На правах рукописи

БУЗМАКОВ
Алексей Владимирович

МОДЕЛИРОВАНИЕ ПРОЦЕССОВ С СОСТОЯНИЯМИ СЛОЖНОЙ СТРУКТУРЫ НА ОСНОВЕ РЕШЁТОК ЗАМКНУТЫХ ОПИСАНИЙ

Специальность 05.13.18 —
Математическое моделирование, численные методы и комплексы программ
(технические науки)

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата технических наук

Москва – 2014

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики»

Научный руководитель:	Кузнецов Сергей Олегович, доктор физико-математических наук, заведующий кафедрой анализа данных и искусственного интелекта Национального исследовательского университета «Высшая школа экономики»
Официальные оппоненты:	Вагин Вадим Николаевич, доктор технических наук, профессор кафедры прикладной математики ФГОБУ ВПО Национального исследовательского университета Московский Энергетический Институт
	Виноградов Дмитрий Вячеславович, кандидат физико-математических наук, старший научный сотрудник отдела теоретических и прикладных проблем информатики ФГБУН Всероссийского института научной и технической информации
Ведущая организация:	ФГБУН Институт проблем управления им. В.А. Трапезникова

Защита состоится «16» февраля 2015 года в 13:00 на заседании диссертационного совета Д 212.048.09, созданного в «Национальном исследовательском университете «Высшая школа экономики» по адресу: 105187, г. Москва, ул. Кирпичная, д.33, ауд. – 503.

С диссертацией можно ознакомиться в библиотеке Национального исследовательского университета «Высшая школа экономики» по адресу: 101990, г. Москва, ул. Мясницкая, д. 20, и на сайте `http://www.hse.ru/sci/diss/`.

Автореферат разослан «__» октября 2014 г.

Ученый секретарь
диссертационного совета
доктор технических наук, профессор Назаров Станислав Викторович

Общая характеристика работы

Актуальность темы. Многие явления окружающего мира могут быть представлены процессами. Например, развитие болезни пациента характеризуется изменением состояния больного, происходящего в том числе под воздействием применяемого подхода к его лечению. При моделировании процессов только некоторые особенности состояний могут быть зафиксированы и обработаны. Последовательности зафиксированных состояний называются реализациями процесса, а всё множество доступных реализаций процесса называется его логом. Таким образом, располагая логами процесса, можно построить его модель, которая может быть использована для решения различных задач, таких как экспертный анализ процесса, а также классификация и кластеризация новых реализаций процесса. В этой работе мы фокусируемся на задаче автоматического построения модели процесса для её последующего анализа экспертом. Например, для задачи анализа процесса госпитализаций пациентов одна реализация может состоять из состояний, каждое из которых описывается рядом параметров, таких как, расположение больницы, применяемые медицинские процедуры, продолжительность госпитализации и др. Модель, построенная по логу процесса госпитализации, может быть использована экспертом для различных стоящих перед ним задач, таких как оптимизация процесса лечения по качеству или скорости, поиск систематических ошибок процесса.

За последнее десятилетие было проведено много исследования моделей процессов и методов их построения с акцентом на события, такие как сети Петри, Yawl и другие (W. van der Aalst, 2011). В этих работах под событиями понимаются переходы из одного состояния в другое. Структура же самих событий в этих работах не рассматривается. В этой работе мы фокусируемся именно на состояниях процесса, в то время как переходы между ними рассматриваются просто как временные отношения следования. Более того, в этой работе мы рассматриваем процессы с состояниями сложной структуры, что подчёркивает тот факт, что одно состояние описывается многими параметрами разной природы. В частности, для процесса госпитализации пациентов одно состояние описывается среди прочих параметров не просто именем боль-

ницы, но также и таксономией больниц по территориальному расположению – древовидной структурой, задающей отношение частное-общее, в которой также присутствуют больницы «обобщённого» вида, такие как «все больницы определённого города». Это позволяет включить в модель как можно более полный лог процесса, что помогает строить модель процесса госпитализации с «обобщением» некоторых параметров, и, таким образом, предоставляет эксперту возможность более подробно исследовать процесс госпитализации.

Для моделей таких процессов могут быть применены различные подходы (M. Plantevit et al., 2010; E. Egho et al., 2014), которые ищут частые последовательности в логе процесса, при этом структура одного элемента последовательности может включать несколько компонент, для каждой из которых задана таксономия. К сожалению, в этих подходах нельзя передать такую важную информацию, как количество повторений определённой процедуры, например, химиотерапии. И, более того, эффективность этих подходов невысокая, при условии, что многие из порождаемых элементарных моделей имеют малую пользу для эксперта. Другим возможным подходом к анализу таких данных является работа (S. Tsumoto, 2014), в которой авторы на основе статистических методов отображают данные, собранные в больницах. Однако данная работа не фокусируется на последовательных зависимостях, и в ней такие зависимости имеют очень грубое представление, привязанное к конкретным временам суток.

Специфика процессов с состояниями сложной структуры требует специального типа моделей, способного представлять средства выражения сходства и различия реализаций процесса с состояниям сложной структуры. В качестве математического аппарата построения моделей процессов с состояниями сложной структуры используются решетки замкнутых описаний, представляемые так называемыми узорными структурами (B. Ganter & S.O. Kuznetsov, 2001), дополненные средствами приближенного описания - проекциями описаний, позволяющим моделировать информацию о большом количестве реализаций процессов с помощью существующих эффективных алгоритмов. Этот математический аппарат позволяет находить классы реализаций процессов схожих между собой, с получением соответствующего сходства. Найденные классы реализаций процессов упорядочены по включению соответствующих

множеств реализаций в каждом классе и формируют так называемую иерархическую модель процесса. Такая модель состоит из множества элементарных моделей процесса, каждая из которых описывает процесс на определённом уровне абстракции, учитывающем только часть информации доступную в логах процесса.

Таким образом, **объектом исследования** являются различные процессы с состояниями сложной структуры. **Предметом исследования** являются математическая модель, алгоритмы её построения и комплекс программ анализа процессов с состояниями сложной структуры с целью экспертного анализа этого процесса для его оптимизации и поиска ошибок.

Целью диссертационного исследования является разработка подходов к построению моделей процессов с состояниями сложной структуры на основе решёток замкнутых описаний. Модели должны строиться за приемлемое время и иметь размер адекватный для экспертного анализа.

В соответствии с целью исследования были поставлены следующие **задачи**:

1. Предложить иерархическую модель процессов с состояниями сложной структуры, которую можно построить за приемлемое время, с целью дальнейшего экспертного анализа.

2. Предложить адекватную, эффективно вычислимую меру качества элементарных моделей процессов с состояниями сложной структуры с целью уменьшения сложности иерархической модели таких процессов.

3. Разработать комплекс программ для анализа процессов с состояниями сложной структуры на основе предложенной модели и апробировать его на данных о процессах госпитализации пациентов.

Следующие особенности работы определяют её **научную новизну**:

1. Предложен класс моделей на основе узорных структур для исследования процессов с состояниями сложной структуры.

2. Подход к моделированию на основе решеток замкнутых описаний был обобщён на более широкий класс проекций описаний, имеющих высокую практическую значимость. Это позволило автоматически строить

модели введённого класса за меньшее время, чем при использовании альтернативных подходов.

3. Впервые для измерения качества элементарных моделей была экспериментально проверена на широкой тестовой базе данных возможность применения меры качества по устойчивости.

4. Были предложены две эффективные оценки устойчивости, которые имеют лучшие вычислительные характеристики и точность, чем существующие аналоги.

5. Впервые создан комплекс программ, позволяющий разрабатывать модели на основе решеток замкнутых описаний, в рамках которого была реализована модель процессов с состояниями сложной структуры. Эта модель была апробирована на процессе госпитализации пациентов.

Теоретическая ценность данной работы состоит

1. в расширении и уточнении аппарата узорных структур для моделирования процессов с состояниями сложной структуры;

2. во введении и исследовании проекции минимальной длины и алфавитной проекции моделей процессов с состояниями сложной структуры;

3. во введении и исследовании эффективных оценок устойчивости элементарных моделей на основе замкнутых описаний.

Практическая ценность работы состоит

1. в разработке класса моделей процессов с состояниями сложной структуры;

2. в разработке быстрых алгоритмов вычисления мер качества элементарных моделей при создание иерархических моделей на основе аппарата узорных структур;

3. в получении значимых результатов исследования процессов госпитализаций для оптимизации процессов лечения;

4. в разработке эффективного программного комплекса, который предоставляет возможность использования методов решеток замкнутых описаний для построения моделей процессов с состояниями сложной структуры.

Положения, выносимые на защиту:

1. Предложен класс иерархических моделей процессов с состояниями сложной структуры и вычислительные методы автоматического построения таких моделей по логу процессов за приемлемое время с учетом специфики предметной области.

2. Аппарат узорных структур и их проекций был расширен на более широкий класс проекций.

3. Экспериментально обоснована возможность применения меры качества устойчивости для выделения важных элементарных моделей в задаче упрощения разработанной иерархической модели процессов с состояниями сложной структуры.

4. Введены две эффективные оценки меры качества устойчивости и экспериментально обоснована возможность их успешного применения.

5. Разработан комплекс программ для анализа процессов с состояниями сложной структуры на основе решеток замкнутых описаний (узорных структур), успешно апробированный в задаче анализа госпитализаций. Разработанный комплекс программ включен в программную систему FCART.

Достоверность полученных результатов опирается на строгость использованных математических моделей и на их экспериментальное подтверждение.

Апробация работы. Основные результаты работы обсуждались и докладывались на следующих конференциях:

1. Симпозиум BioIntelligence 2012, Софи-Антиполис, Франция;

2. Первый международный семинар «Что АФП может сделать для искусственного интеллекта?» (Workshop: What can FCA do for Artificial Intelligence?), 2012, Монпелье, Франция;

3. Девятая международная конференция по решёткам понятий и их приложениям (The Ninth International Conference on Concept Lattices and Their Applications), 2012, Малага, Испания;

4. Симпозиум BioIntelligence 2013, Софи-Антиполис, Франция;

5. Второй международный семинар «Что АФП может сделать для искусственного интеллекта?» (Workshop: What can FCA do for Artificial Intelligence?), 2013, Пекин, Китай (два доклада);

6. Семинар на ECML/PKDD 2013: языки для анализа данных и машинного обучения (ECML/PKDD 2013 Workshop: Languages for Data Mining and Machine Learning), 2013, Прага, Чехия

7. Десятая международная конференция по решёткам понятий и их приложениям (The Tenth International Conference on Concept Lattices and Their Applications), 2013, Ла-Рошель, Франция;

8. Вторая международная конференция по информационным технологиям и численному управления (The Second International Conference on Information Technology and Quantitative Management), 2014, Москва, Россия.

9. Двенадцатая международная конференция по анализу формальных понятий (12th International Conference on Formal Concept Analysis), 2014, Клуж-Напока, Румыния;

Публикации. Основные результаты по теме диссертации изложены в 8 научных работах, 2 из которых изданы в изданиях, рекомендованных ВАК, 6 — в рецензируемых трудах международных конференций.

Структура и объем диссертации. Диссертация состоит из введения, пяти глав, заключения и списка литературы. Общий объем работы — 154 страницы. Список литературы включает 140 названий.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи работы, сформулированы научная новизна и практическая значимость представляемой работы.

Первая глава посвящена обзору подходов по моделированию объектов со сложной структурой, прежде всего процессов с состояниями сложной структуры. В частности рассматриваются классические подходы анализа процессов, которые получают такие модели процессов как сети Петри и системы переходов. Также рассматриваются некоторые иерархические модели, представленные иерархией простых моделей. В этой главе устанавливается, что иерархические модели могут быть получены методами анализа данных, как объединение найденных закономерностей, или элементарных моделей, в одну иерархию. К таким моделям относятся модели последовательностей и графовые модели. В главе показывается, что для обработки процессов с состояниями сложной структуры необходимо создание нового подхода для построения моделей таких процессов. В первой главе также отмечается, что среди моделей последовательностей есть несколько подходов, которые могут быть применены к исследованию процессов с состояниями сложной структуры (M. Plantevit et al., 2010; E. Egho et al., 2014). Графовые модели в работе рассматриваются как способ представление таких процессов, однако методы анализа графовой информации требуют большого времени вычислений и потому не применимы для анализа процессов с состояниями сложной структуры.

Также в этой главе приводится обзор методов по выбору наиболее важных элементарных моделей для уменьшения размера общей иерархической модели. В силу небольшого количества подходов к выбору элементарных моделей процессов, приводятся меры качества более простых закономерностей, таких как описываемых множествами признаков. В главе заключается, что мера устойчивости является оптимальным выбором и позволяет обрабатывать элементарные модели процессов. Также в главе предполагается необходимость сравнения устойчивости с другими мерами качества при ранжировании закономерностей разного вида.

Во **второй главе** подробно разбираются основные понятия анализа формальных понятий и узорных структур, на которых основана предлагаемая иерархическая модель процессов с состояниями сложной структуры. Затем вводится сама модель таких процессов. Эта модель является трудновычислимой и поэтому её необходимо редуцировать. Далее вводится авторское расширение проекций узорных структур и описываются его свойства. В конце данной главы определяются специальные проекции для модели процессов с состояниями сложной структуры, имеющие высокое практическое значение.

Узорные структуры являются расширением анализа формальных понятий для работы со структурными данными, такими как данные, описываемые численными значениями, множествами последовательностей или графов. *Узорной структурой* называется тройка $(G, (D, \sqcap), \delta)$, где G – множество объектов, $(D, \sqcap)$ – полная полурешётка всевозможных описаний, а $\delta\colon G \to D$ – функция, сопоставляющая объекту из множества G его описание из D.

Полурешёточная операция $\sqcap$ соответствует операции сходства между двумя описаниями, определяемой для разных типов описаний в работах Г.К. Финна, С.О. Кузнецова и др. Соответствие Галуа между множествами объектов и множеством описаний определяется следующим образом:

$$A^\diamond := \prod_{g \in A} \delta(g), \qquad \text{для } A \subseteq G$$
$$d^\diamond := \{g \in G \mid d \sqsubseteq \delta(g)\}, \quad \text{для } d \in D.$$

Здесь $\sqsubseteq$ – это отношение поглощения, однозначно задающееся через полурешёточную операцию как: $a \sqsubseteq b \Leftrightarrow a \sqcap b = a$.

Пример. *Каждая реализация процесса госпитализации пациентов соответствует последовательности госпитализаций одного пациента. В этом случае G – множество всех реализаций процесса доступных для анализа, D соответствует множеству всех множеств последовательностей смены состояний, а результат полурешёточной операции между $X, Y \in D$ на нём задаются как множество, последовательности которого являются более общими, чем последовательности множеств X и Y. Например, последовательность $\langle a, b, c\rangle$ является менее общей, чем последовательность $\langle a, b\rangle$, так как реже*

встречается в реализациях процессов. Функция δ в этом случае задаётся как соответствие между идентификаторами записей G и реальным описанием соответствующим этим записям из D.

Узорным понятием узорной структуры $(G, (D, \sqcap), \delta)$ называется пара (A, d), в которой $A \subseteq G$ – подмножество множества объектов, $d \in D$ – одно из описаний из полурешётки, такие что $A^\diamond = d$ и $d^\diamond = A$; A называется объёмом понятия, а d – узорным содержанием.

Для того, чтобы создать модель процессов с состояниями сложной структуры, нам потребуется определить узорную структуру на основе произвольного частичного порядка. Такой частичный порядок соответствует отношениям «часть–целое», «подкласс–класс». Пусть дан некоторый частичный порядок $(P, \leq)$, тогда соответствующая ему решётка $(D, \sqcap)$ задаётся как множество подмножеств P таких, что если $p \in P$ принадлежит элементу решётки $d \in D$, то все меньшие элементы также принадлежат этому элементу решётки, $\forall p \in d, \nexists q \in P, q \leq p : q \notin d$. При этом решёточной операцией является теоретико-множественная операция пересечения. Нетрудно заметить, что результат такой операции между $d_1, d_2 \in D$ даст некоторый элемент из D. Стоит отметить, что на практике множество $d \in D$ может иметь существенный размер и поэтому его эффективнее и осмысленнее представлять максимальными элементами данного множества, $\tilde{d} = \{p \in d \mid \nexists q \in d : q > p\}$.

Теперь мы можем определить первую версию иерархической модели процесса с состояниями сложной структуры. Эта модель основывается на последовательностях сложных элементов, для простоты называемые сложными последовательностями.

Определение 17. *Пусть дана полурешётка $(E, \sqcap)$, называемая алфавитом, тогда последовательностью элементов сложной структуры называется любой упорядоченный список элементов из E. Такая последовательность обозначается как $s = \langle e_1, \ldots, e_n \rangle$.*

Сложной такая последовательность называется потому, что алфавит такой последовательности является полурешёткой, в отличии от простого алфавита, в котором любые два элемента несравнимы, а также в отличии от алфавита

$(\wp(M), \cap)$, часто применяющегося в анализе данных, представленных последовательностями. Далее необходимо исключить те последовательности, которые содержат $\perp$ в качестве элемента, где $\perp = \bigsqcap_{e \in E} e$. Это делается по аналогии с анализом данных, представленных последовательностями, в котором исключаются все последовательности, включающие $\emptyset$.

Определение 18. *Последовательность $s = \langle e_1, \ldots, e_n \rangle$ является допустимой, если для любого i, $e_i \neq \perp$.*

Определение 19. *Последовательность $t = \langle t_1; ...; t_k \rangle$ является подпоследовательностью для последовательности $s = \langle s_1; ...; s_n \rangle$, что обозначается как $t \leq s$, тогда и только тогда, когда $k \leq n$ и существуют $j_1, ..j_k$ такие, что $1 \leq j_1 < j_2 < ... < j_k \leq n$, а также для любого $i \in \{1, 2, ..., k\}$, $t_i \sqsubseteq_E s_{j_i}$.*

Сложными последовательностями моделируются реализации процесса. Определение 19 задаёт частичный порядок на сложных последовательностях, и, таким образом, элемент полурешётки описаний на сложных последовательностей является множеством последовательностей. Такой подход задаёт решётку узорных понятий, которая получается по множеству реализаций. Однако, количество узорных понятий может быть большим. Для того, чтобы иметь возможность обрабатывать такие данные за приемлемое время, необходимо ввести проекции узорных структур, которые также помогают сократить количество ненужных элементарных моделей.

Проекцией полурешётки описаний $(D, \sqcap)$ называется функция $\psi : D \to D$, которая является оператором ядра: монотонной ($x \sqsubseteq y \Rightarrow \psi(x) \sqsubseteq \psi(y)$), сжимающей ($\psi(x) \sqsubseteq x$) и идемпотентной ($\psi(\psi(x)) = \psi(x)$).

Определение 16. *Проекция узорной структуры, полученная из узорной структуры $(G, (D, \sqcap), \delta)$ с помощью проекции ψ – эта такая узорная структура $(G_\psi, (D_\psi, \sqcap_\psi), \delta_\psi)$, в которой $G_\psi = G$, $D_\psi = \psi(D) = \{d \in D \mid d = \psi(d)\}$, с полурешёточной операцией $\forall x, y \in D, x \sqcap_\psi y := \psi(x \sqcap y)$, а $\delta_\psi = \psi \circ \delta$.*

По сравнению с классическим определением проекции узорной структуры (B. Ganter & S.O. Kuznetsov, 2001) , определение 16 также меняет полурешётку описаний в спроецированной узорной структуре. Это позволяет проецировать не только начальное описание объектов, но и полурешёточную операцию

сходства. Что позволяет вводить более широкий класс проекций, имеющий важное прикладное значение.

Здесь нам необходимо показать, что спроецированная узорная структура, является узорной структурой в смысле ранее ведённого определения, то есть что $(D_\psi, \sqcap_\psi)$ является полурешёткой. Это следует из следующего утверждения.

Утверждение 3. *Пусть дана полурешётка $(D, \sqcap)$ и проекция ψ, тогда для всех $x, y \in D$ выполняется $\psi(x \sqcap y) = \psi(\psi(x) \sqcap y)$.*

Следствие 2. *Пусть дана полурешётка $(D, \sqcap)$ и проекция ψ, тогда $(D_\psi, \sqcap_\psi)$, заданная определением 16, является полурешёткой, то есть $\sqcap_\psi$ является коммутативной, ассоциативной и идемпотентной операцией.*

С новым определением спроецированной решётки выполняются все свойства проекций, введённых B. Ganter & S.O. Kuznetsov (2001). В частности верно следующее утверждение, которое показывает связь между узорной структурой и её проекцией.

Утверждение 5. *Для каждого понятия в $(G, (D_\psi, \sqcap_\psi), \delta_\psi)$ существует понятие в $(G, (D, \sqcap), \delta)$ с таким же объёмом. Если d является содержанием в $(G, (D, \sqcap), \delta)$, тогда $\psi(d)$ является содержанием в $(G, (D_\psi, \sqcap_\psi), \delta_\psi)$, при этом $\psi(d)^{\diamond\diamond} \sqsubseteq d$.*

Введём теперь специальные проекции для узорных структур на последовательностях, имеющих важное прикладное значение при моделировании процессов с состояниями сложной структуры.

Определение 20. *Проекция минимальной длины (ПМД) $\ell_{\min}$ узорной структуры на сложных последовательностях задаётся следующей функцией:*

$$\psi(d) = \{s \in d \mid |s| \geq \ell_{\min}\},$$

где $d \in D$ – любой элемент полурешётки описания на сложных последовательностях.

Проекция в определении 20 позволяет ограничить минимальную длину рассматриваемых последовательностей, что позволяет исключить из рассмотрения короткие, заведомо бесполезные последовательности. Следующие определения 21 и 22 задают возможные изменения алфавита. В силу того, что алфавит является полурешёткой, к нему также можно применить проекцию. Такая проекция алфавита индуцирует проекцию узорной структуры на последовательностях. Алфавитная проекция позволяет упростить алфавит сложных последовательностей. В частности, если в алфавите есть несколько компонент, то некоторые из них могут быть исключены из рассмотрения.

Определение 21. *Пусть даны алфавит* $(E, \sqcap)$*, его проекция* ψ_E *и последовательность* $s = \langle e_1, \cdots, e_n \rangle$*, основанная на* E*, то есть* $e_i \in E$*. Тогда алфавитной проекцией последовательности* s *называется последовательность* $\psi_e(s) = \langle \psi_E(e_1), \cdots, \psi_E(e_n) \rangle$.

Определение 22. *Пусть дана узорная структура* $(G, (D, \sqcap), \delta)$ *на сложных последовательностях из* $\mathcal{S}$*, где* $\mathcal{S}$ *множество всех последовательностей основанных на* $(E, \sqcap)$*. Тогда алфавитной проекцией узорной структуры называется проекция, задаваемая следующей функцией от* $d \in D$*:*

$$\psi(d) = \{s \in \mathcal{S} \mid s \text{ – допустимая и } \exists \tilde{s} \in d : s \leq \psi_E(\tilde{s})\}.$$

Определения 20 и 22 задают функции, являющиеся монотонными, расширяющими и идемпотентными, то есть задают проекции узорной структуры (Утверждения 7 и 8).

Проекции узорных структур на последовательностях позволяют существенно уменьшить количество узорных понятий и повысить эффективность вычислений, не теряя при этом в качестве, что позволяет моделировать процессы с состояниями сложной структуры за приемлемое время. Однако, конечная иерархическая модель может содержать большое количество элементарных моделей, количество которых необходимо уменьшить для успешного применения этой модели при экспертном анализе.

Третья глава посвящена мерам качества элементарных моделей. В нашем случае каждое узорное понятие является простой моделью процесса, но количество понятий может быть большим. Поэтому необходимо выбирать наи-

более важные понятия при создания модели процесса. В этой главе изучается мера качества по устойчивости и вводятся эффективные методы её оценки.

Устойчивостью формального понятия $\mathtt{Stab}(\mathcal{C})$ называется отношение количества подмножеств объема понятия ($\mathtt{Ext}(\mathcal{C})$), описание которых совпадает с содержанием ($\mathtt{Int}(\mathcal{C})$), к количеству подмножеств понятия. Здесь и далее $\wp(P)$ означает множество всех подмножеств P.

$$\mathtt{Stab}(\mathcal{C}) := \frac{|\{s \in \wp(\mathtt{Ext}(\mathcal{C})) \mid s^{\diamond} = \mathtt{Int}(\mathcal{C})\}|}{2^{|\mathtt{Ext}(\mathcal{C})|}} \tag{1}$$

Устойчивость формального понятия показывает, насколько сильно содержание формального понятия зависит от выборки данных. Чем выше устойчивость, тем больше комбинаций объектов можно выбросить из формального контекста, не меняя содержание узорного понятия. Также можно доказать, что устойчивость формального понятия может только увеличиться при проецировании узорной структуры.

Утверждение 10. *Пусть дана узорная структура $(G, (D, \sqcap), \delta)$, и её проекция ψ. Пусть также дано понятие $\mathcal{C}_\psi$ в $(G, (D_\psi, \sqcap_\psi), \delta_\psi)$, тогда существует понятие $\mathcal{C}$ в $(G, (D, \sqcap), \delta)$, объём которого совпадает с объёмом $\mathcal{C}_\psi$, $\mathtt{Ext}(\mathcal{C}_\psi) = \mathtt{Ext}(\mathcal{C})$ (смотри утверждение 5), при этом устойчивость понятия $\mathcal{C}$ не превышает устойчивости понятия $\mathcal{C}_\psi$, $\mathtt{Stab}(\mathcal{C}) \leq \mathtt{Stab}(\mathcal{C}_\psi)$.*

Было показано ранее, что нахождение устойчивости для данного понятия является #P-полной задачей (С.О. Кузнецов, 1990). Более того, как показывают вычислительные эксперименты, нахождение индекса устойчивости может требовать существенно больше времени, чем вычисление самой решётки [5]. Значит, эффективная оценка этого индекса является необходимой для применения индекса устойчивости.

Утверждение 11. *Устойчивость формального понятия может быть оценена по формуле:*

$$1 - \sum_{\mathcal{D} \in \mathtt{DD}(\mathcal{C})} \frac{1}{2^{\Delta(\mathcal{C},\mathcal{D})}} \leq \mathtt{Stab}(\mathcal{C}) \leq 1 - \max_{\mathcal{D} \in \mathtt{DD}(\mathcal{C})} \frac{1}{2^{\Delta(\mathcal{C},\mathcal{D})}}, \tag{2}$$

где $\mathtt{DD}(\mathcal{C})$ *– это множество всех прямых наследников понятия* $\mathcal{C}$ *(наибольших понятий меньших* $\mathcal{C}$*), а* $\Delta(\mathcal{C},\mathcal{D}) := |\mathtt{Ext}(\mathcal{C}) \setminus \mathtt{Ext}(\mathcal{D})|$ *– разница в количестве объектов между объёмами понятий* $\mathcal{C}$ *и* $\mathcal{D}$*.*

Теоретическая временная сложность данного подхода совпадает со сложностью нахождения всех непосредственных наследников для данного понятия, то есть $O(n \cdot m^2)$. Данная оценка может быть применена для одного понятия и не требует нахождения всего множества понятий, что особенно важно для больших решёток, в которых нахождение всех понятий не представляется возможным. Назовём этот подход *методом оценивания*.

Эксперименты показывают, что для больших понятий существует много понятий с устойчивостью близкой к 1. Чтобы упростить анализ таких решёток, можно перевести устойчивость в логарифмическую шкалу

$$\mathtt{LStab}(\mathcal{C}) = -\log_2(1 - \mathtt{Stab}(\mathcal{C})) \tag{3}$$

Принимая во внимание (2), получаем следующее:

$$-\log_2(\sum_{\mathcal{D}\in\mathtt{DD}(\mathcal{C})} 2^{-\Delta(\mathcal{C},\mathcal{D})}) \leq \mathtt{LStab}(\mathcal{C}) \leq \Delta_{\min}(\mathcal{C}). \tag{4}$$

Здесь $\Delta_{\min}(\mathcal{C}) = \min\limits_{\mathcal{D}\in\mathtt{DD}(\mathcal{C})} \Delta(\mathcal{C},\mathcal{D})$.

По вышеприведённым оценкам устойчивости нельзя гарантировать наперёд заданную точность оценки, но зато она может быть эффективно вычислена. Другая возможная оценка устойчивости приведена в работе М.А. Бабина и С.О. Кузнецова (2012), где авторы оценивают устойчивость методом Монте-Карло. Этот метод позволяет гарантировать наперёд заданную точность, но процедура вычисления требует большого количество попыток для высокой точности. Эти методы могут быть объединены следующим образом, называемым *комбинированным*. Сначала устойчивость рассчитывается с помощью оценочного метода, и если точность оказывается не ниже заданной, то задача является решенной, иначе устойчивость оценивается методом Монте-Карло с требуемой точностью. Как мы увидим, комбинированный метод оценки устойчивости позволяет существенно уменьшить время вычисления оценки

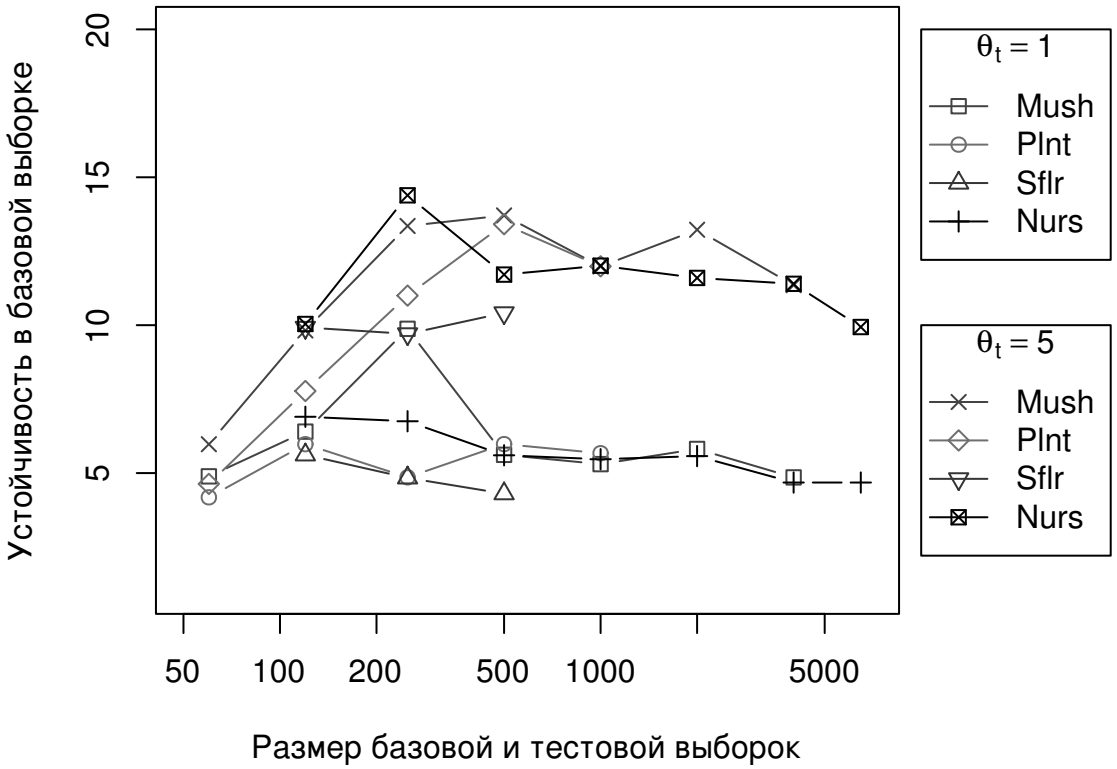

Рисунок 1: Порог устойчивости в базовой выборке, гарантирующий, что 99% понятий в тестовой выборке, похожих на устойчивые понятия базовой выборки, также устойчивы в тестовой для двух порогов $\theta_t = 1$ и $\theta_t = 5$.

по сравнению с методом Монте-Карло, но в отличии от оценочного метода, комбинированный метод гарантирует требуемую точность оценки.

Далее в данной главе исследуется поведение устойчивости. В первом эксперименте доступные выборки данных разделяются случайным образом на две и исследуется, каким образом устойчивость в одной из выборок зависит от устойчивости в другой выборке. В частности, показывается, что устойчивость удобнее использовать в логарифмической шкале. В этом случае устойчивость для понятий с одинаковым содержанием, но порождённых по разным выборкам отличается несущественно. Мы также показываем, что понятия из разных выборок, упорядоченные по устойчивости, имеют схожий порядок, а значит, устойчивость может быть использована для ранжирования понятий.

Ещё один вопрос, который решается этим экспериментом, – *каким должен быть порог устойчивости?* Рисунок 1 показывает, каким должен быть порог в первой выборке, чтобы 99% устойчивых понятий в этой выборке были также устойчивы (по меньшему наперёд заданному порогу) во второй выборке.

Здесь стоит отметить несколько моментов. Во-первых, для выборок небольшого размера разброс данной зависимости от данных к данным может сильно отличаться. Но при этом, начиная примерно с 500-1000 объектов в выборке, пороги устойчивости начинают вести себя похожим образом вне зависимости от данных. Так для того, чтобы содержание устойчивого понятия присутствовало в тестовой выборке, необходимо установить порог устойчивости θ_b в 5-6. Если же мы хотим, чтобы понятие с этим содержанием было устойчивым по порогу $\theta_t = 5$, тогда порог устойчивости в базовой выборке θ_t должен быть равен 11. Данные результаты предполагают, что устойчивость имеет асимптотическое поведение, которое, возможно, может быть доказано формально. На данный момент, насколько это известно, такого формального исследования не проводилось, и, таким образом, такое исследование может стать интересным направлением будущих работ.

Ещё одно направление исследования устойчивости – это сравнение полезности упорядочивания закономерностей с помощью устойчивости и с помощью других индексов полезности. Один из способов такого сравнения – это исследование качества классификации в задаче обучения с учителем с помощью первых закономерностей, выделенных тем или иным методом. Такое сравнение может быть выполнено в рамках следующей методологии:

1. Выбирается выборка данных $\mathcal{D}$.

2. Стократно выборка разделяется на обучающую и тестовую подвыборки случайным образом. При это обучающая выборка содержит 90% $\mathcal{D}$, но не больше чем 1000 объектов, что является ограничением демо-версии программного обеспечения `Magnum Opus` (G. I. Webb & S. Zhang 2005). Это программное обеспечение используется для расчёта меры рычага, участвующей в сравнении.

3. Выбирается целевой класс $\mathcal{C}$.

4. Выбирается целевая мера качества $\mathcal{M}$.

5. В каждой обучающей выборке, полученной на шаге 2, ищутся замкнутые закономерности и они упорядочиваются мерой $\mathcal{M}$. При этом метки классов объектов не учитываются.

6. Среди всего множества закономерностей выделяются гипотезы класса $\mathcal{C}$, известные также как контрастные закономерности (emerging patterns). Гипотезы – это такие закономерности, которые являются характеристическими для одного класса, то есть в поддержке такой закономерности присутствуют в основном (согласно порогу θ) объекты класса $\mathcal{C}$. Такие закономерности предполагаются хорошими для задач классификации. Пусть найдено N гипотез.

7. На этих N гипотезах строятся N классификаторов, основанных на первых $k < N$ гипотезах согласно мере $\mathcal{M}$. Каждый такой классификатор работает следующим образом: для множества гипотез $\{p_1, \cdots, p_k\}$ классификатор относит любой объект к классу $\mathcal{C}$, чьё описание содержит хотя бы одну p_i из множества гипотез.

8. Вычисляются точность и полнота для каждого такого классификатора в тестовом множестве. Эти результаты интерполируются в 21 точке следующего вида: (p, r), где p – это точность, а r – полнота, при этом $r \in \{0, 0.05, \cdots, 0.95, 1\}$. Эти точки задают некоторую кривую.

9. Шаги 6–8 повторяются для каждой пары обучающей и тестовой подвыборки выборки $\mathcal{D}$. Вычисляется усреднённая кривая точности-полноты.

10. Площадь под усреднённой кривой даёт численное значение качества меры $\mathcal{M}$ на выборке $\mathcal{D}$ по отношению к классу $\mathcal{C}$.

11. Шаги 3–10 повторяются для всех классов в $\mathcal{D}$ и для всех тестируемых мер качества.

12. Шаги 1-11 повторяются для все имеющихся выборок данных.

Проведено сравнение следующих мер качества: поддержка, устойчивость, разница, получаемая из верхней оценки устойчивости, и рычаг. Здесь стоит отметить, что устойчивость и разница ведут себя одинаково, подтверждая эффективность введённой оценки. Между устойчивостью и рычагом нет очевидных различий, но устойчивость может быть применена к закономерностям любого типа в том числе как мера качества элементарных моделей процессов, тогда как рычаг используется только для множеств признаков.

В заключении главы приводятся эксперименты эффективности предложенных подходов к оценке устойчивости. С точки зрения временной эффективности предложенные методы оценки оказались существенно более эффективными, чем метод Монте-Карло, предложенный ранее. При этом комбинированный метод позволяет гарантировать точность оценки и существенно повышает точность по сравнению с оценочным методом на неустойчивых понятиях.

С точки зрения точности исследуется, как много ошибочно-устойчивых и ошибочно-неустойчивых понятий может быть найдено, если использовать верхнюю и нижнюю границу оценки в качестве значения устойчивости. Показывается, что нижняя оценка подходит очень близко к истинному значению устойчивости, так как ошибочно-неустойчивых понятий практически нет, в то время как количество ошибочно-устойчивых может доходить до 20%.

Проведённые выше эксперименты показывают, что мера качества элементарных моделей по устойчивости может быть успешно применена для выделения важных элементарных моделей, что позволяет существенно уменьшить размер иерархической модели и делает её пригодной для экспертного анализа.

Четвертая глава посвящена программному комплексу для моделирования объектов сложной структуры на основе узорных структур. Данный комплекс позволяет создавать модели на основе любых узорных структур и комбинировать доступные алгоритмы получения решёток формальных понятий.

Для построения узорной решётки понятий достаточно немного изменить существующие алгоритмы построения решёток формальных понятий. Теоретико-множественная операция пересечения должна быть заменена на полурешёточную операцию сходства, а операция проверки того, что одно множество является подмножеством другого, должна быть заменена на проверку поглощения одного элемента полурешётки другим. Соответственно, операции $(\cdot)'$ должна быть заменена на $(\cdot)^\diamond$.

Математический формализм узорных структур позволяет работать с любыми типами полурёшеток описаний. Полурешётка описаний задаётся неявно через задание операции сходства. Другими словами, пользователь алгоритмически определяет операцию сходства между любыми двумя узорами интересующего его вида. Этот подход с одной стороны позволяет задавать любую

полурешётку описаний, с другой стороны он позволяет сохранять однажды реализованные операции сходства для повторного использования.

Архитектурное решение подбиралось из следующих требований:

Моделирование процессов. Программный комплекс должен позволять моделировать процессы с состояниями сложной структуры.

Гибкость. Программный комплекс должен позволять с минимальными усилиями создавать любые модели на основе узорных структур. В частности необходимо чтобы:

- Любая узорная структура могла быть представленной в проектируемом программном обеспечении.
- Любой алгоритм по построению решётки формальных понятий, который можно адаптировать для построения узорных структур, мог быть добавлен в систему.

Эффективность. Вычислительные затраты для расчёта узорной структуры должны быть минимизированы.

Кроссплатформенность. Необходимо избежать зависимости от окружения программного комплекса, что позволит использовать данный комплекс под такими операционными системами, как Windows, Linux и MacOS.

Общая архитектура программного комплекса показана на рисунке 2. Для поддержки любой узорной структуры в архитектуре выделяется две подсистемы: менеджер узоров – подсистема определения полурешётки описаний, и построитель решётки. На их основании можно создать модель процессов с состояниями сложной структуры.

Предложенная архитектура была реализована в качестве отдельного программного комплекса, а также на её основе узорные структуры могут быть встроены в программный комплекс `FCART`, который разрабатывается на кафедре анализа данных и искусственного интеллекта в высшей школе экономики для моделирования и анализа данных при помощи методов АФП.

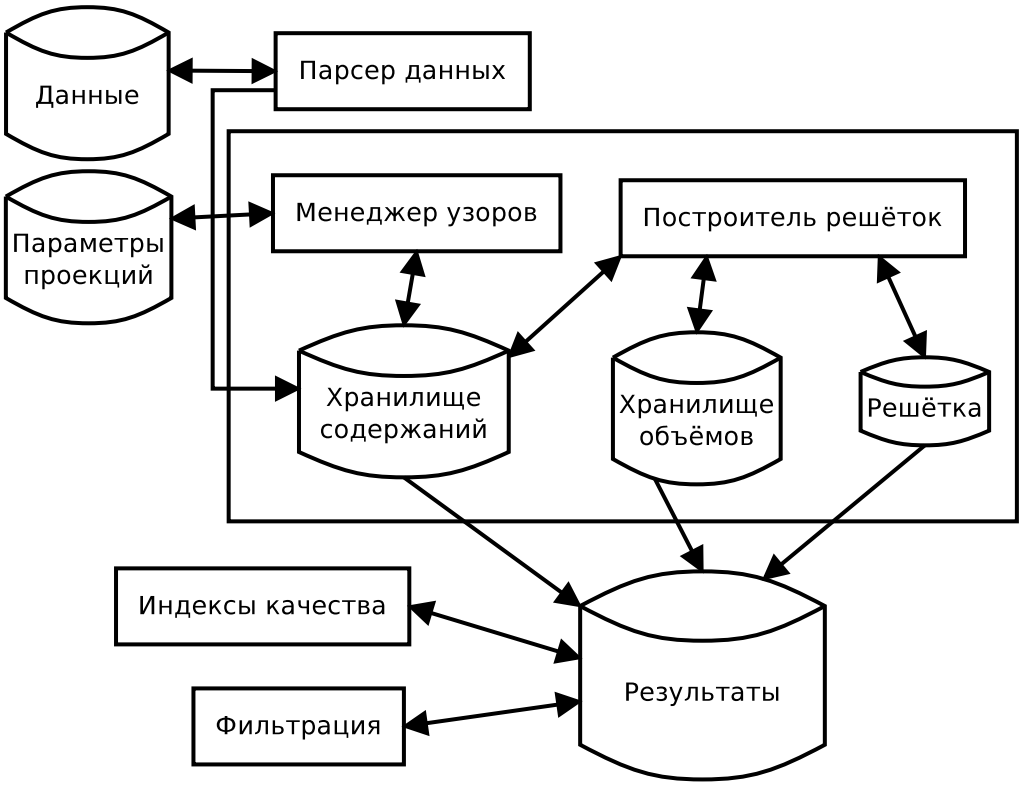

Рисунок 2: Архитектура предлагаемого программного обеспечения для работы с узоными структурами.

Данный программный комплекс реализован на C++ без использования системных библиотек эффективности и возможности кроссплатформенного использования. Программный код этого комплекса занимает около 22 тысяч строк, что соответствует 750 Кб кода, 83 функции являются внешними и могут быть использованы при моделировании на основе узорных структур.

В **пятой главе** приведены результаты экспериментальной апробации модели процессов с состояниями сложной структуры, методов её построения и разработанного комплекса программ, предложенных в предыдущих главах. Эта модель применяется к исследованию посещений веб-страниц пользователями и к процессу госпитализации пациентов. Первый эксперимент является тестированием модели на открытой выборке данных, в то время как второе исследование проводилось во взаимодействии с экспертами предметной области, которые положительно оценили возможности модели.

Для первого эксперимента модель была построена для всех 10^6 записей о посещениях веб-страниц пользователями. Эксперименты показали эффективность введённой ПМД-проекции и её возможности при построении значимых

Тип проекция	Ц!П2	Ц!П3	Ц!ПИ2	Ц!ПИ3	Р!Ц!2	Р!Ц!3
Время работы (s)	18	8	417	15	10	6
Число понятий (в тысячах)	34.7	8.67	1856	93.2	4.2	2.2
Число устойчивых (Stab $\geq$ 0.97) понятий	615	192	1117	311	131	45

Таблица 1: Результаты экспериментов для разных типов проекций.

моделей. Также показано, что описанный в этой работе подход для построения модели по простым данным большого объёма имеет приемлемое время вычислений. Далее разработанный в этой работе подход применялся для моделирования процесса госпитализации пациентов. В этой задаче каждому пациенту соответствует последовательность госпитализаций, задающая реализацию процесса лечения для этого пациента. Каждая госпитализация описывается местоположением больницы, причиной госпитализации и множеством медицинских процедур, которые были применены при лечении больного. Госпитализации типичного больного может выглядеть следующим образом:

$$\langle [\text{CH}_1, \text{Рак}, \{P_1, P_2\}]; [\text{CH}_2, \text{Хим.Тер.}, \{\}] * [10] \rangle .$$

Это означает, что пациент, сначала проходил лечение в больнице $\texttt{CH}_1$, в которой у него детектировали Рак, после чего он прошёл 10 курсов химиотерапии в $\texttt{CH}_2$. Данная выборка содержит 2400 пациентов.

Таблица 1 показывает времена построения решёток узорных понятий, количество понятий в этих решётках, а также количество устойчивых понятий для различных типов проекций. Так первый столбец соответствует проекции Ц!П2, для которой решётка была построена за 18 секунд, в этой решётке около 34700 понятий, среди которых устойчивых только 615. В имени проекции Р соответствует расположению больницы, Ц – цели госпитализации, П – множеству медицинских процедур, а И – интервалам изменений количества госпитализаций на курс химиотерапии. Число в самом конце задаёт минимально допустимую длину подпоследовательности, т. е. параметр ПМД-проекции. Данные проекции задают постановку тех экспериментов, которые отвечают на важные вопросы экспертов предметной области. Также эксперименты показывают, что для каждой из проекции количество ложно-устойчивых понятий, полученных по верхней оценки устойчивости, не превышает 5% от действи-

тельно устойчивых понятий, и, таким образом, для данной выборки можно получать устойчивые понятия, основываюсь только на оценке сверху. Мы видим, что данная модель может быть построена за приемлемое время и её размер не превышает 1000 узорных понятий или элементарных моделей, что является приемлемым для экспертного анализа.

#	Проекция	Содержание	Уст.	Объём
2	Ц!П2	⟨[Рак, {Апп.}]; [Хим.Подготовка, {}]; [Хим.Тер., {}]⟩	4	293
4	Ц!ПИЗ	⟨[Рак, {}]; [Хим.Подготовка, {}]; [Хим.Тер., {}] $*$ [8, 24]⟩	4	193
5	Р!Ц!З	⟨[Регион А,Рак]; [Регион А,Хим.Подготовка]; $\cdots$ $\cdots$ [Конкретная больница в А,Хим.Тер.]⟩	5	29

Таблица 2: Содержания интересных понятий из решёток для разных проекций. `Уст.` – сокращения для порядкового номера по индексу устойчивости. `Хим.Тер.` означает химиотеропию, `Хим.Подготовка` – подготовку к химиотерапии, `Апп.` – оперативное лечение аппендицита.

Таблица 2 показывает содержания некоторых важных понятий, полученных для определённых проекций с соответствующими поддержкой и рангом устойчивости (порядковым номером по индексу устойчивости). Такие важные понятия организованы в иерархическую структуру и при экспертном анализе исследуются от более общих к менее общим, пока не будут найдены элементарные модели важные для решения экспертом конкретной задачи. Например, узор #2 соответствует известной в медицине практике, согласно которой при обнаружении у пациента аппендицита, необходимо также проверить его на раковые заболевания органов пищеварения. В узоре #4 мы можем увидеть количество необходимых сеансов химиотерапии – от 8 до 24. Узор #5 позволяет найти тот факт, что рак у некоторых пациентов был обнаружен в различных больницах региона A, в то время как химиотерапию они все проходили в конкретной клинике. Такого рода узоры помогают как найти скрытые знания о природе заболеваний, как в случаях #2 и #4, так и выявить ошибки или особенности, как в случае #5.

В конце данной главы исследуется возможность сходных работ к исследованию таких процессов. Для этого сходные подходы должны быть модифицированы, чтобы иметь возможность обрабатывать такие процессы. В частности описывается, как сложная структура должна быть переведена в представле-

ние, в котором каждое состояние описывается множеством двоичных признаков. При таком представлении алгоритм `CloSpan` (X. Yan & J. Han, 2003) не может построить модель за приемлемое время с нужной точностью, в то время как алгоритм `M`^{3}`SP` (Plantevit et al., 2010) может построить модель лишь для некоторых проекций с порогом по частоте в 5% за схожее время что и наш подход (проекция `Ц!П2`). Например, для проекции `Ц!ПИ2` подход `M`^{3}`SP` не может построить модель для порога по частоте в 50%.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

- Разработан класс иерархических моделей процессов с состояниями сложной структуры, позволяющих исследовать такие процессы. Метод построения моделей предложенного класса основан на математическом аппарате решеток замкнутых описаний (узорных структур). Одной из ключевых составляющих предлагаемого подхода являются проекции, предоставляющие средства приближения описаний. Они позволяют сохранять важные особенности модели и существенно ускоряют вычисления. Проекции были введены в работе B. Ganter & S.O. Kuznetsov (2001), но в диссертации класс возможных проекций был существенно расширен. Это позволило вводить проекции, которые имеют высокую практическую значимость при моделировании процессов с состояниями сложной структуры. В диссертации были введены и исследованы два важных класса проекций, используемые при построении моделей процессов с состояниями сложной структуры. Первый из них – это проекции минимальной длины, которые позволяют исключать короткие закономерности из иерархической модели, что позволяет существенно сократить время расчёта модели без потери её качества. Второй введенный вид проекции – это проекция на алфавите состояний процессов, позволяющая гибко управлять включаемой в модель информацией, что позволяет быстрее строить иерархическую модель, в которой остаются только важные для эксперта элементарные модели – элементы иерархии.

- Иерархическая модель процессов состоит из элементарных моделей, не все из которых важны. В данной работе экспериментально показывается

возможность использование меры качества по устойчивости для выделения таких значимых элементарных моделей. Во-первых, на широкой тестовой базе было проверено, что устойчивость выделяет схожие элементарные модели при построении иерархических моделей по различным данным, порождённых по одной генеральной совокупности. Во-вторых, устойчивость сравнивается с некоторыми другими мерами качества на задаче задаче классификации. В этом сравнении устойчивость является одним из лидеров, однако расчёт устойчивости при построении модели может занимать существенное время, и поэтому для практического использования устойчивости как меры качества при моделировании реальных процессов, были предложены методы приближённого вычисления. Введены две оценки меры устойчивости, а эффективность этих оценок для отбора моделей была подтверждена при построении моделей реальных процессов с состояниями сложной структуры. Обе введённые оценки имеют важное практическое значение и являются вычислительно более эффективными, чем ранее известные.

- Предложенная математическая модель процессов с состояниями сложной структуры, а также методы её построения реализованы в рамках программного комплекса. Отличительной особенностью предложенной архитектуры этого комплекса является её модульность, которая позволяет исследовать различные типы процессов с состояниями сложной структуры. Более того в рамках разработанного комплекса реализован общий подход к моделированию на основании узорных структур и их проекций, что позволяет с небольшими усилиями создавать и исследовать новые модели, основанные на аппарате узорных структур.

- Разработанные модель, вычислительные методы и комплекс программ были применены для исследования процесса госпитализации пациентов. Результаты исследований были признаны значимыми экспертами в предметной области как позволяющие повысить качество лечения больных. Были рассмотрены различные параметры проекций модели (проекция минимальной длины и проекция алфавита), которые позволяют построить упрощённую, но адекватную модель процессов госпитализации.

Публикации автора по теме диссертации

Публикации в изданиях, входящих в **перечень ВАК**:

1. *Buzmakov A.*, *Kuznetsov S. O.*, *Napoli A.* Scalable Estimates of Concept Stability // Form. Concept Anal. Vol. 8478 / ed. by C. V. Glodeanu, M. Kaytoue, C. Sacarea. — Springer Berlin Heidelberg, 2014. — Pp. 157–172. — (Lecture Notes in Computer Science). — ISBN 978-3-319-07247-0.
2. *Бузмаков А. В.* Узорные структуры для анализа сложных последовательностей // Научно-техническая информация. серия 2 информационные процессы и системы. — 2013. — Т. 10. — С. 27—39.

Прочие публикации:

3. *Buzmakov A.*, *Egho E.*, *Jay N.*, *Kuznetsov S. O.*, *Napoli A.*, *Raïssi C.* FCA and pattern structures for mining care trajectories // Work. Notes FCA4AI. — 2013. — Pp. 7–14.
4. *Buzmakov A.*, *Egho E.*, *Jay N.*, *Kuznetsov S. O.*, *Napoli A.*, *Raïssi C.* On Projections of Sequential Pattern Structures (with an application on care trajectories) // Proc. 10th Int. Conf. Concept Lattices Their Appl. — 2013. — Pp. 199–208.
5. *Buzmakov A.*, *Egho E.*, *Jay N.*, *Kuznetsov S. O.*, *Napoli A.*, *Raïssi C.* The representation of sequential patterns and their projections within Formal Concept Analysis // Work. Notes LML. — 2013. — Pp. 65–79.
6. *Buzmakov A.*, *Kuznetsov S. O.*, *Napoli A.* A New Approach to Classification by Means of Jumping Emerging Patterns // Work. Notes FCA4AI. — 2012. — Pp. 15–22.
7. *Buzmakov A.*, *Kuznetsov S. O.*, *Napoli A.* Is Concept Stability a Measure for Pattern Selection? // Procedia Comput. Sci. — 2014. — Vol. 31. — Pp. 918–927. — ISSN 1877-0509.
8. *Buzmakov A.*, *Neznanov A. A.* Practical Computing with Pattern Structures in FCART Environment // Work. Notes FCA4AI. — 2013. — Pp. 49–52.

Лицензия ЛР № 020832 от «15» октября 1993 г.
Подписано в печать «_» _____ ___ г. Формат 60x84/16
Бумага офсетная. Печать офсетная.
Усл. печ. л. 1.
Тираж 100 экз. Заказ №__ Типография издательства НИУ ВШЭ,
125319, г. Москва, Кочновский пр-д., д. 3.