Отдельно выделено среднее значение в каждом квартиле. Оно используется при обратном переходе от конечнозначных временных рядов к действительным с целью определения прогнозируемого значения исходно ряда.

Таблица 6. Результаты дискретизации

Ряд	От	До	Значение из E_4	Среднее
Adobe	$-\infty$	−0,11750	0	−0,29070
Adobe	−0,11750	0,00650	1	−0,05033
Adobe	0,00650	0,12750	2	0,06096
Adobe	0,12750	∞	3	0,29942
BMC	$-\infty$	−0,05500	0	−0,14016
BMC	−0,05500	0,00000	1	−0,02573
BMC	0,00000	0,05250	2	0,02359
BMC	0,05250	∞	3	0,14402
Business Objects	$-\infty$	−0,09500	0	−0,27475
Business Objects	−0,09500	0,00000	1	−0,04382
Business Objects	0,00000	0,09250	2	0,04332
Business Objects	0,09250	∞	3	0,26859
Cognos	$-\infty$	−0,09250	0	−0,21977
Cognos	−0,09250	0,00250	1	−0,04161
Cognos	0,00250	0,09500	2	0,04365
Cognos	0,09500	∞	3	0,23489
Computer Associate	$-\infty$	−0,06500	0	−0,16745
Computer Associate	−0,06500	0,00250	1	−0,02772
Computer Associate	0,00250	0,07250	2	0,03438
Computer Associate	0,07250	∞	3	0,17251

Модели экспоненциального сглаживания получали на вход исходные действительные временные ряды, что позволило этим методам использовать всю доступную информацию.

При сравнении предложенного метода и экспоненциального сглаживания каждый из методов осуществил 20 прогнозов на один момент времени вперед. Средний квадрат ошибки каждого из методов представлен в табл. 8.

Как видно из табл. 8, при прогнозировании курса акций предложенный метод превосходит по качеству прогнозирования экспоненциальное сглаживание при некоторых параметрах.

Помимо прогнозирования значений пучка временных рядов предложенный в работе алгоритм поиска плавно изменяющихся закономерностей позволил получить представление о характере структурных изменений в пучках временных рядов.

Например, для целевого ряда, описывающего поведение курса акций компании Rational, была найдена следующая плавно меняющаяся закономерность. Шаги плавно меняющейся закономерности — это стационарные закономерности для трех отрезков пучка. Первый отрезок начинается датой 13 мая 2002 г., второй — 21 февраля 2003 г., третий — 28 ноября 2003 г.

Приведенную закономерность можно описать и в терминах действительных временных рядов с помощью табл. 6 и 7. Например, первый столбец закономерности может быть интерпретирован следующим образом: если на последнем временном интервале (час) курса акций Rational упал более чем на 0,0675 доллара США и курс акций Computer Associates

Таблица 7. Результаты дискретизации

Ряд	От	До	Значение	Прогноз
Novell	$-\infty$	−0,02500	0	−0,07158
Novell	−0,02500	−0,00175	1	−0,01273
Novell	−0,00175	0,02500	2	0,01024
Novell	0,02500	∞	3	0,07677
Oracle	$-\infty$	−0,04000	0	−0,09399
Oracle	−0,04000	−0,00025	1	−0,01922
Oracle	−0,00025	0,03750	2	0,01709
Oracle	0,03750	∞	3	0,10058
Peoplesoft	$-\infty$	−0,07000	0	−0,18676
Peoplesoft	−0,07000	−0,00175	1	−0,03260
Peoplesoft	−0,00175	0,06500	2	0,02850
Peoplesoft	0,06500	∞	3	0,19325
Rational	$-\infty$	−0,06750	0	−0,20687
Rational	−0,06750	0,00500	1	−0,02739
Rational	0,00500	0,07500	2	0,03584
Rational	0,07500	∞	3	0,19421

Таблица 8. Средний квадрат ошибки

Целевой Предложенный ряд	Экспоненциальное сглаживание $\alpha = 0{,}1$	Экспоненциальное сглаживание $\alpha = 0{,}3$	метод
Adobe	$9{,}05 \cdot 10^{-2}$	$7{,}32 \cdot 10^{-2}$	$6{,}89 \cdot 10^{-2}$
BMC	$13{,}24 \cdot 10^{-3}$	$11{,}15 \cdot 10^{-3}$	$9{,}72 \cdot 10^{-3}$
Business Objects	$17{,}42 \cdot 10^{-2}$	$7{,}19 \cdot 10^{-2}$	$3{,}74 \cdot 10^{-2}$
Cognos	$6{,}73 \cdot 10^{-2}$	$3{,}08 \cdot 10^{-2}$	$2{,}39 \cdot 10^{-2}$
Computer Associates	$4{,}49 \cdot 10^{-2}$	$2{,}87 \cdot 10^{-2}$	$1{,}91 \cdot 10^{-2}$
Novell	$7{,}62 \cdot 10^{-3}$	$4{,}18 \cdot 10^{-3}$	$2{,}54 \cdot 10^{-3}$
Oracle	$8{,}61 \cdot 10^{-3}$	$6{,}77 \cdot 10^{-3}$	$5{,}45 \cdot 10^{-3}$
Peoplesoft	$3{,}24 \cdot 10^{-3}$	$2{,}94 \cdot 10^{-3}$	$1{,}26 \cdot 10^{-3}$
Rational	$5{,}19 \cdot 10^{-2}$	$3{,}43 \cdot 10^{-2}$	$2{,}06 \cdot 10^{-2}$

Таблица 9. Первый шаг плавно меняющейся закономерности ряда Rational

Rational (t-1)	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3
Computer Associates (t-8)	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Rational (t)	0	0	0	0	0	0	0	0	3	0	3	3	3	3	3	3

уменьшился более чем на 0,065 доллара, то курс акций Rational на следующем временном интервале упадет примерно на 0,2 доллара.

Таблица 10. Второй шаг плавно меняющейся закономерности ряда Rational

Rational (t-1)	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3
Computer Associates (t-**6**)	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Rational (t)	0	0	0	0	**1**	**1**	0	**1**	**2**	**1**	**2**	**2**	3	3	3	**2**

Жирным шрифтом выделены изменения при переходе к следующему шагу плавно меняющейся закономерности.

Таблица 11. Третий шаг плавно меняющейся закономерности ряда Rational

Rational (t-1)	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3
Computer Associates (t-6)	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Rational (t)	0	0	**1**	0	1	1	**1**	1	**1**	**2**	2	2	3	3	**2**	**3**

При переходе от первого шага ко второму произошли более существенные структурные изменения закономерности по сравнению с переходом от второго шага к третьему. Последующий анализ событий на рынке акций показал, что указанное структурное изменение могло быть вызвано покупкой компанией IBM компании Rational, произошедшей 20 февраля 2003 г. Данное событие повлияло на структуру компании Rational и на отношение инвесторов к данной компании. В свою очередь указанные изменения отразились на закономерностях, определяющих поведение временного ряда курса акций компании Rational.

Таким образом, технический анализ пучка временных рядов с использование алгоритма поиска плавно меняющихся закономерностей позволяет выявлять события, которые влияют на закономерности, определяющие поведение рядов.

Результаты испытаний показали, что предложенный в настоящей работе подход может быть более эффективен при краткосрочном прогнозировании, чем другие алгоритмы прогнозирования. При этом алгоритм не только позволяет сделать прогноз, но и осуществляет поиск скрытых закономерностей, описывающих явление.

Заключение

В настоящей работе рассматривается подход к поиску закономерностей в пучках конечнозначных временных рядов. Этот подход позволяет выявлять закономерности, которые подвергаются «плавным» структурным изменениям с течением времени. Для определения подобного рода изменений в работе описана мера сходства закономерностей и описано ее применение как одного из весов на графе закономерностей.

Разрабатываемый алгоритм поиска плавно меняющихся закономерностей в пучках временных рядов решает задачу как алгоритм интеллектуального анализа данных. Он не только позволяет прогнозировать процесс, но и осуществляет поиск скрытых закономерностей в данных и дает возможность в явном виде описать закономерность. Найденные закономерности могут быть использованы как для прогнозирования следующих элементов пучка временных рядов, так и для детального анализа явления, описанного пучком временных рядов, и моделирования явления. Это делает возможным применение предло-

женного алгоритма в широком пласте задач прогнозирования временных рядов, а также в задачах изучения и описания процессов, которые могут представлены пучком временных рядов.

Предложенный в настоящей работе подход был реализован в программной системе и протестирован на модельных и реальных задачах. Испытания на модельных задачах с использованием разработанного экспериментального стенда показали, что алгоритм, основанный на введенных в работе мерах сходства и функционалах качества, позволяет эффективно находить заложенные закономерности, в том числе при достаточно высоком уровне шума.

Эксперименты на модельных пучках временных рядов показали, что использование введенной меры сходства закономерностей в функционале качества существенно повышает качество прогнозирования. Вместе с тем был получен диапазон весов, при котором достигается максимальное качество распознавания.

Анализ реальных временных рядов с применением предложенного алгоритма также свидетельствовал об эффективности алгоритма при краткосрочном прогнозировании. Вместе с тем алгоритм решает и задачу интеллектуального анализа данных, предложив закономерности, описывающие взаимосвязь одномерных временных рядов.

Таким образом, апробация предложенного подхода к прогнозированию процессов с плавно меняющимися закономерностями на модельных и реальных данных позволяет судить о достаточной эффективности предложенных алгоритмов при анализе пучков временных рядов с плавно меняющимися закономерностями.

Литература

[1] *Андерсон Т.* Статистический анализ временных рядов. М.: Мир, 1976.

[2] *Бокс Дж., Дженкинс Г.* Анализ временных рядов, прогноз и управление. М.: Мир, 1974.

[3] *Хеннан Э.* Многомерные временные ряды. М.: Мир, 1974.

[4] *Engle R. F., Kroner K. F.* Multivariate Simultaneous Generalized ARCH // *Econometric Theory*, 1993. Vol. 11, P. 122–150.

[5] *Лукашин Ю. П.* Адаптивные методы краткосрочного прогнозирования временных рядов. М.: Финансы и статистика, 2003. 416 с.

[6] *Agrawal R., Imielinski T., Swamiet A.* Mining association rules between sets of items in large databases // *Conference Management of Data Proceedings*, 1993. P. 207–216.

[7] *Agrawal R., Srikant R.* Mining sequential patterns // *11th Conference (International) on Data Engineering Proceedings*, 1995. P. 3–14.

[8] *Mannila H., Toivonen H., Verkamo A. I.* Discovery of frequent episodes in event sequences // *Data Mining Knowledge Discovery*, 1997. Vol. 1, No. 3. P. 259–289.

[9] *Morchen F., Ultsch A.* Efficient mining of understandable patterns from multivariate interval time series // *Data Mining Knowledge Discovery*, 2007. Vol. 15, No. 2. P. 181–215.

[10] *Das G., Lin K., Mannila H., et al.* Rule discovery from time series // *4th Conference (International) on Knowledge Discovery and Data Mining Proceedings*, 1998. P. 16–22.

[11] *Sayal M.* Detecting time correlations in time-series data streams. Palo Alto: HP Labs, 2004.

[12] *Morchen F., Ultsch A.* Optimizing time series discretization for knowledge discovery // *11th Conference (International) on Knowledge Discovery and Data Mining Proceedings*, 2005. P. 660–665.

[13] *Brown R. G.* Smoothing forecasting and prediction of discrete time series. N.Y.: Prentice-Hall, 1963.

[14] *Trigg D. W., Leach A. G.* Exponential smoothing with an adaptive response rate // *Operat. Res. Quart.*, 1967. Vol. 18, No. 1. P. 53–59.

[15] *Engle R. F.* ARCH: Selected readings. Oxford: Oxford Univ. Press, 1995.

[16] *Zadeh L. A., Ragazzini J. R.* The analysis of sampled-data systems // *Applic. Industry (AIEE)*, 1952. P. 225–234.

[17] *Rao A. G., Shapiro A.* Adaptive smoothing using evolutionary spectra // *Management Sc.*, 1970. Vol. 17, No. 3. P. 208–218.

[18] *Filipenkov N. V.* Data mining in non-stationary multidimensional time series using a rule similarity measure // *IADIS European Conference on Data Mining Proceedings*, 2008. P. 92–96.

[19] *Филипенков Н. В.* Об одном методе поиска плавно меняющихся закономерностей в пучках временных рядов // *Ж. вычисл. матем. и матем. физики*, 2009. Т. 49, № 11. С. 2020–2040.

[20] *Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И.* Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004.

[21] *Журавлев Ю. И.* Избранные научные труды. М.: Магистр, 1998.

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

[23] *Кристофидес Н.* Теория графов. Алгоритмический подход. М.: Мир, 1978. 432 с.

References

[1] *Anderson T. W.* 1971. The Statistical analysis of time series. N.Y: John Wiley & Sons.

[2] *Box G., Jenkins G.* 1970. Time series analysis: Forecasting and control. San Francisco: Holden-Day.

[3] *Hannan E. G.* 1970. Multiple time series. N.Y: John Wiley & Sons.

[4] *Engle R. F., Kroner K. F.* 1993. Multivariate Simultaneous Generalized ARCH. *Econometric Theory* 11:122–150.

[5] *Lukashin Yu. P.* 2003. Adaptive methods for short-term forecasting of Time Series. Moscow: Finansy i Statistika. 416 p. (in Russ.)

[6] *Agrawal R., Imielinski T., Swamiet A.* 1993. Mining association rules between sets of items in large databases. *Conference Management of Data Proceedings* 207–216.

[7] *Agrawal R., Srikant R.* 1995. Mining sequential patterns. *11th Conference (International) on Data Engineering Proceedings* 3–14.

[8] *Mannila H., Toivonen H., Verkamo A. I.* 1997. Discovery of frequent episodes in event sequences. *Data Mining Knowledge Discovery* 1(3):259–289.

[9] *Morchen F., Ultsch A.* 2007. Efficient mining of understandable patterns from multivariate interval time series. *Data Mining Knowledge Discovery* 15(2):181–215.

[10] *Das G., Lin K., Mannila H., et al.* 1998. Rule discovery from time series. *4th Conference (International) on Knowledge Discovery and Data Mining Proceedings* 16–22.

[11] *Sayal M.* 2004. Detecting time correlations in time-series data streams. Palo Alto: HP Labs.

[12] *Morchen F., Ultsch A.* 2005. Optimizing time series discretization for knowledge discovery. *11th Conference (International) on Knowledge Discovery and Data Mining Proceedings* 660–665.

[13] *Brown R. G.* 1963. Smoothing forecasting and prediction of discrete time series. New York: Prentice-Hall.

[14] *Trigg D. W., Leach A. G.* 1967. Exponential smoothing with an adaptive response rate. *Operat. Res. Quart.* 18(1):53–59.

[15] *Engle R. F.* 1995. ARCH: Selected readings. Oxford: Oxford Univ. Press.

[16] *Zadeh L. A., Ragazzini J. R.* 1952. The analysis of sampled-data systems. *Applic. Industry (AIEE)* 225–234.

[17] *Rao A. G., Shapiro A.* 1970. Adaptive smoothing using evolutionary spectra. *Management Sc.* 17(3): 208–218.

[18] *Filipenkov N. V.* 2008. Data mining in non-stationary multidimensional time series using a rule similarity measure. *IADIS European Conference on Data Mining Proceedings* 92–96.

[19] *Filipenkov N. V.* 2009. A method for finding smoothly varying rules in multidimensional time series. *Computational Mathematics Mathematical Physics* 49(11):1930–1948.

[20] *Barsegyan A. A., Kupriyanov M. S., Stepanenko V. V., Kholod I. I.* 2004. Methods and models of data analysis: OLAP and Data Mining. St. Petersburg: BKhV-Peterburg. (in Russ.)

[21] *Zhuravlev Yu. I.* 1998. Selected scientific works. Moscow: Magistr. (in Russ.)

[22] *Rudakov K. V.* 1992. D.Sc. Diss. Moscow: Dorodnicyn Computing Centre of the Russian Academy of Sciences. (in Russ.)

[23] *Christofides N.* 1975. Graph theory: An algorithmic approach. Orlando, FL: Academic Press. 400 p.