САРАПАС Владимир Викторович

Алгебраические методы синтеза алгоритмов классификации элементов временных рядов

Специальность 05.13.17 – теоретические основы информатики

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва – 2010

Работа выполнена на кафедре теоретической информатики и дискретной математики математического факультета Московского педагогического государственного университета

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

Официальные оппоненты:

доктор физико-математических наук
СМЕТАНИН Юрий Геннадьевич

кандидат физико-математических наук,
доцент ГУРОВ Сергей Исаевич

Ведущая организация:
Московский физико-технический институт
(государственный университет)

Защита состоится «11» марта 2011 г. в «16» часов на заседании диссертационного совета Д 212.154.32 при Московском педагогическом государственном университете по адресу: 107140, г. Москва, ул. Краснопрудная, д. 14, математический факультет МПГУ, ауд. 301.

С диссертацией можно ознакомиться в библиотеке Московского педагогического государственного университета: 119992, Москва, ул. Малая Пироговская, д. 1.

Автореферат разослан « 7 » февраля 2011 г.

Ученый секретарь
диссертационного совета МУРАВЬЁВА О.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Создание и исследование систем распознавания образов является трудоёмкой теоретической и технической задачей, необходимость её решения возникает в различных областях. Одним из направлений теории распознавания образов является алгебраический подход, включающий в себя методы построения проблемно-ориентированных теорий синтеза корректных алгоритмов распознавания образов на базе параметрических семейств операторов.

Алгебраический подход возник в результате исследований, в которых соответствующие решаемым задачам алгоритмы преобразования информации строились в контексте отсутствия подходящих математических моделей конкретных ситуаций.

Академик РАН Ю.И. Журавлёв в 70-х годах XX века заложил основы алгебраического подхода к синтезу корректных алгоритмов распознавания образов. Алгебраический подход к проблеме распознавания образов позволил по-новому и эффективно решать многие задачи классификации, прогнозирования и, вообще говоря, задачи преобразования информации.

В основу алгебраического подхода легла идея выбирать некоторые алгоритмы из эвристических семейств и, используя подходящие корректирующие операции над ними, получать оптимальные алгоритмы для решаемых задач. Необходимо подчеркнуть, что вышеупомянутая идея активно использовалась и используется различными группами исследователей (Н.Г. Белецкий, В.С. Казанцев, В.Д. Мазуров, Л.А. Растригин, Р.Х. Эренштейн и др.). В основополагающих работах Ю.И. Журавлёва по алгебраическому подходу к распознаванию образов помимо использования и развития этой идеи были введены такие важные понятия алгебраического подхода, как регулярность и полнота.

При дальнейших исследованиях были получены важные результаты для многих семейств алгоритмов и корректирующих операций над ними (А.Р. Ашуров, Ю.И. Журавлёв, И.В. Исаев, В.В. Краснопрошин, К.В. Рудаков, В.В. Рязанов и др.). В итоге всех вышеперечисленных исследований алгебраический подход стал общетеоретической базой для решения задач распознавания и используемых при этом математических конструкций и методов.

Отметим, что применение алгебраических конструкций было обосновано на базе принятия некоторых дополнительных метрических и статистических гипотез. Исследования первого типа проводились Ю.И. Журавлёвым и его учениками, а исследования второго типа, для которых был создан специальный тонкий математический аппарат, были проведены академиком РАН В.Л. Матросовым. Им были устранены некоторые внешние противоречия между статистической теорией и алгебраическим подходом.

Основополагающие идеи академика РАН Ю.И. Журавлёва развил в своих трудах член-корреспондент РАН К.В. Рудаков. Он разработал алгебраическую теорию универсальных и локальных ограничений для алгоритмов распознавания, чем расширил границы применимости идей

алгебраического подхода. Им была исследована проблема разрешимости и регулярности задач классификации и получен общий необходимый и достаточный критерий регулярности, который для отдельных конкретных систем универсальных ограничений сводится к легко проверяемым на практике условиям. Также К.В. Рудаковым были получены критерии полноты для моделей алгоритмов как на общем уровне, так и для конкретных систем универсальных ограничений; выявлены критерии полноты для моделей алгоритмических операторов и семейств корректирующих операций и сформулировано понятие корректности семейств решающих правил.

При решении многих прикладных задач часто возникает необходимость выделения внутри временного ряда так называемых трендов. Как правило, трендом называют интервал временного ряда, не содержащий точек экстремума. Задачи выделения трендов временных рядов находятся в центре многих отраслей науки и прикладных сфер.

В реферируемой диссертации под выделением трендов подразумевается решение задачи классификации, в которой каждой точке ряда сопоставляется номер класса из заранее определенного множества классов или, говоря иначе, метка из фиксированного словаря разметки. Анализ распределенных во времени элементов требует применения комплексных решений, что и обеспечивают алгебраические методы синтеза алгоритмов классификации. Алгебраический подход как способ построения проблемно-ориентированных теорий синтеза корректных алгоритмов распознавания образов на базе параметрических семейств операторов, позволяет построить такую теорию над предметной областью, объектами изучения которой будут являться обучаемые алгоритмы выделения трендов, семейства алгоритмов и операции над ними.

В рамках алгебраического подхода К.В. Рудаков и Ю.В. Чехович разработали общую теорию задач с теоретико-множественными ограничениями. Также была разработана специализация этой теории для решения задач синтеза алгоритмов, описывающих отображения из пространства начальных информаций – конечных множеств точек на плоскости – во множество финальных информаций – множество конечных наборов «меток» – слов в конечном алфавите., то есть алгоритмов выделения трендов временных рядов.

При этом не были построены конкретные примеры моделей, удовлетворяющие разработанной теории. Поэтому построение, теоретическое и экспериментальное исследование таких семейств оставалось нерешённой **актуальной задачей**. В реферируемой диссертации такие модели предложены, а также проведено их теоретическое и экспериментальное исследование.

Объектом исследования являются методы алгебраического подхода к решению задач синтеза обучаемых алгоритмов выделения трендов временных рядов.

Предмет исследования составили параметрические семейства алгоритмов, семейства корректирующих операций и решающих правил для задач синтеза обучаемых алгоритмов выделения трендов временных рядов.

Целью данной работы является синтез и исследование параметрических семейств алгоритмов классификации элементов временных рядов, допускающих минимум ошибок на прецедентах и удовлетворяющих наборам дополнительных ограничений. Для достижения указанной цели в работе поставлены и решены следующие **задачи**:

1) конкретизировать основные принципы и методы алгебраического подхода для задач выделения трендов;
2) построить параметрические модели алгоритмов классификации элементов временных рядов, задать семейства корректирующих операций над этими алгоритмами и семейства решающих правил;
3) исследовать построенные параметрические модели алгоритмов классификации элементов временных рядов;
4) разработать соответствующее программное обеспечение;
5) провести серию экспериментов и сделать необходимые выводы.

Методы исследования. В настоящей работе использовались методы алгебраического подхода к распознаванию образов, теории задач с теоретико-множественными ограничениями, теории оптимизации, системного анализа, для проведения экспериментов использовались специально разработанные программные средства.

Научная новизна исследования обусловлена тем, что в его рамках созданы новые модели алгоритмов, проведён их анализ с помощью аппарата теории задач с теоретико-множественными ограничениями и тем, что впервые разработана программа для синтеза обучающихся алгоритмов выделения трендов временных рядов.

Теоретическая значимость обусловлена, прежде всего, тем, что впервые для анализа моделей алгоритмов используется теория задач с теоретико-множественными ограничениями. Данное исследование развивает и дополняет алгебраическую теорию синтеза обучаемых алгоритмов выделения трендов временных рядов.

Практическая значимость заключается в том, что результаты работы могут быть использованы в дальнейших научных исследованиях, посвящённых проблеме выделения трендов временных рядов, а также в возможности использования результатов исследования в таких практических сферах, как экономика, астрономия, медицинская диагностика, геологический анализ.

Основные положения, выносимые на защиту:

1) семейства фильтрующих алгоритмов выделения трендов временных рядов полны при системе локальных аксиом;
2) семейство фильтрующих алгоритмов с переменным параметром выделения трендов временных рядов полно при расширенной системе аксиом;

3) алгебраическая коррекция результатов работы семейств алгоритмов в сплит-модели позволяет добиться улучшения результатов классификации;
4) результаты экспериментального исследования демонстрируют практическую применимость разработанных конструкций.

Апробация работы. Результаты диссертационного исследования и его отдельные положения были представлены в докладах и выступлениях на кафедре теоретической информатики и дискретной математики Московского педагогического государственного университета, в отделе «Интеллектуальные системы» Вычислительного центра имени А.А. Дородницына РАН, обсуждались на научно-практической конференции преподавателей, аспирантов и сотрудников математического факультета МПГУ (Москва, 2007, 2008), заседаниях круглого стола молодых ученых по приоритетным направлениям развития науки (Москва, 2007, 2008), Девятой международной научно-практической конференции «Высокие технологии, исследования, промышленность» (Санкт-Петербург, 2010), XI Всероссийской научно-практической конференции студентов, аспирантов и молодых учёных с международным участием «Молодежь и наука XXI века» (Красноярск, 2010), II Международной научно-практической конференции «Наука и современность – 2010» (Новосибирск, 2010). По теме диссертации опубликовано пять работ, в том числе одна в журнале, включённом в список ВАК.

Структура диссертации соответствует логике научного исследования, определяется его целью и основными задачами: работа (общим объёмом 102 страниц) состоит из введения, трёх глав, заключения, списка литературы и приложения. Библиография включает 103 наименования научной литературы.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** даётся общая характеристика работы: обосновывается актуальность темы диссертации, выбор объекта и предмета исследования, раскрывается его научная новизна, теоретическая и практическая значимость, определяются цель и задачи, а также положения, выносимые на защиту; характеризуется общая методологическая база работы.

В **первой главе** рассматриваются общие вопросы задач распознавания и классификации в рамках алгебраического подхода.

О постановке задачи распознавания имеет смысл говорить тогда, когда применение классических математических методов и построение формальных теорий чем-либо осложнено или вовсе невозможно. Располагая информацией о классах, благодаря параметризации моделей алгоритмов распознавания, можно синтезировать корректные алгоритмы для некоторых узких подклассов задач, при отсутствии точной модели изучаемых объектов и явлений. Важно отметить, что использование достаточно «богатых» многопараметрических моделей влечёт за собой обычно трудноразрешимую

проблему поиска глобально оптимального алгоритма, причём часто получается так, что при использовании более «бедной» модели, в рамках которой легко отыскать оптимальный алгоритм, возможно получение более точного решения задачи распознавания, чем при использовании локально-экстремального алгоритма из «богатой» сложноустроенной модели.

В контексте алгебраического подхода к синтезу корректных алгоритмов распознавания образов, классификации и прогнозирования в первой главе диссертации рассматривается класс задач, характеризующийся наличием явным образом заданных теоретико-множественных ограничений на множество допустимых ответов алгоритма.

В соответствии с теорией члена-корреспондента РАН К.В. Рудакова, опишем задачу классификации в виде задачи синтеза алгоритма преобразования информации. Будем рассматривать некоторое множество $\mathfrak{S}=\{S\}$, элементы которого называются объектами. Описания объектов $D(S)$ образуют пространство начальных информаций $\mathfrak{I}_i=\{D(S)|\ S\in\mathfrak{S}\}$, элементы которого обозначаются I_i, так что $\mathfrak{I}_i=\{I_i\}$.

В первой главе рассматривается задача синтеза алгоритмов A, реализующих отображения из пространства начальных информаций $\mathfrak{I}_i$ в пространство финальных информаций $\mathfrak{I}_f=\{I_f\}$. Далее мы не будем различать алгоритмы и реализуемые ими отображения. Решение синтезируется в рамках модели алгоритмов $\mathfrak{M}$, где $\mathfrak{M}\subseteq\{A|\ A:\mathfrak{I}_i\to\mathfrak{I}_f\}$. Задачи определяются структурными информациями I_s, выделяющими из $\mathfrak{M}$ подмножества допустимых отображений, обозначаемые $\mathfrak{M}[I_s]$. Любой алгоритм A, реализующий произвольное допустимое отображение, называется корректным для задачи, определяемой структурной информацией I_s, и является ее решением.

Конструкции алгебраического подхода к проблеме синтеза корректных алгоритмов основаны на использовании «промежуточного» по отношению к $\mathfrak{I}_i$ и $\mathfrak{I}_f$ пространства оценок $\mathfrak{I}_e=\{I_e\}$. При этом корректные алгоритмы синтезируются на базе эвристических информационных моделей, т.е. параметрических семейств отображений из $\mathfrak{I}_i$ в $\mathfrak{I}_f$, представляющих собой специальные суперпозиции алгоритмических операторов (отображений из $\mathfrak{I}_i$ в $\mathfrak{I}_e$) и решающих правил (отображений из $\mathfrak{I}_e^p$ в $\mathfrak{I}_f$, p – арность решающего правила).

Модели $\mathfrak{M}$ определяются моделями алгоритмических операторов $\mathfrak{M}^0$, где $\mathfrak{M}^0\subseteq\mathfrak{M}_*=\{B|\ B:\mathfrak{I}_i\to\mathfrak{I}_e\}$, и решающих правил $\mathfrak{M}^1$, где $\mathfrak{M}^1\subseteq\bigcup_{p=0}^{\infty}\{C|\ C:\mathfrak{I}_e^p\to\mathfrak{I}_f\}$ следующим образом:

$$\mathfrak{M} = \mathfrak{M}^1 \circ \mathfrak{M}^0 = \{C \circ (B_1 \times ... \times B_p)_\Delta \mid C \in \mathfrak{M}^1, B_1, ..., B_p \in \mathfrak{M}^0\}.$$

Для синтеза корректных алгоритмов используются также множества $\mathfrak{F}$ корректирующих операций, определённых над множеством отображений $\mathfrak{M}_*$. Корректирующие операции F, рассматриваемые в настоящей работе, индуцируются операциями $\tilde{F}$ над пространством оценок $\mathfrak{I}_e$:

$$F(B_1,...,B_p)(I_i) = \tilde{F}(B_1(I_i),...,B_p(I_i)),$$

где I_i пробегает пространство начальных информаций $\mathfrak{I}_i$, алгоритмические операторы $B_1,...,B_p$ – произвольные отображения из $\mathfrak{I}_i$ в $\mathfrak{I}_e$, и $\tilde{F}$ – операция над $\mathfrak{I}_e$.

Схема построения модели алгоритмов $\mathfrak{M}$ представлена на следующей коммутативной диаграмме:

$$\begin{array}{ccc} \mathfrak{I}_i & \xrightarrow{\mathfrak{M}} & \mathfrak{I}_f \\ \mathfrak{M}^0 \downarrow & & \uparrow \mathfrak{M}^1 \\ \mathfrak{I}_e^p & \xrightarrow{\mathfrak{F}} & \mathfrak{I}_e \end{array}$$

При зафиксированном пространстве возможных начальных информаций $\mathfrak{I}_i$ и пространстве финальных информаций $\mathfrak{I}_f$ любую из задач определяет соответствующая структурная информация I_s, являющаяся системой ограничений, которые выделяют из $\mathfrak{M}_*$ подмножество допустимых для задачи отображений $\mathfrak{M}[I_s]$. Прецедентные ограничения, то есть наборы пар вида $((I_i^1, I_f^1),...,(I_i^q, I_f^q))$, сопровождаемые требованием $A(I_i^k) = I_f^k$ при $k \in \{1,...,q\}$, являются характерной частью структурной информации в рассматриваемых задачах. Кроме того, структурная информация может содержать и дополнительные ограничения на вид отображений, реализуемых корректными алгоритмами. Член-корреспондент РАН К.В. Рудаков предложил рассматривать прецедентные и дополнительные ограничения как абсолютно равноправные части структурной информации. Прецедентные и дополнительные к ним ограничения имеют принципиально важное отличие: рассматривая алгоритм как «чёрный ящик» можно легко проверить, удовлетворяет ли он прецедентным ограничениям, однако осуществить такую проверку для дополнительных ограничений обычно невозможно.

В рамках алгебраического подхода преимущественно используется понятие регулярности, которое является обобщением понятия разрешимости. Это обусловлено, прежде всего, тем, что объектами теоретического исследования являются не отдельные задачи, а классы в некотором смысле однородных задач. При таком подходе возможно из факта принадлежности

конкретной задачи определённому классу делать выводы о её разрешимости методами, разработанными для данного класса задач. Другими словами, *проблема регулярности задач синтеза корректных алгоритмов является* обобщением проблемы разрешимости.

Пусть имеется разбиение множества $\mathfrak{Z}$ изучаемых задач с общей системой универсальных ограничений на классы эквивалентности по некоторому отношению «≈»: $Z_1 \approx Z_2$ если задачи Z_1 и Z_2 неразличимы в момент выбора и анализа модели алгоритмов.

Задача Z из множества $\mathfrak{Z}$ называется регулярной, если она разрешима и разрешимы все задачи из класса эквивалентности по отношению «≈», в который она входит.

Задача Z из множества $\mathfrak{Z}$ называется полной относительно семейства $\mathfrak{M}$ отображений из $\mathfrak{I}_i$ в $\mathfrak{I}_f$, если в $\mathfrak{M}$ содержатся допустимые отображения для всех задач из класса эквивалентности, содержащего Z; задача Z называется регулярной, если для нее существует семейство отображений $\mathfrak{M}$, относительно которого она полна.

Так как подклассы регулярных задач обычно меньше классов разрешимых задач, то требование полноты является более мягким, чем требование разрешимости для всех в принципе разрешимых задач, поэтому синтез полных моделей алгоритмов – как правило, более реальная на практике задача, нежели построение моделей для решения всех разрешимых задач.

Во **второй главе** ставится задача выделения трендов временных рядов, производится формализация этого класса задач, описываются методы их решения, доказывается полнота некоторых моделей алгоритмов.

В основу исследования были положены принципы алгебраического подхода к синтезу корректных алгоритмов выделения трендов с использованием теории задач с теоретико-множественными ограничениями.

Необходимо отметить, что решение задачи выделения трендов некоторого временного ряда достаточно сильно зависит от мнения эксперта и поэтому, как правило, не имеет единственного решения. Следовательно, представляется рациональным рассматривать алгоритмы выделения трендов, настраиваемых на определённый тип анализа.

В соответствии с теорией, изложенной в работах Ю.В. Чеховича, рассмотрим конечные наборы точек на плоскости $S(v,t) \in R^2$. Конечной плоской конфигурацией (КПК) назовём вектор $\bar{S}^d = (S^1,...,S^d) = ((t^1,v^1),...,(t^d,v^d))$, где $t \in R, v \in R, d \geq 1$. При этом значения $t^i, i = 1,...,d$ таковы, что $t^i < t^{i+1}$, либо $t^i \leq t^{i+1}$, и если $t^i = t^{i+1}$, то $v^i < v^{i+1}$. В случае строгого неравенства будем называть конфигурацию однозначной, во втором случае – неоднозначной. Множество всех

однозначных конфигураций определим как $K_1 = \bigcup_{d=1}^{\infty} K_1^d$, а множество неоднозначных как $K = \bigcup_{d=1}^{\infty} K^d$. Очевидно, что $K_1 \subseteq K$. Мощностью конфигурации будем называть количество её точек.

Словарём разметки или множеством меток называется конечное множество $M = \{\mu_1, ..., \mu_r\}, r \geq 1$.

Расширенным множеством меток или расширенным словарём разметки называется множество $M_\Delta = M \cup \{\Delta\}, \Delta \notin M$, где Δ - специальная метка, означающая «не размечено».

Например, разметка может быть такой: $M = \{\min, \max, down, up, plt\}$, где «min» - метка для обозначения точки минимума, «max» - точки максимума, «down» - точки убывания, «up» - точки возрастания, «plt» - точки плато.

Размеченным объектом называется пара (S, μ), где $\mu \in M$. Размеченной конфигурацией называется пара $(\bar{S}^d, \bar{\mu}^d)$, где $\bar{S}^d \in K^d$ для неоднозначной и $\bar{S}^d \in K_1^d$ для однозначной конфигураций, а $\bar{\mu}^d \in M^d$. $\bar{\mu}^d$ называется разметкой или полной разметкой конфигурации $\bar{S}^d$. Если $\bar{\mu}^d \in M_\Delta^d$, то пара $(\bar{S}^d, \bar{\mu}^d)$ называется частично размеченной конфигурацией, разметка $\bar{\mu}^d$ при этом называется частичной разметкой конфигурации $\bar{S}^d$.

Алгоритмом разметки называется всякий алгоритм A, реализующий отображение $A : C \to \tilde{M}$ такое, что для любого $d \geq 1$ верно $A(\bar{S}^d) = \bar{\mu}^d$, где $\bar{S}^d \in K^d, \bar{\mu}^d \in M^d$.

Аксиомами или правилами разметки называется набор $\Pi = \{\pi_1, ..., \pi_k\}$ эффективно вычислимых предикатов: $\pi_i : \bigcup_{d=1}^{\infty} (K^d \times M^d) \to \{0,1\}$.

Пусть фиксирована система аксиом разметки $\Pi = \{\pi_1, ..., \pi_k\}$. Разметка $\bar{\mu}^d$ называется подходящей для $\bar{S}^d$, если $\Pi(\bar{S}^d, \bar{\mu}^d) = 1$. Частичная разметка $\bar{\mu}_0^d \in M_\Delta^d$ называется подходящей для $\bar{S}^d$ тогда и только тогда, когда существует полная подходящая разметка $\bar{\mu}^d$, являющаяся продолжением $\bar{\mu}_0^d$.

Система аксиом разметки $\Pi = \{\pi_1,...,\pi_k\}$ является непротиворечивой тогда и только тогда, когда выполнено следующее условие: $\exists\ d \ \exists\ \bar{S}^d \ \exists\ \bar{\mu}^d : \Pi(\bar{S}^d, \bar{\mu}^d) = 1$, то есть когда существует по меньшей $d \geq 1 \quad K^d \quad M$ мере одна конфигурация, для которой существует подходящая разметка.

Система аксиом разметки $\Pi = \{\pi_1,...,\pi_k\}$ является независимой тогда и только тогда, когда выполнено условие: $\forall\ w \ \exists\ d \ \exists\ \bar{S}^d \ \exists\ \bar{\mu}^d : (\Pi^w(\bar{S}^d, \bar{\mu}^d) = 1) \wedge (\Pi(\bar{S}^d, \bar{\mu}^d) = 0)$, где $w = 1...k \quad d \geq 1 \quad K^d \quad M$

$\Pi^w = \{\pi_1,...,\pi_{w-1},\pi_{w+1},...,\pi_k\}$, то есть, в том случае, когда при изъятии из системы любой аксиомы существует хотя бы одна конфигурация и существует разметка этой конфигурации, являющаяся подходящей в смысле усеченного набора аксиом и неподходящей в смысле исходного набора.

Система аксиом разметки $\Pi = \{\pi_1,...,\pi_k\}$ называется покрывающей тогда и только тогда, когда выполнено следующее условие: $\forall \bar{S}^d \ \exists\ \bar{\mu}^d : \Pi(\bar{S}^d, \bar{\mu}^d) = 1$, то есть когда для любой конфигурации $K \quad M^d$ существует хотя бы одна подходящая разметка.

Произвольное конечное множество пар вида $H = \{(\bar{S}_i^{d_i}, \bar{\mu}_i^{d_i}) : \bar{S}_i^{d_i} \in K^{d_i}; \bar{\mu}_i^{d_i} \in M_\Delta^{d_i}; i = 1,...,q\}$ называется набором прецедентов.

Задача Z выделения трендов заключается в синтезе подходящего алгоритма A, такого, что при всех $i = 1,...,q$ полная разметка $A(\bar{S}_i^{d_i})$ является продолжением разметки $\bar{\mu}_i^{d_i}$, или, иначе говоря, такого, что выполнено условие:

$$\forall j \ \underset{\{1,...,d_i\}}{} (\mu_i^j \in \bar{\mu}_i^{d_i}) \Rightarrow ((\mu_i^j \neq \Delta) \Rightarrow (\mu_i^j = \gamma_i^j)), \qquad (1)$$

где $\bar{\gamma}_i^{d_i} = A(\bar{S}_i^{d_i})$.

Из того, что алгоритм A подходящий следует, что для всех $\bar{\gamma}_i^{d_i} = A(\bar{S}_i^{d_i})$ выполнено условие: $\Pi(\bar{S}_i^{d_i}, \bar{\gamma}_i^{d_i}) = 1$.

Алгоритм A, удовлетворяющий условию (1) будем называть корректным для задачи Z.

Задача выделения трендов Z называется разрешимой тогда и только тогда, когда для нее существует подходящий корректный алгоритм A.

В данной главе рассматривается задача синтеза корректного алгоритма для разметки произвольных конфигураций. Для неё разработана сплит-модель, позволяющая практически решать задачу выделения трендов. Данная модель содержит в себе два параметрических семейства алгоритмов A^l и A^w.

Для проведения одной из серий экспериментов, которая будет описана ниже, был зафиксирован словарь разметки $M = \{\min, \max, \text{non}\}$ и система П из четырёх аксиом:

A1: $\underset{\{2,\ldots,d-1\}}{\forall i} : (v_{i-1} < v_i) \wedge (v_{i+1} > v_i) \Rightarrow \mu_i \neq \text{"min"} \wedge \mu_i \neq \text{"max"}$

A2: $\underset{\{2,\ldots,d-1\}}{\forall i} : (v_{i-1} > v_i) \wedge (v_{i+1} < v_i) \Rightarrow \mu_i \neq \text{"min"} \wedge \mu_i \neq \text{"max"}$

A3: $\underset{\{2,\ldots,d-1\}}{\forall i} : (v_{i-1} > v_i) \wedge (v_{i+1} > v_i) \Rightarrow \mu_i \neq \text{"max"}$

A4: $\underset{\{2,\ldots,d-1\}}{\forall i} : (v_{i-1} < v_i) \wedge (v_{i+1} < v_i) \Rightarrow \mu_i \neq \text{"min"}$

Первое параметрическое семейство алгоритмов A^l является однопараметрическим с параметром δ, в результате его работы формируется новая конфигурация, полученная «вычёркиванием» некоторых точек исходной, это можно представить в виде ломаной с вершиной в некоторых точках конфигурации. Семейство устроено следующим образом:

1) фиксируется стартовая точка с индексом $i = 1$, она же отмечается как первая вершина ломаной;
2) задаётся k (начальное значение равно 2), через точки конфигурации с индексами i и $i+k$ проводится прямая l;
3) вычисляются расстояния d_j, где $i < j < i+k$, от точек с индексами j до прямой l;
4) если все $d_j \leq \delta$, то k увеличивается на единицу и осуществляется переход к пункту 2, если в конфигурации остались нерассмотренные точки;
5) если хотя бы одно $d_j > \delta$, то новой стартовой точкой становится точка с индексом $i+k-1$, она теперь рассматривается как точка с индексом i, она же отмечается как очередная вершина ломаной, k присваивается его начальное значение и осуществляется переход к пункту 2, если в конфигурации остались нерассмотренные точки;
6) последняя точка конфигурации отмечается как очередная вершина ломаной;
7) точки конфигурации, которые являются вершинами ломаной, размечаются в соответствии с их взаимным расположением, то есть, если

$(v_{i-1} > v_i) \wedge (v_{i+1} > v_i)$, то точка с индексом i размечается как «min», если $(v_{i-1} < v_i) \wedge (v_{i+1} < v_i)$, то точка с индексом i размечается как «max», всем остальным точкам присваивается метка «non».

Второе параметрическое семейство A^W имеет несколько параметров: $\omega, \sigma, v_1, ..., v_r, r \geq 1$, в рассматриваемом ниже случае $r = 3$ (по количеству меток в словаре разметки). Данное семейство устроено следующим образом:

1) для каждой точки конфигурации формируется вектор $\overline{m}_i$ размерности ω, который будет заполняться значениями в процессе работы алгоритма;
2) фиксируется ω первых идущих подряд точек конфигурации;
3) среди зафиксированных точек проводится поиск минимума и максимума;
4) для найденных минимума и максимума в соответствующие векторы $\overline{m}_i$ добавляются метки «min» и «max», для остальных точек в векторы $\overline{m}_i$ добавляются метки «non»;
5) происходит смещение «окна» выделенных точек на σ точек вправо, если в конфигурации ещё остались нерассмотренные точки, таким образом, фиксируются очередные ω точек, осуществляется переход к пункту 3;
6) после того, как все точки конфигурации рассмотрены, и векторы $\overline{m}_i$ заполнены значениями меток, производится вычисление значений координат векторов $\overline{w}_i$ размерности 3 (по количеству меток в словаре разметки), которые сформированы для каждой точки конфигурации, по следующей схеме: первая координата – это произведение v_1 и количества меток «min» в соответствующем по индексу векторе $\overline{m}_i$, вторая – произведение v_2 и количества меток «max» и так далее.
7) в каждом векторе $\overline{w}_i$ производится поиск максимального значения координаты и в соответствии с этим каждой точке конфигурации присваивается метка, т.е. если наибольшее значение имеет первая координата вектора $\overline{w}_i$, то соответствующая точка размечается как «min», если наибольшее значение имеет вторая координата, то соответствующая точка размечается как «max» и так далее.

На основании вышеизложенного материала разработана компьютерная программа. В качестве входных данных используются наборы прецедентов, то есть КПК и разметки этих конфигураций экспертом, выходными данными является разметка алгоритма и некоторая статистическая информация. Схема работы этой программы приведена в рассматриваемой главе.

В соответствии с теорией, разработанной К.В. Рудаковым и Ю.В. Чеховичем, в данной главе рассмотрены требования к семействам алгоритмов, выполнение которых обеспечивало бы полноту этих семейств.

Вводится набор $\Pi=\{\pi_1,...,\pi_k\}$ предикатов $\pi_i:\mathfrak{I}_i\times\mathfrak{I}_f\to\{0,1\}$ для формализации понятия теоретико-множественных ограничений.

Пусть I_i – произвольный элемент пространства $\mathfrak{I}_i$. Положим $\Pi(I_i)=\{I_f \mid I_f\in\mathfrak{I}_f, \underset{1...k}{\forall} j:\pi_j(I_j,I_f)=1\}$ – множество всех допустимых значений корректных алгоритмов для начальной информации I_i.

Определение 2.3.1. Множество

$$PREC=\{((I_i^1,...,I_i^q),(I_f^1,...,I_f^q)) \mid$$

$$\mid q\in\mathbb{N},(I_i^1,...,I_i^q)\in\mathfrak{I}_i^q, I_i^j\neq I_i^k \text{ при } j\neq k,(I_f^1,...,I_f^q)\in\mathfrak{I}_f^q,$$

$$I_f^j\in\Pi(I_i^j) \text{ при } j=1,...,q\}$$

Определение 2.3.2. Модель $\mathfrak{M}$ называется Π-полной, если выполнены условия (2) и (3):

$$\underset{\mathfrak{I}_i}{\forall} I_i:\mathfrak{M}(I_i)=\{A(I_i) \mid A\in\mathfrak{M}\}\subseteq\Pi(I_i); \tag{2}$$

$$\underset{PREC}{\forall}((I_i^1,...,I_i^q),(I_f^1,...,I_f^q))\ \underset{\mathfrak{M}}{\exists} A:\ \underset{\{1,...,q\}}{\forall} j:A(I_i^j)=I_f^j. \tag{3}$$

Определение 2.3.3. Семейство решающих правил $\mathfrak{M}^1$ называется Π-полным, если существуют модель алгоритмических операторов $\mathfrak{M}^0$ и семейство корректирующих операций $\mathfrak{F}$ такие, что модель $\mathfrak{M}=\bigcup\limits_{\lambda\in L}\bigcup\limits_{\omega\in W(\lambda)}\mathfrak{M}^1\circ\mathfrak{F}^\lambda(\mathfrak{M}^0_{\lambda,\omega})$ является Π-полной.

Определение 2.3.4. При фиксированном Π-полном семействе решающих правил $\mathfrak{M}^1$ семейство корректирующих операций $\mathfrak{F}$ называется $\mathfrak{M}^1-\Pi$-полным, если существует модель алгоритмических операторов $\mathfrak{M}^0$ такая, что модель $\mathfrak{M}=\bigcup\limits_{\lambda\in L}\bigcup\limits_{\omega\in W(\lambda)}\mathfrak{M}^1\circ\mathfrak{F}^\lambda(\mathfrak{M}^0_{\lambda,\omega})$ является Π-полной.

Определение 2.3.5. При фиксированных Π-полном семействе решающих правил $\mathfrak{M}^1$ и $\mathfrak{M}^1-\Pi$-полном семействе корректирующих операций $\mathfrak{F}$ модель алгоритмических операторов $\mathfrak{M}^0$ называется $\mathfrak{F}-\mathfrak{M}^1-\Pi$-полной, если модель $\mathfrak{M}=\bigcup\limits_{\lambda\in L}\bigcup\limits_{\omega\in W(\lambda)}\mathfrak{M}^1\circ\mathfrak{F}^\lambda(\mathfrak{M}^0_{\lambda,\omega})$ является Π-полной.

Рассмотрим непустое семейство решающих правил $\mathfrak{M}^1=\bigcup\limits_{p=0}^{\infty}\mathfrak{M}^1_p$,

где при любом p из $\mathbb{N}_0$ выполнено соотношение $\mathfrak{M}^1_p\subseteq\{C \mid C:\mathfrak{I}_e^p\to\mathfrak{I}_f\}$. При этом для любого $X\subseteq\mathfrak{I}_e$ оказывается, естественно, выполненным условие

$$\mathfrak{M}^1(X)=\bigcup_{p=0}^{\infty}\mathfrak{M}^1_p(X^p)=\bigcup_{p=0}^{\infty}\bigcup_{C\in\mathfrak{M}^1_p}\bigcup_{\bar{x}\in X^p}C(\bar{x}).$$

Определение 2.3.6. Пусть $p\in\mathbb{N}_0$. Для произвольного I_i из $\mathfrak{I}_i$ множеством $\alpha_p(\mathfrak{M}^1,I_i)$ называется пересечение в p-ой декартовой степени пространства оценок $\mathfrak{I}_e$ всех полных прообразов множества $\Pi(I_i)$ относительно решающих правил арности p:

$$\alpha_p(\mathfrak{M}^1,I_i)=\bigcap_{C\in\mathfrak{M}^1_p}C^{-1}(\Pi(I_i))=\{I_e|\ I_e\in\mathfrak{I}_e^p,\ \underset{\mathfrak{M}^1_p}{\forall}C:C(I_e)\in\Pi(I_i)\}. \qquad (4)$$

Определение 2.3.7. Пусть $p\in\mathbb{N}_0$. Для семейства $\mathfrak{M}^1$ и элемента I_i пространства $\mathfrak{I}_i$ подмножество $X(I_i)$ пространства оценок $\mathfrak{I}_e$ называется допустимой p-проекцией, если выполнены условия (5) и (6):

$$X(I_i)^p\subseteq\alpha_p(\mathfrak{M}^1,I_i); \qquad (5)$$

$$\neg\exists Z\subseteq\mathfrak{I}_e:(X(I_i)\subset Z)\wedge(Z^p\subseteq\alpha_p(\mathfrak{M}^1,I_i)). \qquad (6)$$

Множество всех допустимых p-проекций для семейства $\mathfrak{M}^1$ и элемента I_i обозначим $\xi_p(\mathfrak{M}^1,I_i)$.

Для произвольного I_i из $\mathfrak{I}_i$ введем множество $\Phi(\mathfrak{M}^1,I_i)$ функций выбора допустимых проекций:

$$\Phi(\mathfrak{M}^1,I_i)=\{\varphi|\ \varphi:\mathbb{N}_0\to B(\mathfrak{I}_e),\underset{\mathbb{N}_0}{\forall}p:((\mathfrak{M}^1_p=\varnothing)\Rightarrow\varphi(p)=\mathfrak{I}_e)\wedge$$

$$\wedge((\mathfrak{M}^1_p\neq\varnothing)\Rightarrow(\varphi(p)\in\xi_p(\mathfrak{M}^1,I_i)))\},$$

где $B(\mathfrak{I}_e)$ – множество всех подмножеств множества $\mathfrak{I}_e$.

Для каждой функции выбора допустимых проекций φ из $\Phi(\mathfrak{M}^1,I_i)$ положим $X(I_i,\varphi)=\bigcap_{p=0}^{\infty}\varphi(p)$. Отметим, что

$$\mathfrak{M}^1(X(I_i,\varphi))=\bigcup_{r=0}^{\infty}\bigcup_{C\in\mathfrak{M}^1_r}C((\bigcap_{p=0}^{\infty}\varphi(p))^r).$$

Пусть $\tilde{\Phi}(\mathfrak{M}^1,I_i)=\{\varphi|\ \varphi\in\Phi(\mathfrak{M}^1,I_i),X(I_i,\varphi)\neq\varnothing\}$.

На приведённых выше определениях К.В. Рудакова и Ю.В. Чеховича базируются две предложенные ими следующие теоремы.

Теорема 2.3.1. При всех I_i из $\mathfrak{I}_i$ выполнено соотношение (7):

$$\bigcup_{\varphi\in\tilde{\Phi}(\mathfrak{M}^1,I_i)}\mathfrak{M}^1(X(I_i,\varphi))\subseteq\Pi(I_i). \qquad (7)$$

Теорема 2.3.2. (Критерий П-полноты для семейств решающих правил). Для П-полноты семейства решающих правил $\mathfrak{M}^1$ необходимо и достаточно, чтобы при любом I_i из $\mathfrak{I}_i$ было выполнено условие (8):

$$\bigcup_{\varphi\in\tilde{\Phi}(\mathfrak{M}^1,I_i)}\mathfrak{M}^1\,(X(I_i,\varphi))=\Pi(I_i). \qquad (8)$$

Далее во второй главе диссертации приведены определения и теоремы, сформулированные и доказанные автором данного исследования.

Рассмотрены так называемые фильтрующие алгоритмы выделения трендов временных рядов. Далее подразумевается использование словаря разметки $\{\min,\max,non\}$.

Определение 2.4.1. Назовём алгоритм разметки A_δ фильтрующим с параметром δ, если результатом его работы является новая КПК S^l и множество индексов Ind_d^l.

Для фильтрующего алгоритма A_δ имеем следующее (определение 2.3.6):

$$\alpha_1(\mathfrak{M}^1,I_i)=C^{-1}(\Pi(I_i)).$$

Теорема 2.4.1. Для A_δ верно, что $C(C^{-1}(\Pi(I_i)))=\Pi(I_i)$.

Определение 2.4.2. Назовём алгоритм $A_{\delta(t)}$ фильтрующим с переменным параметром, если этот алгоритм является фильтрующим по определению 2.4.1 и использует новый параметр на каждой итерации.

Теорема 2.4.2. Семейство алгоритмов $\{A_{\delta(t)}\}$ полно.

В **третьей главе** диссертации проводится экспериментальное исследование методов синтеза алгоритмов выделения трендов, даётся описание проведённых экспериментов.

В этой главе рассматриваются результаты работы нашей программы на пяти различных КПК. Данные для конфигураций – курсы валют, опубликованные Центробанком РФ.

Далее во всех случаях под верной разметкой точки конфигурации алгоритмом подразумевается, что алгоритм разметил точку так же, как эксперт. В том случае, если эксперт не разметил точку конфигурации, то любая её разметка алгоритмом считается верной, если она удовлетворяет аксиомам разметки.

Первая КПК:

$\delta=0{,}03289;\ \omega=5;\ \sigma=1;\ \nu_1=0{,}2;\ \nu_2=0{,}1;\ \nu_3=0{,}5.$

Первый базовый алгоритм верно разметил 84 из 96 точек КПК, функционал качества второго алгоритма, рассмотренный в пункте 5 описания параметрического семейства A^W, равен 18,8, результат работы третьего результирующего алгоритма 89 из 96 верно размеченных точек, $\alpha=0{,}05$.

Улучшение качества работы алгебраической композиции алгоритмов относительно первого базового здесь составляет 5,21%.

Вторая КПК:

$\delta = 0{,}06927;\ \omega = 3;\ \sigma = 1;\ \nu_1 = 0{,}2;\ \nu_2 = 0{,}1;\ \nu_3 = 0{,}1.$

Первый базовый алгоритм верно разметил 270 из 285 точек КПК, функционал качества второго алгоритма, рассмотренный в пункте 5 описания параметрического семейства A^W, равен 26,5, результат работы третьего результирующего алгоритма 273 из 285 верно размеченных точек, $\alpha = 0{,}55$. Улучшение качества работы алгебраической композиции алгоритмов относительно первого базового здесь составляет 1,05%.

Третья КПК:

$\delta = 0{,}34436;\ \omega = 7;\ \sigma = 1;\ \nu_1 = 0{,}3;\ \nu_2 = 0{,}3;\ \nu_3 = 0{,}2.$

Первый базовый алгоритм верно разметил 78 из 87 точек КПК, функционал качества второго алгоритма, рассмотренный в пункте 5 описания параметрического семейства A^W, равен 14,6, результат работы третьего результирующего алгоритма 83 из 87 верно размеченных точек, $\alpha = 0{,}05$. Улучшение качества работы алгебраической композиции алгоритмов относительно первого базового здесь составляет 5,74%.

Далее рассмотрим результаты экспериментов, полученные на наборах прецедентов, каждый из которых состоит из двух КПК.

Первый набор:

$\delta = 0{,}23571;\ \omega = 6;\ \sigma = 1;\ \nu_1 = 0{,}3;\ \nu_2 = 0{,}2;\ \nu_3 = 0{,}2.$

Первый базовый алгоритм верно разметил 159 из 183 точек набора, функционал качества второго алгоритма, рассмотренный в пункте 5 описания параметрического семейства A^W, равен 30,5, результат работы третьего результирующего алгоритма 175 из 183 верно размеченных точек, $\alpha = 0{,}35$. Улучшение качества работы алгебраической композиции алгоритмов относительно первого базового здесь составляет 8,74%.

Второй набор:

$\delta = 0{,}09224;\ \omega = 8;\ \sigma = 1;\ \nu_1 = 0{,}5;\ \nu_2 = 0{,}3;\ \nu_3 = 0{,}4.$

Первый базовый алгоритм верно разметил 652 из 719 точек набора, функционал качества второго алгоритма, рассмотренный в пункте 5 описания параметрического семейства A^W, равен 89,8, результат работы третьего результирующего алгоритма 702 из 719 верно размеченных точек, $\alpha = 0{,}15$. Улучшение качества работы алгебраической композиции алгоритмов относительно первого базового здесь составляет 6,95%.

Проведённые эксперименты показывают, что улучшение качества работы алгебраической композиции относительно первого базового алгоритма (отметим, что аналогичное улучшение наблюдается и относительно второго базового алгоритма) не зависит от размера КПК или их набора, а также от характера действий эксперта при их разметке (при

условии, что разметка корректна с точки зрения используемых аксиом). Это позволяет говорить о получении стабильных результатов в контексте задачи синтеза обучаемых алгоритмов выделения трендов временных рядов.

Далее в этой главе рассматривается работа экспериментальной программы в схемах и графиках.

В **заключении** диссертации приведены основные результаты, полученные в работе.

В **приложение** вынесен исходный текст экспериментальной программы, разработанной автором диссертации и описанной во второй главе реферируемой работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

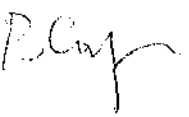
1. Построены эффективные в рамках рассматриваемого класса задач выделения трендов временных рядов параметрические семейства алгоритмических операторов, решающих правил и корректирующих операций.
2. Изложены требования к семействам алгоритмов, выполнение которых обеспечивает их полноту, доказаны теоремы о полноте.
3. Разработана сплит-модель, позволяющая практически решать задачу выделения трендов.
4. Разработана программа для синтеза обучаемых алгоритмов выделения трендов временных рядов на основе предложенной модели, описан принцип её работы.
5. Описаны серии проведённых экспериментов с использованием разработанной программы, на схемах подробно представлены принципы её работы. Сделаны выводы о результатах экспериментов.

Основное содержание диссертации отражено в работах:

1. **Sarapas V.V. Experimental Study of Synthesis Methods of Algorithms of Trends Allocation. // Pattern Recognition and Image Analysis. A Journal of Russian Academy of Sciences. City of Dover: Pleiades Publishing, c/o МАИК «Наука / Interperiodica», 2010. Vol. 20. №2. P. 145 – 151. – 0,75 п.л.**
2. Сарапас В.В. Параметрические семейства алгоритмических операторов для решения задач выделения трендов. // Высокие технологии, исследования, промышленность. Сборник трудов девятой международной научно-практической конференции. Санкт-Петербург:

Изд-во Политехнического университета, 2010. Т. 3. С. 124 – 125. – 0,25 п.л.

3. Сарапас В.В. О классификации элементов временных рядов. // Математика, информатика, физика и их преподавание. М.: МПГУ, 2009. С. 163 – 164. – 0,2 п.л.
4. Сарапас В.В. Алгебраический подход к задаче выделения трендов временных рядов. // Молодежь и наука XXI века. Материалы XI Всероссийской научно-практической конференции студентов, аспирантов и молодых учёных с международным участием. Красноярск: Изд-во КГПУ, 2010. Т. 1. С. 229 – 233. – 0,5 п.л.
5. Сарапас В.В. Об алгебраических методах синтеза алгоритмов выделения трендов временных рядов. // Наука и современность – 2010. Сборник материалов II Международной научно-практической конференции. Новосибирск: Изд-во «СИБПРИНТ», 2010. Ч. 3. С. 77 – 82. – 0,5 п.л.



Подп. к печ. 20.12.2010 Объем 1 п.л. Заказ № 148 Тир 100 экз.
Типография МПГУ