

На правах рукописи

Филипенков Николай Владимирович

Об алгоритмах прогнозирования процессов с плавно меняющимися закономерностями

Специальность 05.13.17 — теоретические основы информатики

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва
2010

Работа выполнена в Учреждении Российской академии наук Вычислительный центр им. А. А. Дородницына РАН.

Научный руководитель: доктор физико-математических наук,
профессор, член-корреспондент РАН
Рудаков Константин Владимирович.

Официальные оппоненты: доктор технических наук,
профессор
Дедус Флоренц Федорович,
кандидат физико-математических наук
Гуревич Игорь Борисович.

Ведущая организация: Московский Педагогический Государственный
Университет.

Защита диссертации состоится «22» апреля 2010 г. в 15 часов на заседании диссертационного совета Д 002.017.02 в Учреждении Российской академии наук Вычислительный центр им. А. А. Дородницына РАН по адресу: 119333, Москва, ул. Вавилова, 40.

С диссертацией можно ознакомиться в библиотеке Учреждения Российской академии наук Вычислительный центр им. А. А. Дородницына РАН.

Автореферат разослан «18» марта 2010 г.

Ученый секретарь
диссертационного совета
Д 002.017.02, д.ф.-м.н., профессор

В. В. Рязанов

Общая характеристика работы

Актуальность темы. Анализ и прогнозирование процессов, протекающих во времени, всегда были крайне актуальными задачами. Важность задач прогнозирования обусловила как накопление статистической информации о значениях показателей процессов в различные моменты времени, так и стимулировала развитие методов анализа этих данных.

Еще большую значимость приобрел анализ пучков временных рядов или многомерных временных рядов, который предполагает исследование временных рядов в их взаимосвязи и взаимовлиянии. Пучки временных рядов позволяют описать процесс или систему процессов наиболее полно.

Задачи прогнозирования и поиска закономерностей в пучках временных рядов возникают в различных сферах деятельности человека: медицине, экономике, физике, химии, метеорологии, кибернетике. Пучки временных рядов могут, например, описывать процессы жизнедеятельности человека, стоимость акций на бирже, курсы валют, сигналы, погодные условия и т. д.

Для решения задач анализа временных рядов было предложено большое число методов. В том числе различные методы сглаживания и фильтрации (Р. Г. Браун, А. Лич, Д. Тригг, Ч. Хольт), авторегрессии и скользящего среднего (Дж. Бокс, Г. Дженкинс), модели, учитывающие гетероскедастичность (Р. Энгл). Прорабатывались алгоритмы основанные на спектральных характеристиках (Д. Ваттс, Ф. Ф. Дедус, Г. Дженкинс, Л. Заде, Дж. Рагаззини), статистические модели (С. А. Айвазян, Т. Андерсон, В. М. Бухштабер, М. Кендэл, Г. Крамер, А. Стюарт, Э. Хеннан). Были разработаны различные модели, представляющие закономерности в виде правил: ассоциативных (Р. Агравал, Р. Шрикант), эпизодических (Х. Маннила), иерархических (Ф. Морхен). Для поиска правил также создавались методы дискретизации временных рядов (Г. Дас). Алгебраический подход, предложенный академиком РАН Ю. И. Журавлевым, был применен к задачам анализа

временных рядов и выделения трендов (К. В. Рудаков, Ю. В. Чехович).

Цель работы состоит в разработке алгоритма поиска закономерностей в пучках временных рядов, базирующегося на предположении о плавном изменении закономерностей с течением времени.

В настоящей работе рассматривается задача поиска закономерностей в пучках дискретных нестационарных временных рядов с конечным алфавитом значений. Для ее решения предлагается подход, который позволяет выявлять закономерности, подвергающиеся «плавным» структурным изменениям с течением времени. Для определения подобного рода изменений предлагается мера сходства закономерностей и описывается ее применение как веса на графе закономерностей.

Методика исследований. В диссертации использован экспериментально-теоретический подход к задаче поиска закономерностей как задаче интеллектуального анализа данных. Также использованы методы теории оптимизации и теории графов, комбинаторика. Проведены эксперименты на модельных и реальных пучках временных рядов.

Научная новизна. Все результаты, полученные в диссертации, являются новыми. В работе предложены модель закономерности и подход, который позволяет учитывать «плавное» изменение закономерностей с течением времени. Для определения плавного изменения вводится понятие меры сходства закономерностей, указывающей на близкие закономерности. При этом задача поиска закономерностей рассматривается как задача интеллектуального анализа данных, что позволяет в явном виде описывать найденные закономерности.

Апробация работы. Результаты, изложенные в диссертации, докладывались на XIII Международной научной конференции «Ломоносов» (Москва, 2006 г.); 49-й и 50-й конференциях МФТИ (Долгопрудный,

2006 г. и 2007 г.); Международной конференции «Pattern Recognition and Information Processing»–2007 (Минск, Республика Беларусь, 2007 г.) Всероссийских конференциях «Математические методы распознавания образов» XIII и XIV (Москва, 2007 г. и 2009 г.); Международной конференции «Интеллектуализация обработки информации»–2008 (Алушта, Украина, 2008 г.); Международной конференции «IADIS European Conference on Data Mining»–2008 (Амстердам, Нидерланды, 2008 г.); а также на научных семинарах ВЦ РАН.

Основные результаты диссертации

- 1) Введены и исследованы понятия маски, функции и закономерности в пучках конечнозначных временных рядов.
- 2) Построен алгоритм поиска постоянных закономерностей в пучках временных рядов.
- 3) Получены оценки необходимой длины пучка временных рядов для поиска закономерностей, определенных на всех наборах значений аргументов.
- 4) Предложены меры сходства масок, функций и закономерностей. Получены условия, при которых указанные меры сходства являются метриками.
- 5) Введено и исследовано понятие графа закономерностей. Предложен подход к поиску плавно меняющихся закономерностей как кратчайших путей на графе закономерностей.
- 6) Рассмотрены модификации алгоритма поиска плавно меняющихся закономерностей для выявления периодических закономерностей.
- 7) Реализован программный стенд, позволяющий импортировать и генерировать временные ряды, осуществлять поиск стационарных

и плавно меняющихся закономерностей, а также производить прогнозирование пучков временных рядов.

- 8) Проведена серия экспериментов на модельных пучках временных рядов. На рассмотренных примерах показана высокая эффективность метода. Выработаны практические рекомендации по установке параметров алгоритма.
- 9) Проведены практические эксперименты для сравнения эффективности с другими методами на реальных временных рядах. Показано превосходство предложенного метода. С использованием предложенного алгоритма проведено выявление скрытых закономерностей, определяющих поведение пучка временных рядов, и дана их содержательная интерпретация.

Практическая и теоретическая ценность. Предложенный подход позволяет выявлять плавно меняющиеся закономерности в пучках временных рядов.

Алгоритмы, предложенные в работе, могут быть использованы для прогнозирования временных рядов. Вместе с тем, алгоритмы позволяют решать задачу интеллектуального анализа данных.

Таким образом, подход позволяет выявлять скрытые закономерности в данных и в явном виде выписывать их. Найденные закономерности в дальнейшем могут быть использованы экспертами для детального анализа явления, а также для моделирования процесса, описываемого пучком временных рядов.

Вместе с тем в работе приводятся теоретические результаты, касающиеся предложенных алгоритмов. В том числе, оценивается необходимая длина пучка временных рядов для поиска наиболее универсальных закономерностей. Также рассматриваются условия, при которых предложенные в работе меры сходства являются метриками.

Публикации. По теме диссертации автором опубликованы 10 работ, в том числе 1 работа в ЖВМиМФ без соавторов.

Структура и объем работы. Работа состоит из введения, трех глав и списка литературы из 69 наименований. Общий объем работы — 82 страницы, включая 5 рисунков и 14 таблиц.

Содержание диссертационной работы

Во **введении** дается краткий обзор методов анализа временных рядов. Описывается общая структура диссертации и приводится краткое изложение содержания работы.

В **первой главе** вводятся основные определения и обозначения, рассматривается постановка задачи, вводятся основные понятия маски и закономерности, оценивается необходимая длина пучка временных рядов. В этой главе также рассматривается способ построения закономерности по маске, вводится понятие полноты системы закономерностей, описывается алгоритм поиска постоянных закономерностей.

Пучком временных рядов $\mathfrak{S}$ называется совокупность взаимосвязанных временных рядов S_i , $i \in \{1, 2, \dots, N\}$. Каждый ряд S_i представляет собой последовательность чисел конечнозначной логики $E_{k_i} = \{0, 1, \dots, k_i\}$. 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 , $i \in \{1, 2, \dots, N\}$ в момент времени $t \in \{1, 2, \dots, T\}$ обозначим через $a(i, t)$ или $a_{i,t}$.

Маской ω на прямоугольнике $N \times \Delta$ называется булева матрица размера $N \times \Delta$ (здесь параметр Δ определяет максимальный отступ

по времени). Число единиц в маске ω называется *мощностью* маски и обозначается $||\omega||$. Элемент маски, находящийся в i -й строке и j -м столбце, обозначается $\omega(i, j)$ или $\omega_{i,j}$. *Закономерностью* R называется набор (p, ω, f) с такими особенностями:

- 1) число $p \in \{1, 2, \dots, N\}$ 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 .

$$f: E_{k_{i_1}} \times \dots \times E_{k_{i_{||\omega||}}} \rightarrow E_{k_p} \cup \{\lambda\},$$

где $\omega(i_1, j_1), \dots, \omega(i_{||\omega||}, j_{||\omega||})$ — 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 ($E_{k_1} = \dots = E_{k_N} = E_k$), то функция f принадлежит множеству P_k^* всех частично определенных функций k -значной логики.

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

Рассматривается один из подходов к поиску закономерностей в пучках временных рядов, который предполагает отсутствие изменений в закономерностях с течением времени. Для простоты выкладок предполагается, что $k_i = k$, $i = 1, 2, \dots, N$, т.е. рассматриваются N k -значных временных рядов длины T .

Пусть заданы пучок временных рядов $\mathfrak{S} \in E_k^{N \times T}$, целевой ряд $p \in \{1, 2, \dots, N\}$ и маска $\omega \in E_2^{N \times \Delta}$ с единичными элементами $\omega(i_1, j_1), \dots, \omega(i_{||\omega||}, j_{||\omega||})$, упорядоченными лексикографически. Тогда

при фиксированном целевом ряде p на основании $\mathfrak{S}$ и ω строится множество пар (α_t, v_t) , $t \in \{1, 2, \dots, T - \Delta\}$, где

$$\alpha_t = \{a(i_1, t + j_1 - 1), \dots, a(i_{\|\omega\|}, t + j_{\|\omega\|} - 1)\} \in E_k^{\|\omega\|},$$

$$v_t = a(p, t + \Delta) \in E_k.$$

Пусть α — произвольный набор из $E_k^{\|\omega\|}$ и $v \in E_k$. Частотой $\nu(\alpha, v, p, \omega, \mathfrak{S})$ называется число раз, которое пара (α, v) встречается среди пар (α_t, v_t) , $t \in \{1, 2, \dots, T - \Delta\}$. Достоверностью $Conf_{set}(\alpha, v, p, \omega, \mathfrak{S})$ значения v на наборе α при маске ω на пучке $\mathfrak{S}$ называется частота появления значения v на наборе α :

$$Conf_{set}(\alpha, v, p, \omega, \mathfrak{S}) = \frac{\nu(\alpha, v, p, \omega, \mathfrak{S})}{\sum_{i=0}^{k-1} \nu(\alpha, i, p, \omega, \mathfrak{S})}.$$

Поддержкой $Supp_{set}(\alpha, p, \omega, \mathfrak{S})$ набора α при маске ω на пучке $\mathfrak{S}$ называется частота появления набора α в исследуемой выборке:

$$Supp_{set}(\alpha, p, \omega, \mathfrak{S}) = \frac{\sum_{i=0}^{k-1} \nu(\alpha, i, p, \omega, \mathfrak{S})}{T - \Delta}.$$

Достоверностью $Conf(R, \mathfrak{S})$ закономерности $R = (p, \omega, f)$ на пучке временных рядов $\mathfrak{S}$ называется доля правильных прогнозов закономерности R на пучке временных рядов $\mathfrak{S}$:

$$Conf(R, \mathfrak{S}) = Conf(p, \omega, f, \mathfrak{S}) = \frac{\sum_{\alpha \in E_k^{\|\omega\|}} \nu(\alpha, f(\alpha), p, \omega, \mathfrak{S})}{T - \Delta} =$$

$$= \sum_{\alpha \in E_k^{\|\omega\|}} Conf(\alpha, f(\alpha), p, \omega, \mathfrak{S}) Supp(\alpha, p, \omega, \mathfrak{S}).$$

Поиск постоянных закономерностей на пучке временных рядов $\mathfrak{S}$, определяющих поведение целевого ряда $p = 1, 2, \dots, N$, состоит в последовательном переборе масок. Для каждой маски ω определяется закономерность $R = (p, \omega, f)$, которая наиболее точно описывает исследуемый пучок временных рядов $\mathfrak{S}$, т.е. которая максимизирует достоверность закономерности R . При фиксированном целевом ряде p и маске ω оптимальная закономерность на пучке временных рядов $\mathfrak{S}$ строится путем выбора функции f .

Теорема 1.1. При фиксированном целевом ряде p и маске ω максимальную достоверность $Conf(R, \mathfrak{S})$ на пучке временных рядов $\mathfrak{S}$ имеет закономерность $R_0 = (p, \omega, f_0)$, где значения $f_0(\alpha)$ выбираются следующим образом:

$$f_0(\alpha) = \arg \max_{v \in E_k} \nu(\alpha, v, \omega, \mathfrak{S}).$$

Малая длина T пучка временных рядов $\mathfrak{S}$ приводит к появлению неопределенных значений λ на большом числе наборов функции f . Это способно сделать найденную закономерность $R = (p, \omega, f)$ не пригодной для применения на практике. Ниже приводится оценка необходимой длины пучка временных рядов T при фиксированных k и $\|\omega\|$, которая основана на следующих заключениях. Если набор $\alpha \in E_k^{|\omega|}$ не появился в множестве $\{\alpha_t\} \ t \in \{1, 2, \dots, T - \Delta\}$, то значение функции f на нем равно λ . Поэтому информативной оценкой достаточности длины пучка временных рядов является вероятность того, что в множестве $\{\alpha_t\}$ присутствуют все наборы из $E_k^{|\omega|}$. В работе рассмотрена оценка этой вероятности при равновероятном появлении всех наборов в пучке временных рядов и доказана следующая теорема.

Теорема 1.2. Пусть все наборы из $E_k^{|\omega|}$ появляются в множестве $\{\alpha_t\}$ с равной вероятностью. Обозначим через $M = k^{|\omega|}$ число всех наборов из $E_k^{|\omega|}$, через $L = T - \Delta$ — число элементов в множестве $\{\alpha_t\}$ ($M < L$). Вероятность того, что в множестве $\{\alpha_t\}$ присутствуют все наборы из $E_k^{|\omega|}$, обозначим через P . Тогда $P = M!S(L, M)/M^L$, где $S(L, M)$ — число Стирлинга II рода.

В работе приводятся примеры необходимой длины пучка временных рядов для различных значений параметров P и Δ .

Алгоритм поиска постоянных закономерностей на пучке временных рядов $\mathfrak{S}$ состоит в последовательном переборе масок, максимизирующем достоверность $Conf(R, \mathfrak{S}_{valid})$ построенных закономерностей. Здесь $\mathfrak{S}_{valid}$ — отрезок пучка временных рядов $\mathfrak{S}$, используемый для валидации закономерностей.

В качестве методов оптимизации предлагается использовать методы селекции признаков, применяемые в распознавании образов. Это обусловлено сходством задачи выбора оптимального подмножества признаков и задачи выбора единичных элементов маски. В настоящей работе применяется алгоритм плавающего поиска.

Во **второй главе** предлагается подход к поиску изменяющихся закономерностей, базирующийся на алгоритме поиска постоянных закономерностей. В этой главе определяется ключевое понятие меры сходства закономерностей, описывается методика поиска изменяющихся закономерностей, рассматриваются вопросы сложности алгоритмов и иллюстрируются возможности модификации подхода для поиска периодических закономерностей.

Закономерность, определяющая поведение целевого ряда, может меняться с течением времени. Описанный выше алгоритм не способен адекватно реагировать на подобного рода изменения. Наиболее часто встречающейся в литературе идеей при поиске нестационарных закономерностей является использование «старых» наборов с меньшим весом по отношению к «более новым» наборам. Но данная методика не позволяет выделить закономерности, подвергающиеся структурным изменениям, т.е. изменениям маски или функции. Предлагаемый подход призван исправить этот недостаток. Он учитывает возможные структурные изменения и ориентирован не только на прогнозирование значений ряда, но и на поиск закономерностей, которые затем могут быть использованы при анализе и моделировании процесса, описываемого пучком временных рядов.

Идея подхода состоит в разбиении исходного пучка временных рядов на отрезки, на каждом из которых применяется алгоритм поиска постоянных закономерностей. Наиболее близкие в смысле некоторой меры сходства закономерности, полученные на различных отрезках, считаются этапами эволюции одной закономерности. Ключевую роль в алгоритме

играет мера сходства на закономерностях.

Во второй главе вводится и исследуется понятие меры сходства на закономерностях, полученных в результате анализа k -значных временных рядов. Предполагается, что закономерности в процессе структурной деформации не могут менять номер целевого ряда. Поэтому мера сходства вводится на закономерностях с одинаковым параметром p .

Построение меры сходства масок происходит в несколько шагов. На первом этапе определяется мера сходства $\rho_m(\omega_1, \omega_2)$ для масок ω_1, ω_2 одинаковой мощности. Указанная мера сходства отражает минимальную стоимость деформации одной маски в другую. Доказывается теорема о том, что

Теорема 2.1. *Описанное отображение ρ_m является метрикой.*

Мера сходства ρ_m обобщается на случай масок произвольной мощности. Идея состоит в том, чтобы переставить элементы меньшей маски, получив подмножество элементов большей маски. Определяется мера сходства ρ'_m для масок произвольной мощности.

Теорема 2.2. *Для меры сходства ρ'_m выполнены все аксиомы метрики, за исключением неравенства треугольника.*

Показывается, что введенную меру сходства ρ'_m можно сделать метрикой, изменив слагаемое, отвечающее за различие в мощности масок. Полученная метрика обозначается ρ''_m .

Далее приводится следующее описание нормировки метрики ρ''_m . Пусть в рамках рассматриваемой задачи поиска закономерностей фиксировано максимальное количество аргументов функции. Это означает, что существует некоторый параметр μ , который определяет максимальное количество единиц в маске. Тогда метрика ρ''_m нормируется с использованием параметра μ . Итоговая мера сходства масок обозначается ρ^μ_m .

После исследования меры сходства масок во второй главе определяются и исследуются меры сходства на функциях.

Вначале определяется вспомогательная мера $\hat{\rho}$. Мера сходства $\hat{\rho}$ с

параметром $w_\lambda \in \mathbb{R}$ вводится на множестве $\{0, 1, \dots, k-1, \lambda\}$. Для этого доопределяется модуль разности, действующий на множестве $\{0, 1, \dots, k-1\}$, следующими правилами:

- 1) $\hat{\rho}(\lambda, x) = \hat{\rho}(x, \lambda) = w_\lambda > 0 \quad \forall x \in \{0, 1, \dots, k-1\}$,
- 2) $\hat{\rho}(\lambda, \lambda) = 0$.

Теорема 2.3. *Описанное отображение $\hat{\rho}$ является метрикой тогда и только тогда, когда $k \leq 2w_\lambda + 1$.*

Следствие 1. Минимальное w_λ , при котором $\hat{\rho}$ является метрикой, равно $(k-1)/2$.

На основе меры $\hat{\rho}$ определяется метрика ρ_f на функциях, которые зависят от одинакового числа переменных. Она вводится как метрика на векторах фиксированной длины, координатами которых могут быть элементы множества $\{0, 1, \dots, k-1, \lambda\}$, где λ обозначает отсутствие значения функции на данном наборе. Предполагается, что фиксирован некоторый порядок переменных. Справедлива следующая теорема.

Теорема 2.4. *Минимальное w_λ , при котором ρ_f является метрикой, равно $(k-1)/2$. При $0 \leq w_\lambda \leq (k-1)/2$ выполнены все аксиомы метрики, за исключением неравенства треугольника.*

Метрика на произвольных частично определенных функциях из P_k^* определяется так же, как и для функций, которые зависят от одинакового числа переменных. Для этого добавляются фиктивные переменные к функции, у которой их меньше.

Основываясь на содержательных соображениях, формулируются два требования к мере сходства на закономерностях. Во-первых, она должна отражать близость масок, а во-вторых — близость функций.

Пусть $R_1 = (p, \omega_1, f)$, $R_2 = (p, \omega_2, g)$ — закономерности, порожденные k -значными временными рядами, т.е. функции f и g принадлежат P_k^* . Определим отображение ρ на закономерностях, используя понятия

метрики на масках и метрики на частично определенных функциях, введенные ранее:

$$\rho(R_1, R_2) = \varkappa_m \rho_m^\mu(\omega_1, \omega_2) + \varkappa_f \rho_f(f, g).$$

Здесь $\varkappa_m$ и $\varkappa_f$ — веса мер сходства, удовлетворяющие следующим условиям: $0 \leq \varkappa_m \leq 1$, $0 \leq \varkappa_f \leq 1$, $\varkappa_m + \varkappa_f = 1$.

Затем в работе вводится понятие последовательности отрезков. *Отрезком* $\mathfrak{S}^1$ на пучке временных рядов $\mathfrak{S}$ с *началом* $t_b \in \{0, 1, \dots, T\}$ и *концом* $t_e \in \{0, 1, \dots, T\}$ ($t_b < t_e$) назовем матрицу размерности $N \times \theta$, где $\theta = t_e - t_b + 1$, составленную из последовательных столбцов матрицы $\mathfrak{S}$, первым из которых является столбец с номером t_b , а последним — столбец с номером t_e . Величина θ называется *длиной* отрезка $\mathfrak{S}^1$. Множество отрезков $\mathfrak{S}^1, \dots, \mathfrak{S}^m$ на пучке временных рядов $\mathfrak{S}$ с началами $t_b^1, \dots, t_b^m$ и концами $t_e^1, \dots, t_e^m$ соответственно называется *последовательностью отрезков*, если $\forall i, j \in \{1, 2, \dots, m\}$ справедливо $\{(i < j) \Rightarrow ((t_b^i < t_b^j) \& (t_e^i < t_e^j))\}$.

На основе понятия последовательности отрезков и стационарной закономерности вводится понятие плавно меняющейся закономерности.

Пусть $\mathfrak{S}^1, \dots, \mathfrak{S}^m$ — последовательность отрезков на пучке временных рядов $\mathfrak{S}$. Введем понятия изменяющейся закономерности и ее длины.

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

Длиной $l(\tilde{R})$ изменяющейся закономерности $\tilde{R}$ называется сумма мер сходства «соседних» шагов — закономерностей, составляющих меняющуюся закономерность.

Пусть каждый из отрезков $\mathfrak{S}^j$, $j = 1, 2, \dots, m$, разбит на две части: обучение $\mathfrak{S}_{train}^j$ и валидацию $\mathfrak{S}_{valid}^j$. Тогда алгоритм поиска постоянных

закономерностей, примененный к каждому из отрезков, порождает наборы закономерностей: $R_1^1, \dots, R_{q_1}^1$ - на отрезке $\mathfrak{S}^1$, ..., $R_1^m, \dots, R_{q_m}^m$ - на отрезке $\mathfrak{S}^m$. Для каждой закономерности R_i^j , $i = 1, 2, \dots, q_j$, $j = 1, 2, \dots, m$ определены значения показателей качества: достоверность на обучении $Conf(R_i^j, \mathfrak{S}_{train}^j)$, достоверность на валидации $Conf(R_i^j, \mathfrak{S}_{valid}^j)$, поддержка на обучении $Supp(R_i^j, \mathfrak{S}_{train}^j)$ и т.п. Найденные закономерности можно представить в виде следующего графа закономерностей.

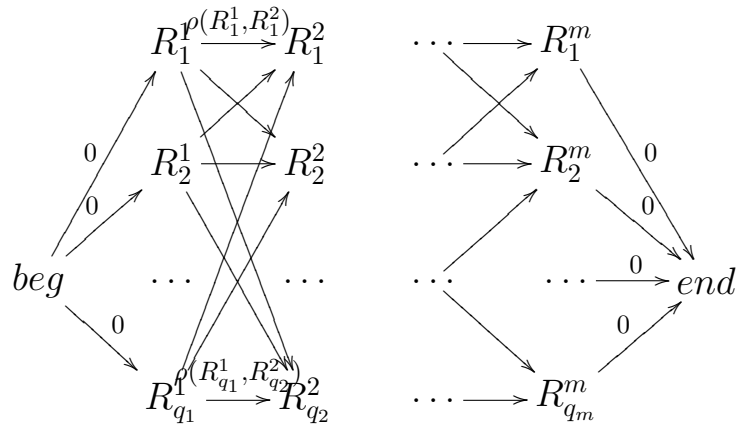


Рисунок 1. Граф закономерностей

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.

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 $(1 - Q_{step})$, где Q_{step} — функционал качества шага изменяющейся закономерности $\tilde{R}$, который также определяется в работе.

В конце второй главы рассматриваются показатели качества плавно меняющихся закономерностей. Отдельно исследуется вопрос изменения структуры графа для поиска плавно меняющихся закономерностей.

В третьей главе изложены результаты экспериментов на модельных и реальных пучках временных рядов.

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.

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

Первая серия проводилась с целью анализа стабильности предложенного алгоритма поиска изменяющихся закономерностей к шуму в исходных данных. Вторая серия экспериментов позволила оценить влияние меры сходства закономерностей при поиске плавно меняющихся закономерностей.

Результаты первой серии экспериментов на модельных рядах 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.

Вторая серия экспериментов показала, что добавление меры сход-

ства закономерностей в функционал качества позволяет существенно повысить точность распознавания в пучках с плавно меняющимися закономерностями. При этом были определены рекомендуемые значения одного из ключевых параметров алгоритма: веса меры сходства закономерностей.

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, работающих в сфере информационных технологий. Результаты испытаний на реальных временных рядах 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.

В **заключении** подводится краткий итог диссертации.

Список публикаций по теме диссертации

1. *Филипенков Н. В.* Поиск плавно меняющихся закономерностей в пучках временных рядов // XIII Международная научная конференция «Ломоносов»: Тез. докл. — М.: Изд-во МГУ, 2006. — С. 56–57.
2. *Филипенков Н. В.* О задачах анализа пучков временных рядов с изменяющимися закономерностями // Искусственный интеллект. — 2006. — № 2. — С. 125–129.
3. *Филипенков Н. В.* Об одном методе выявления плавно меняющихся закономерностей в k -значных временных рядах // 49-я научная конференция МФТИ: Тез. докл. — М.: МФТИ, 2006. — С. 270–271.

4. *Филипенков Н. В.* Об оптимальном выборе закономерностей, составляющих плавно меняющуюся закономерность // Математические методы распознавания образов XIII: Тез. докл. — М.: МАКС Пресс, 2007. — С. 223–225.
5. *Филипенков Н. В.* Поиск плавно меняющихся ассоциативных правил // 50-я научная конференция МФТИ: Тез. докл. — Часть VII. Управление и прикладная математика. Том 2. — М.: МФТИ, 2007. — С. 117–119.
6. *Филипенков Н. В.* Об эволюционирующих алгоритмах классификации и прогнозирования // Интеллектуализация обработки информации: Тез. докл. — Симферополь. Крымский НЦ НАН Украины, 2008. — С. 228–230.
7. *Филипенков Н. В.* О некоторых аспектах интеллектуального анализа пучков временных рядов // Математические методы распознавания образов XIV: Тез. докл. — М.: МАКС Пресс, 2009. — С. 204–207.
8. *Филипенков Н. В.* Об одном методе поиска плавно меняющихся закономерностей в пучках временных рядов // **Ж. вычисл. матем. и матем. физ.** — 2009. — Т. 49, № 11. — С. 2020–2040.
9. *Filipenkov N. V.* On the mining of slightly changing patterns in multidimensional time series // Proc. 9th Internat. Conf on Pattern Recognition and Information Processing. — Minsk, 2007. — Vol. 1 — pp. 123–127.
10. *Filipenkov N. V.* Data mining in non-stationary multidimensional time series using a rule similarity measure // Proc. IADIS European Conf. on Data Mining. — Amsterdam, 2008. — pp. 92–96.