На правах рукописи

КАШНИЦКИЙ
Юрий Савельевич

МЕТОДЫ ЗАМКНУТЫХ ОПИСАНИЙ В ЗАДАЧЕ КЛАССИФИКАЦИИ ДАННЫХ СО СЛОЖНОЙ СТРУКТУРОЙ

Специальность 05.13.17 —
Теоретические основы информатики
(технические науки)

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата технических наук

Москва – 2018

Работа выполнена в Департаменте анализа данных и искусственного интеллекта федерального государственного автономного образовательного учреждения высшего образования "Национальный исследовательский университет "Высшая школа экономики"

Научный руководитель:	**Кузнецов Сергей Олегович**, доктор физико-математических наук, заведующий департамента анализа данных и искусственного интеллекта Национального исследовательского университета "Высшая школа экономики"
Официальные оппоненты:	**Богатырев Михаил Юрьевич**, доктор технических наук, профессор кафедры информационной безопасности Тульского государственного университета
	Чувилин Кирилл Владимирович, кандидат технических наук, доцент кафедры математических основ управления факультета управления и прикладной математики Московского физико-технического института
Ведущая организация:	Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук (ИПУ РАН)

Защита состоится "____" _____ 201__ г. в ____ часов на заседании диссертационного совета Д 002.073.05 при Федеральном государственном учреждении "Федеральный исследовательский центр "Информатика и управление" Российской академии наук" (ФИЦ ИУ РАН) по адресу: 119333, г. Москва, ул. Вавилова, д.44, кор. 2.

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН, Москва, ул. Вавилова, д. 40, и на официальном сайте ФИЦ ИУ РАН `http://www.frccsc.ru`.

Ученый секретарь диссертационного совета Д 002.073.05
Д.ф.-м.н. Рязанов В.В.

Общая характеристика работы

Актуальность темы. Чаще всего задачи анализа данных формулируются для данных, которые можно представить объектно-признаковыми таблицами. Если посмотреть на задачи машинного обучения в корпоративной среде или соревнования по анализу данных[1], то за редким исключением они сводятся к анализу объектно-признаковых таблиц. При этом данные со сложной структурой (тексты, изображения) тоже представляются в некотором признаковом пространстве (TF-IDF, word2vec, нейросетевые признаки изображений и т.д.). Однако в последнее время активно развиваются методы анализа сложно структурированных данных, для которых теоретически сложно либо практически неэффективно составлять признаковые описания, зато можно судить о свойствах объектов на основе сходства их описаний. Такие задачи встречаются в химической информатике (Misra et al., 2011), анализе текстов (Jurafsky et al., 2000) и изображений (Navarin, 2014) . Далее в этой работе под сложно структурированными данными мы будем понимать данные, для которых можно определить узорную структуру.

Важным аспектом в решении задач классификации является интерпретируемость полученных результатов. Во многих приложениях, особенно в медицине, необходима интерпретация результатов классификации в виде понятных человеку правил, к которым можно применить экспертный анализ и на его основе судить о релевантности используемых моделей, алгоритмов и мер сходства объектов в конкретной задаче. В разных задачах интерпретируемость определяется по-разному, но в данной работе под интерпретируемостью алгоритмов мы будем понимать их возможность объяснить классификацию тестовых примеров. Конкретней, под локальной интерпретируемостью классификации мы пониманием среднюю длину посылок правил, с помощью которых делается прогноз для тестового примера. В (Holte, 1993) показано, что методы классификации на основе коротких правил хорошо работают на большинстве наборах данных популярного репозитория UCI, при этом методы хорошо интерпретируемы, то есть полученные правила могут анализироваться экспертами.

Одним из успешных инструментов для анализа сложно структурированных данных является ДСМ-метод автоматического восстановления зависимостей из эмпирических данных (Финн В.К., 1983), (Финн В.К, 2010), (Кузнецов С.О., 1991), (Дюкова Е.В., 2002). Классификация на основе ДСМ-метода относится к интерпретируемым подходам, поскольку позволяет анализировать структурное сходство тестового примера и обучающих. Однако по качеству классификации, определяемому по метрике типа доли верных ответов на кросс-валидации или отложенной выборке, такой подход уступает ядерным методам (kernel methods) (T. Hofmann et al., 2008), в особенности, методу опорных векторов (C. Cortes, V. Vapnik, 1995). Было предложено множество ядерных функций для оценки сходства объектов со сложной структурой – строковые ядра (H. Lodhi et al., 2002), ядра для последовательностей (C. Cortes, 2008), и графовые ядра (S. Vishwanathan et al., 2010). Недостатком метода опорных векторов является плохая интерпретируемость полученных результатов.

Необходимость анализа данных со сложными структурными описаниями и решения связанных с ними задач классификации делает актуальным применение методов, позволяющих работать со структурным сходством и использовать эффективные приближения описаний. Методы анализа формальных понятий и решеток замкнутых описаний (узорных структур) предоставляют удобный и эффективный математический аппарат для построения моделей в решении целого ряда важных

[1] www.kaggle.com/competitions

научных и прикладных задач. В задачах обучения без учителя эти методы актуальны, поскольку позволяют находить и интерпретировать сходство произвольного множества объектов, а в задачах обучения с учителем – потому что с их помощью можно получить наборы классифицирующих правил, понятных человеку (интерпретируемых) и позволяющих далее применять к ним экспертный анализ. Аппарат проекций узорных структур позволяет эффективно работать с приближенными описаниями сложно структурированных объектов, учитывая основные свойства структуры и понижая вычислительную и временную сложность обработки таких описаний.

Таким образом, **объектом исследования** являются данные со сложной структурой. **Предметом исследования** являются методы, алгоритмы и программы для классификации данных со сложной структурой с помощью классифицирующих правил, а также их экспертного анализа.

Целью диссертационного исследования является разработка единого подхода к классификации данных со сложной структурой. Результатами работы алгоритма должны быть как приемлемое для конкретной задачи качество классификации, так и интерпретируемый вывод алгоритма в виде коротких классифицирующих правил, подходящий для дальнейшего экспертного анализа.

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

1. Предложить универсальный подход к классификации данных со сложной структурой на основе решеток замкнутых описаний;

2. В частном случае описаний в виде бинарных, категориальных и количественных признаков предложить подход к классификации на основе правил, решающий задачу классификации лучше (по точности), чем деревья решений, и порождающий более короткие правила, чем алгоритм случайного леса;

3. Разработать комплекс программ для классификации данных со сложной структурой и апробировать его в задачах классификации как с бинарными, категориальными и количественными признаками, так и с описаниями со сложной структурой в виде последовательностей и графов.

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

1. Предложен новый подход к классификации данных со сложной структурой на основе узорных структур;

2. Предложен специальный вид проекций узорных структур для данных с количественными признаками, обобщающий подход к обучению на основе деревьев решений;

3. Создан комплекс программ для классификации данных со сложной структурой на основе решеток замкнутых описаний. Соответствующие алгоритмы были апробированы на многих наборах данных с категориальными и количественными признаками, а также на данных по токсичности химических веществ со сложной структурой в виде молекулярных графов.

Теоретическая ценность данной работы состоит

1. в представлении методов классификации числовых данных, в том числе деревьев решений, с помощью проекций интервальных узорных структур;

2. в представлении подхода к классификации на основе правил, гарантирующего нахождение правил с лучшим значением выбранного критерия информативности, чем правила, полученные с помощью деревьев решений;

3. во введении и исследовании дискретизирующей проекции для интервальных узорных структур.

Практическая ценность работы состоит

1. в получении качественных (по доле правильных ответов) и интерпретируемых решений задач классификации данных в виде последовательностей и графов;

2. в получении качества классификации в экспериментах с реальными данными, статистически значимо лучшего, чем у алгоритмов построения деревьев решений;

3. в представлении алгоритма классификации на основе правил, более коротких по длине (числу признаков в посылке), а потому легче интерпретируемых, чем правила, построенные алгоритмом случайного леса;

4. в разработке программного комплекса, позволяющего анализировать сложно структурированные данные и решать для них задачи классификации с помощью интерпретируемых наборов правил, подходящих для дальнейшего экспертного анализа.

Положения, выносимые на защиту:

1. Предложен универсальный подход к классификации данных со сложной структурой на основе решеток замкнутых описаний. При этом для каждого объекта порождаются наборы коротких и интерпретируемых классифицирующих правил;

2. Показано, что предложенный алгоритм классификации на основе правил демонстрирует более высокое качество классификации (в терминах средней доли правильных ответов и F1-метрики на кросс-валидации), чем деревья решений. Также он порождает в среднем более короткие и интерпретируемые правила, чем алгоритм случайного леса;

3. Показано, что для любого объекта можно найти подходящее классифицирующее правило, такое что его посылка будет замкнутым множеством признаков, а качество правила (измеряемое с помощью критерия типа прироста информации) – выше, чем у любого подходящего правила, построенного деревом решений.

4. Предложен вид приближений числовых описаний (в терминах проекций интервальных узорных структур), на основе которых представлены посылки правил, полученных с помощью деревьев решений. Эффективность использования таких проекций экспериментально подтверждена в задаче классификации для нескольких наборов данных с количественными признаками;

5. Разработан комплекс программ для анализа данных со сложной структурой на основе решеток замкнутых описаний. Поддерживаются 4 типа данных: числовые (бинарные, категориальные и количественные признаки), интервальные, последовательности и помеченные графы.

Достоверность полученных результатов опирается на строгость использованных математических моделей, их экспериментальное подтверждение и практическую эффективность программных реализаций.

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

1. Семинар Межфакультетской кафедры математического моделирования и компьютерных исследований МГУ имени М.В. Ломоносова 31 октября 2017 года, г. Москва;
2. Семинары отдела Интеллектуальных систем ВЦ РАН им. А.А. Дородницына 20 октября 2016 года, 7 июля 2017 года и 14 сентября 2017 года, г. Москва;
3. 23-ий Международный симпозиум по методологиям интеллектуальных систем (ISMIS 2017), июнь 2017 г., г. Варшава, Польша.
4. Семинары Департамента Анализа Данных и Искусственного Интеллекта НИУ ВШЭ (6 выступлений в мае и октябре 2015-2016 гг., а также в декабре 2016 г. и марте 2017 г.), г. Москва;
5. Пятнадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2016), сентябрь 2016 г., г. Смоленск;
6. Семинар "What can FCA do for Artificial Intelligence?" при Европейской конференции по искусственному интеллекту ECAI, август 2016 г., г. Гаага, Нидерланды;
7. 13-ая международная конференция по решеткам понятий и их приложениям (The 13th International Conference on Concept Lattices and Their Applications), июль 2016 г., г. Москва;
8. Конференция "Технологии Больших Данных" (ТБД-2016), июнь 2016 г., г. Москва;
9. Пятая международная конференция по Анализу Изображений, Сетей и Текстов АИСТ 2016, г. Екатеринбург (награда за лучший доклад в секции "Data Analysis, Graphs & Complex Data");
10. Семинар "What can FCA do for Artificial Intelligence?" при международной объединенной конференции по искусственному интеллекту IJCAI, июль 2015 г., г. Буэнос-Айрес, Аргентина;
11. Ph.D.-семинар при Европейской конференции по машинному обучению и теоретическим основам и практике обнаружения знаний в базах данных ECML/PKDD, 2014 г., г. Нанси, Франция;
12. Семинар "What can FCA do for Artificial Intelligence?" при Европейской конференции по искусственному интеллекту ECAI, июль 2014 г., г. Прага, Чехия;
13. Третья международная конференция по Анализу Изображений, Сетей и Текстов АИСТ, апрель 2014 г., г. Екатеринбург;

Публикации. Основные результаты по теме диссертации изложены в 8 научных работах, 1 из которых издана в издании, рекомендованном ВАК, 7 — в рецензируемых трудах международных конференций, индексируемых в базе данных научного цитирования Scopus.

Диссертация состоит из введения, 4 глав, заключения, списка литературы, а также списков рисунков, таблиц и приложений. Общий объем работы — 111 страниц. Список литературы включает 107 наименований.

Содержание работы

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

Первая глава посвящена обзору базовых понятий теории решеток и Анализа Формальных Понятий, приводится обзор методов классификации в машинном обучении, основанных на ассоциативных правилах, а также рассматриваются критерии отбора классифицирующих правил. Кроме того, обсуждаются подходы к решению задачи классификации, основанные на Анализе Формальных Понятий. Деревья решений интерпретируются в терминах АФП и показывается, как с помощью алгоритма построения решетки формальных понятий предложить отбор правил в задаче классификации, при котором гарантируется, что каждый объект тестовой выборки классифицируется правилом не хуже (в терминах выбранного критерия информативности, такого как неопределенность Джини или прирост информации), чем при классификации на основе дерева решений (Теорема 1).

Во **второй главе** рассматривается аппарат узорных структур (Pattern Structures) (B. Ganter, S.O. Kuznetsov, 2001), который позволяет расширить методы Анализа Формальных Понятий на случай, когда объекты задаются не бинарными признаками, а сложными описаниями. Такими описаниями могут быть интервалы числовых значений, множества последовательностей, строк или графов. Формулируется теорема, аналогичная Теореме 1), но для случая количественных признаков и интервальных узорных структур. Для этого предложена дискретизирующая проекция интервальной узорной структуры (Определение 7). Рассматриваются подходы к классификации данных со сложной структурой на основе ядерных функций и метода опорных векторов, а также на основе узорных структур и их проекций.

Определение 1. *Узорная структура – это тройка* $(G,(D, \sqcap), \delta)$*, где* G *– множество объектов,* $(D, \sqcap)$ *– полная полурешетка всевозможных описаний, а* $\delta\colon G \to D$ *– функция, которая сопоставляет каждому объекту из множества* G *его описание из* D*.*

Соответствие Галуа между подмножествами множества объектов и множеством описаний для узорной структуры $(G, (D, \sqcap), \delta)$ записывается следующим образом:

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

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

Определение 2. *Узорное понятие узорной структуры* $(G, (D, \sqcap), \delta)$ *– это пара* (A, d)*, в которой* $A \subseteq G$ *– подмножество множества объектов,* $d \in D$ *– одно из описаний из полурешетки* $(D, \sqcap)$*, такие что* $A^\diamond = d$ *и* $d^\diamond = A$*. Множество объектов* A *называется узорным объемом понятия, а* d *– его узорным содержанием.*

Количество формальных понятий в решетке, построенной по формальному контексту, может быть экспоненциальным от количества объектов (R. Wille, B. Ganter, 1997). Формальный контекст

– это частный случай узорных структур, и поэтому количество узорных понятий в решетке, построенной для некоторой узорной структуры, может быть экспоненциальным от количества объектов в множестве G. Значит, построение полной полурешетки узорных понятий может быть очень вычислительно сложным. Более того, большинство найденных узорных понятий не интересны для дальнейшего исследования, хотя занимают существенную часть времени вычислений. В случае, когда сама полурешёточная операция сходства вычислительно сложна, построение решетки узорных понятий может стать невозможным. Например, в качестве полурешеточной операции сходства на узорной структуре на графах нужно определять изоморфизм подграфу (С.О. Кузнецов, 2005), что является NP-полной задачей. Для сокращения времени работы алгоритмов построения узорных решеток были введены проекции узорных структур (B. Ganter, S.O. Kuznetsov, 2001). Проекция может быть рассмотрена как способ фильтрации полурешетки описаний с определенными математическими свойствами. Эти свойства позволяют задать связь между понятиями в спроецированной и начальной узорных структурах. К тому же полурешетка, построенная для спроецированной узорной структуры может оказаться значительно меньше исходной, что упрощает ее построение и исследование.

Определение 3. *Проекция полурешетки* $(D, \sqcap)$ *– это функция* $\psi : D \to D$*, которая является оператором ядра, т.е. для любых двух* $x, y \in D$ *верно:*

- $x \sqsubseteq y \Rightarrow \psi(x) \sqsubseteq \psi(y)$ *(монотонность)*
- $\psi(x) \sqsubseteq x$ *(сжимаемость)*
- $\psi(\psi(x)) = \psi(x)$ *(идемпотентность)*

Определение 4. *Проекция узорной структуры, полученная из узорной структуры* $(G, (D, \sqcap), \delta)$ *с помощью проекции* ψ *– это 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 полурешеточной операцией* $\sqcap_\psi$ *такой, что* $\forall x, y \in D\ x \sqcap_\psi y := \psi(x \sqcap y)$*, а* $\delta_\psi = \psi \circ \delta$*.*

Для анализа данных с вещественными значениями признаков в Анализе Формальных Понятий вводятся интервальные узорные структуры.

Описания D объектов узорной структуры образуют полную полурешетку $(D, \sqcap)$, где $\sqcap$ – коммутативная, ассоциативная и идемпотентная операция, определенная на описаниях объектов. Интуитивный смысл этой операции – "сходство" описаний. Для интервалов операция сходства $\sqcap$ определяется следующим образом (M. Kaytoue et al., 2011):

Определение 5. *Пусть* $[a_1,b_1]$ *и* $[a_2,b_2]$ *– два интервала на множестве действительных чисел, т.е.* $a_1, b_1, a_2, b_2 \in \mathbb{R}, a_1 \leq b_1, a_2 \leq b_2$*. Тогда операция сходства для двух интервалов определяется как* $[a_1, b_1] \sqcap [a_2, b_2] = [min(a_1, a_2), max(b_1, b_2)]$*.*

Определение 6. *Для множества объектов* G*, множества описаний* $D = \langle [a_i, b_i] \rangle_{i \in [1,m]} (a_i, b_i \in \mathbb{R}, a_i \leq b_i)$*, полурешеточной операции* $\sqcap$ *и функции* $\delta(G) := \{\delta(g) \mid g \in G\}$ *соответствующая узорная структура* $(G, (D, \sqcap), \delta)$ *называется* ***интервальной узорной структурой****.*

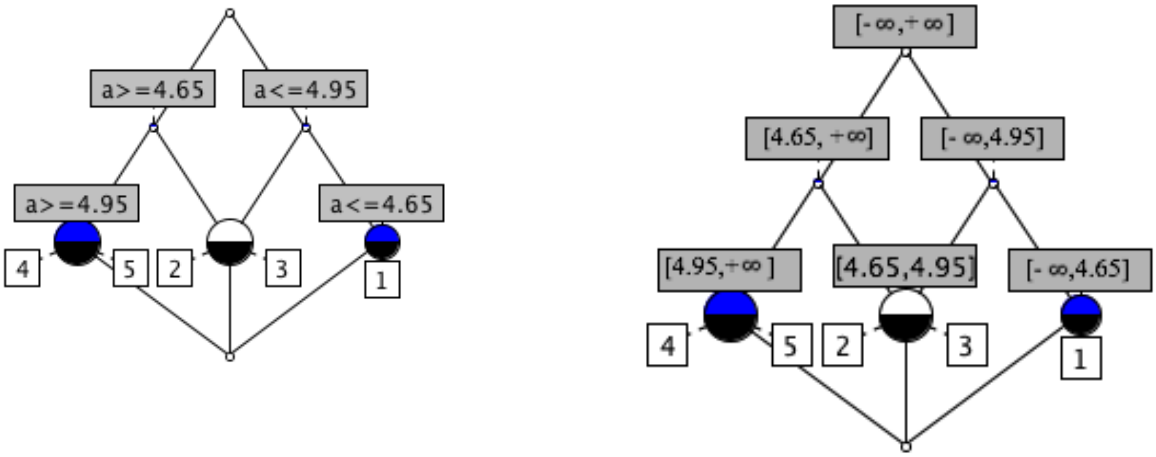


Рисунок 1: Решетка формальных понятий для контекста из Таблицы 1 и изоморфная ей решетка узорных понятий для узорной структуры.

Интервальные узорные структуры были успешно применены для анализа экспрессии генов (M. Kaytoue et al, 2011). В этой задаче каждый ген описывается степенью своей экспрессии в определенных условиях. Таким образом, задано несколько признаков одного гена, соответствующих условиям и имеющие численные значения.

Для последующего сравнения алгоритмов классификации на основе деревьев решений и на основе АФП введем специальный вид проекций для интервальных узорных структур.

Определение 7. *Пусть* $(G, (D, \sqcap), \delta)$ *– интервальная узорная структура и* m *– размерность векторов описаний (см. Определение 6). Пусть* $T_i = \{\tau_{i1}, \ldots, \tau_{it_i}\}$ $(\tau_{ij} \in \mathbb{R}, i \in [1,m], j \in [1,t_i], t_i \in \mathbb{N})$ *– множества вещественных чисел. Тогда,* $\psi(\langle [a_i,b_i] \rangle_{i \in [1,m]}) =$
$\langle [max\{\tau \mid \tau \in T_i \cup \{-\infty, +\infty\}, \tau \leq a_i\}, min\{\tau \mid \tau \in T_i \cup \{-\infty, +\infty\}, \tau \geq b_i\}] \rangle$ *называется* ***дискретизирующей проекцией*** *для интервальной узорной структуры* $(G, (D, \sqcap), \delta)$.

	a
1	4.6
2	4.7
3	4.9
4	5.0
5	5.1

	$a \leq 4.65$	$a \leq 4.95$	$a \geq 4.65$	$a \geq 4.95$
1	×	×		
2		×	×	
3		×	×	
4			×	×
5			×	×

Таблица 1: Простой многозначный контекст и контекст, полученный дискретизированием признака a порогами 4.65 и 4.95.

Пример 1. *Возьмем признак* a *и дискретизируем его порогами* $T = \{4.65, 4.95\}$. *Полученный формальный контекст представлен Таблицей 1, а соответствующая решетка формальных понятий показана на Рисунке 1 (слева).*

$\psi([a,b]) = [max\{\tau \mid \tau \in T^+, \tau \leq a\}, min\{\tau \mid \tau \in T^+, \tau \geq b\}]$ *с* $T^+ = \{-\infty, 4.65, 4.95, +\infty\}$ *– это проекция полурешетки из прошлого примера, а соответствующая решетка узорных понятий изо-*

морфна решетке формальных понятий дискретизированного контекста (Рис. 1 (слева)) и показана на Рис 1 (справа).

Проекция ψ сопоставляет каждому узорному понятию из прошлого примера узорное понятие спроецированной узорной структуры.

Далее покажем, что при помощи CbO-дерева, используемого алгоритмом построения множества формальных понятий "Замыкай по-Одному" (С.О. Кузнецов, 1993) в задаче бинарной классификации можно находить классифицирующие правила с приростом информации, не меньшим, чем в дереве решений соответствующей глубины.

Определение 8. *Пусть дан формальный контекст $\mathbb{K} = (G,M,I)$ и признаки из множества M пронумерованы, т.е. для множества признаков M задан порядок $(\alpha(M), <), \forall m \in M\ \alpha(m) \in [1, |M|]$. Пусть для $B \subseteq M$ $min(B)$ выдает признаки из B с минимальным номером:*
$min(B) = \{m \mid m \in B, \alpha(m) < \alpha(\tilde{m})\ \forall \tilde{m} \in B \backslash \{m\}\}$.

*Обозначим $suc(B)$ – множество всех наследников множества B: понятий с содержанием вида $(B \cup \{i\})''$, таких что $min((B \cup \{i\})'' \setminus B) = \{i\}$. **Признаковым CbO-деревом** для формального контекста $\mathbb{K}$ называется дерево, состоящее из всевозможных множеств $suc(B)$, дуги которого задаются отношением $(B, suc(B))$.*

Теорема 1. *Пусть решается задача бинарной классификации, и обучающая выборка задана формальным контекстом $\mathbb{K}_{+-} = (G_+ \cup G_-,\ M \cup \tau,\ I_+ \cup I_-)$. Пусть также множество признаков дихотомизировано: $M = M_0 \cup \neg M_0$. Пусть для данного формального контекста построено признаковое CbO-дерево T_{CbO}. Для любого пути решения $\langle m_1, \ldots, m_j \rangle$ дерева решений T глубины k $(j \leq k)$ с приростом информации $IG(\langle m_1, \ldots, m_j \rangle)$ найдется замкнутое множество признаков, являющееся вершиной CbO-дерева на глубине не более k, а также посылкой классифицирующего правила с не меньшим приростом информации, чем у $\langle m_1, \ldots, m_j \rangle$.*

Говоря про реализацию алгоритма поиска посылок классифицирующих правил среди формальных понятий, отметим, что доказанное утверждение означает, что для любого правила, построенного деревом решений и имеющего мощность посылки k, можно найти правило с не меньшим приростом информации при построении CbO-дерева с глубиной рекурсии k. Легко показать, что аналогичные утверждения верны и для неопределенности Джини.

В первой и второй главах показано, что Анализ Формальных Понятий предлагает удобный формализм для того чтобы, с одной стороны, выразить на этом языке многие алгоритмы, основанные на классифицирующих ассоциативных правилах, а с другой, чтобы обобщить эти алгоритмы на случай данных со сложной структурой. В **третьей главе** предлагается алгоритм классификации произвольных данных со сложной структурой, для которых можно ввести полурешёточную операцию сходства. Отдельно и с подробными примерами рассматриваются частные случаи, когда данные представлены бинарными, количественными и интервальными признаками, а также помеченными графами.

Предлагаемый подход в случае бинарных признаков в обучающей и тестовой выборке описан в Алгоритме 1 – CoLiBRi (Concept Lattice-Based Rule-learner, классификация на основе правил с помощью решеток формальных понятий). Для категориальных признаков предлагается использовать

One Hot Encoding, то есть для каждого категориального признака порождать бинарные признаки в количестве, равном уникальному числу значений этого категориального признака.

На вход алгоритму подаются обучающий и тестовый формальные контексты $\mathbb{K}_{train} = (G_{train}, M_0 \cup \overline{M}_0 \cup c_{train}, I_{train})$ и $\mathbb{K}_{test} = (G_{test}, M_0 \cup \overline{M}_0, I_{test})$. Множество признаков M дихотомизировано: $M = M_0 \cup \overline{M}_0$, где $\forall g \in G_{train}, m \in M_0 \; \exists\, \overline{m} \in \overline{M}_0 : gI_{train}m \to \neg(gI_{train}\overline{m})$. Также алгоритм использует модификацию программной реализации In-Close 2 (Andrews, 2009) алгоритма "Замыкай по-Одному" ($CbO(K, min\_supp)$) (С.О. Кузнецов, 1993), в которой выдаются все формальные понятия формального контекста K, поддержки которых ограничены снизу значением параметра min_supp. Для выбора классифицирующих правил используется критерий $inf : M \cup c_{train} \to \mathbb{R}$ типа неопределенности Джини или энтропийного прироста информации (в программной реализации по умолчанию – неопределенность Джини). Параметры алгоритма: min_supp и n – минимальная поддержка классифицирующих правил и число правил, используемых для классификации тестового объекта.

Алгоритм состоит из следующих шагов:

1. Инициализировать c_{test} пустым списком, а r_{test} – пустым словарем. В c_{test} будут добавляться предсказанные значения целевого признака для тестовых объектов, а в r_{test} – правила для каждого тестового объекта (ключ в словаре – номер объекта, значение – список правил).

2. Посчитать долю положительных объектов в выборке $c_{pos} = \frac{|c'_{train}|}{|G_{train}|}$.

3. С помощью алгоритма $CbO(K, min\_supp)$ найти все формальные понятия обучающего контекста $\mathbb{K}_{train}$ со значением поддержки не менее min_supp. Параллельно с этим для каждого формального понятия вычислять значение качества соответствующего классифицирующего правила inf. Таким образом, получится словарь $\mathcal{S}$, ключами которого будут содержания формальный понятий, а значениями – соответствующие значения функционала inf.

4. Отсортировать все формальные понятия $\mathcal{S}$ по посчитанным значениям критерия inf в порядке "улучшения", то есть по возрастанию inf, если малые значения критерия говорят о хороших правилах (как в случае неопределенности Джини) или по убыванию, если, наоборот, большие значения критерия свидетельствуют о хороших правилах (прирост информации, среднее уменьшение Джини).

5. Для каждого тестового объекта $g_t \in G_{test}$:

   – Отобрать n_{rules} "подходящих" содержаний формальных понятий, то есть $\{B_i\}_{i \in [1, n_{rules}]} = \{B \mid (A,B) \in \mathcal{S}, g'_t \subseteq B\}$

   – Для каждого из отобранных содержаний формальных понятий $\{B_i\}_{i \in [1, n_{rules}]}$ определить долю положительных объектов $c_i = \frac{|B'_i \cap c'_{train}|}{|B'_i|}$

   – Сформировать таким образом набор правил $\{B_i \to_{c_i} +\}_{i \in [1, n_{rules}]}$ с достоверностями c_i. Записать его в словарь r_{test} для ключа t (номер объекта g_t)

   – Предсказанное значение целевого признака c_{train_t} определить как индикатор того, что средняя арифметическая достоверность найденных правил превышает долю положительных объектов во всей выборке:

$G\backslash M$	**os**	$\neg os$	**oo**	$\neg oo$	**or**	$\neg or$	**th**	$\neg th$	**tm**	$\neg tm$	**tc**	$\neg tc$	$\neg hn$	**hn**	**w**	$\neg w$	**play**
1	×			×		×	×			×		×	×			×	
2	×			×		×	×			×		×	×		×		
3		×	×			×	×			×		×	×			×	×
4		×		×	×			×	×			×	×			×	×
5		×		×	×			×		×	×			×		×	×
6		×		×	×			×		×	×			×	×		
7		×	×			×		×		×	×			×	×		×
8	×			×		×		×	×			×	×			×	
9	×			×		×		×		×	×			×		×	×
10		×		×	×			×	×			×		×		×	×
11	×			×		×		×	×			×		×	×		?
12		×	×			×		×	×			×	×		×		?
13		×	×			×	×			×		×		×		×	?
14		×		×	×			×	×			×	×		×		?

Таблица 2: Пример классификационного контекста.

$$c_{train_t} = [\frac{1}{n\_rules} \sum_{i=1}^{n\_rules} c_i \geq c_{pos}].$$

Добавить это значение в c_{test}.

Algorithm 1 Concept Lattice-Based Rule-learner (CoLiBRi) – случай бинарных признаков.

Input: $\mathbb{K}_{train} = (G_{train}, M_0 \cup \overline{M}_0 \cup c_{train}, I_{train})$
$\mathbb{K}_{test} = (G_{test}, M_0 \cup \overline{M}_0, I_{test})$
$min\_supp \in \mathbb{R}^+$, $n_{rules} \in \mathbb{N}$;
$CbO(K, min\_supp) : K \to \mathcal{S}$;
$inf : M \cup c_{train} \to \mathbb{R}$;
$sort(\mathcal{S}, inf) : \mathcal{S} \to \mathcal{S}$

Output: c_{test}, r_{test}

$c_{test} = \emptyset, r_{test} = \emptyset$
$c_{pos} = \frac{|c'_{train}|}{|G_{train}|}$
$\mathcal{S} = \{(A,B) : inf(B, c_{train}) \mid A \subseteq G_{train}, B \subseteq M, A' = B, B' = A, |A| \geq min\_supp\} = CbO(\mathbb{K}_{train}, min\_supp)$
$\mathcal{S} = sort(\mathcal{S}, inf)$
for $g_t \in G_{test}$ **do**
  $\{B_i\}_{i \in [1, n_{rules}]} = \{B \mid (A,B) \in \mathcal{S}, g'_t \subseteq B\}$
  $c_i = \frac{|B'_i \cap c'_{train}|}{|B'_i|}$
  $r_{test}[i] = \{B_i \to_{c_i} +\}_{i \in [1, n_{rules}]}$
  $c_{test}[i] = [\frac{1}{n\_rules} \sum_{i=1}^{n\_rules} c_i \geq c_{pos}]$
end for

Пример 2. *Продемонстрируем работу алгоритма для набора данных из Таблицы 2. Здесь:*

– $\mathbb{K}_{train} = (G_{train}, M_0 \cup \overline{M}_0 \cup c_{train}, I_{train})$

– $G_{train} = \{1,2,\ldots, 10\}$

	$\{\overline{w}, \overline{tm}\}$	
	Yes	**NO**
$play$	3	3
$\neg play$	1	3

Таблица 3: Таблица сопряженности для $\{\overline{w}, \overline{tm}\}$ и целевого признака $play$.

- M_0 = {or,oo,os,tc,tm,th,hn,w} – *множество признаков Outlook=rainy, Outlook=overcast, Outlook=sunny, Temperature=cool, Temperature=mild, Temperature=hot, Humidity=normal, Windy соответственно.*
- $\overline{M}_0 = \{\overline{or},\overline{oo},\overline{os},\overline{tc},\overline{tm},\overline{th},\overline{hn},\overline{w}\}$ – *множество "отрицаний" признаков из* M_0.
- $I_{train} \subseteq G_{train} \times M_0 \cup \overline{M}_0 \cup c_{train}$ – *бинарное отношение, показанное в Таблице 2 в строках 1–10.*
- $\mathbb{K}_{test} = (G_{test}, M_0 \cup \overline{M}_0, I_{test})$.
- G_{test} = {11,12,13,14}
- $I_{test} \subseteq G_{train} \times M_0 \cup \overline{M}_0$ – *бинарное отношение, показанное в Таблице 2 в строках 11–14.*
- *Зафиксируем среднее значение неопределенности Джини как критерий отбора классифицирующих правил* $inf : M \cup c_{train} \to \mathbb{R}$.
- *Выберем параметры алгоритма* $min\_supp = 0.4$ *и* $n = 3$. *Это значит, что каждый тестовый объект будет классифицироваться 3 правилами, посылками которых будут замкнутые множества признаков с относительной поддержкой не менее 0.4.*

Заметим, что в обучающем контексте доля положительных объектов равна 0.6 (6 из 10).

Построим все формальные понятия обучающего контекста $\mathbb{K}_{train}$ *с мощностью объемов не менее 4 (т.к.* $min\_supp * |G_{train}| = 0.4 * 10 = 4$*). Также для всех формальных понятий посчитаем среднее значение неопределенности Джини соответствующего классифицирующего правила.*

Поясним, как это делается, на примере формального понятия ({1,3,5,9},$\{\overline{w}, \overline{tm}\}$).

- *Составим сводную таблицу по одновременному наличию признаков* $\{\overline{w}, \overline{tm}\}$, *а также по наличию признака целевого класса play. См. Таблицу 3.*
- *Поскольку большинство объектов, имеющих признаки* $\{\overline{w}, \overline{tm}\}$ *одновременно, положительны (также имеют признак "play"), породим с помощью формального понятия* ({1,3,5,9},$\{\overline{w}, \overline{tm}\}$) *классифицирующее правило "*$\overline{w}, \overline{tm} \to play$*".*
- *Для такого правила среднее значение неопределенности Джини равно* $\frac{1+3}{10} * Gini(\frac{1}{4}, \frac{3}{4}) + \frac{3+3}{10} * Gini(\frac{1}{2}, \frac{1}{2}) = 0.4 * (1 - (\frac{1}{4})^2 - (\frac{3}{4})^2) + 0.4 * (1 - (\frac{1}{2})^2 - (\frac{1}{2})^2) = 0.45$.

Топ-10 классифицирующих правил в порядке возрастания средней неопределенности Джини правила (т.е. в порядке "ухудшения" правил) показаны в Таблице 4.

Чтобы определить метки тестового объекта 11, проведем следующие действия согласно Алгоритму 1:

	Классифицирующее правило	**Средняя неопределенность Джини**
1	$os, \neg tc, \neg hn \rightarrow_{(1)} +$	0.171
2	$\neg os, \neg w \rightarrow_{(1)} +$	0.267
3	$\neg oo, \neg tm, w \rightarrow_{(1)} -$	0.3
4	$os, \neg tc, \neg hn, \neg w \rightarrow_{(1)} -$	0.3
5	$os, th, \neg hn, \rightarrow_{(1)} -$	0.3
6	$os \rightarrow_{(0.75)} -$	0.317
7	$\neg oo, \neg tc, \neg hn \rightarrow_{(0.75)} -$	0.317
8	$\neg or, \neg tc, \neg hn \rightarrow_{(0.75)} -$	0.317
9	$\neg os \rightarrow_{(0.83)} +$	0.317
10	$or, \neg th, \neg w \rightarrow_{(1)} +$	0.343

Таблица 4: 10 лучших классифицирующих правил, полученных нахождением формальных понятий контекста из Таблицы 2.

Классифицирующее правило	**Средняя неопределенность Джини**
$os \rightarrow_{(0.75)} -$	0.317
$\neg oo \rightarrow_{(0.5)} -$	0.4
$\neg th, hn \rightarrow_{(0.5)} -$	0.4

Таблица 5: 3 “лучших” правила для классификации объекта *Outlook=sunny, Temperature=mild, Humidity=normal, Windy=true*

1. *Отбираем среди найденных 3 первых формальных понятия, содержания которых являются подмножествами множества признаков объекта 11 (Outlook=sunny, Temperature=mild, Humidity=normal, Windy=true)* – $\{\bar{or}, \bar{oo}, os, \bar{tc}, tm, \bar{th}, hn, w\}$

2. *Составляем на их основе 3 “лучших” правила, которые показаны в Таблице 5.*

3. *Найденные правила определяют значение 0 целевого признака для объекта “Outlook=sunny, Temperature=mild, Humidity=normal, Windy=true”, поскольку* $\frac{1}{3}(0.25 + 0.5 + 0.5) \approx 0.41 < 0.6$.

Модификация подхода к классификации с помощью формальных (узорных) понятий для данных со сложной структурой описана в Алгоритме 2.

На вход алгоритму подаются обучающая и тестовая узорные структуры $PS_{train} = (G_{train}, ((D, \sqcap), c_{train}), \delta_{train})$ и $PS_{test} = (G_{test}, (D, \sqcap), \delta_{test})$. Алгоритм использует модификацию алгоритма “Замыкай по-Одному” ($CbO_{PS}(PS, min\_supp)$) (С.О. Кузнецов, 1993), в которой выдаются все узорные понятия узорной структуры PS, поддержки которых ограничены снизу значением параметра min_supp. Для выбора классифицирующих правил используется критерий $inf : D \times c_{train} \rightarrow \mathbb{R}$ типа неопределенности Джини или энтропийного прироста информации (в программной реализации по умолчанию – среднее значение неопределенности Джини). Параметры алгоритма: min_supp и n – минимальная поддержка классифицирующих правил и число правил, используемых для классификации тестового объекта.

Алгоритм состоит из следующих шагов:

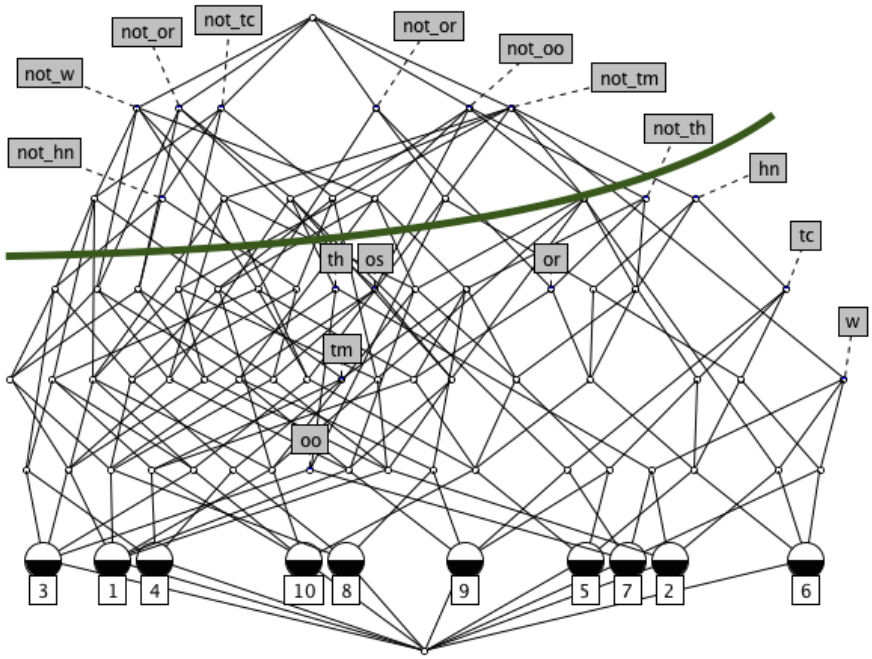


Рисунок 2: Решетка формальных понятий, соответствующая обучающему контексту из Примера 2. Выше зеленой линии лежат формальные понятия с минимальной относительной поддержкой 0.4.

1. Инициализировать c_{test} пустым списком, а r_{test} – пустым словарем. В c_{test} будут добавляться предсказанные значения целевого признака для тестовых объектов, а в r_{test} – правила для каждого тестового объекта (ключ в словаре – номер объекта, значение – список правил).

2. Посчитать долю положительных объектов в выборке $c_{pos} = \frac{|c'_{train}|}{|G_{train}|}$.

3. С помощью алгоритма $CbO_{PS}(PS, min\_supp)$ найти все узорные понятия обучающей узорной структуры PS_{train} со значением поддержки не менее min_supp. Параллельно с этим для каждого узорного понятия вычислять значение качества соответствующего классифицирующего правила inf. Таким образом, получится словарь $\mathcal{S}$, ключами которого будут содержания узорных понятий, а значениями – соответствующие значения функционала inf.

4. Отсортировать все узорные понятия $\mathcal{S}$ по посчитанным значениям критерия inf в порядке “улучшения” (то есть по возрастанию inf, если малые значения критерия говорят о хороших правилах (как в случае неопределенности Джини) или по убыванию, если, наоборот, большие значения критерия свидетельствуют о хороших правилах (прирост информации, среднее уменьшение Джини)).

5. Для каждого тестового объекта $g_t \in G_{test}$:

   – Отобрать n_{rules} “подходящих” содержаний формальных понятий, то есть $\{d_i\}_{i\in[1,n_{rules}]} = \{d \mid (A,d) \in \mathcal{S}, g_t^{\diamond} \sqsubseteq B\}$

   – Для каждого из отобранных содержаний формальных понятий $\{d_i\}_{i\in[1,n_{rules}]}$ определить долю положительных объектов $c_i = \frac{|d_i^{\diamond} \cap c'_{train}|}{|d_i^{\diamond}|}$

– Сформировать таким образом набор правил $\{d_i \to_{c_i} +\}_{i\in[1,n_{rules}]}$. Записать его в словарь r_{test} для ключа t (номер объекта g_t).

– Предсказанное значение целевого признака c_{train_t} определить как индикатор того, что усредненное заключение найденных правил превышает долю положительных объектов во всей выборке:

$$c_{train_t} = [\frac{1}{n\_rules}\sum_{i=1}^{n\_rules} c_i \geq c_{pos}].$$

Добавить это значение в c_{test}.

Algorithm 2 Concept Lattice-Based Rule-learner (CoLiBRi) – случай данных со сложной структурой.

Input: $PS_{train} = (G_{train}, ((D, \sqcap), c_{train}), \delta_{train})$
$PS_{test} = (G_{test}, (D, \sqcap), \delta_{test})$
$min\_supp \in \mathbb{R}^+$, $n_{rules} \in \mathbb{N}$;
$CbO_{PS}(PS, min\_supp) : PS \to \mathcal{S}$;
$inf : D \times c_{train} \to \mathbb{R}$;
$sort(\mathcal{S}, inf) : \mathcal{S} \to \mathcal{S}$

Output: c_{test}, r_{test}

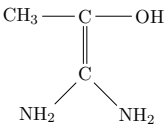
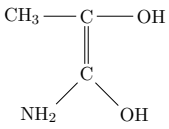
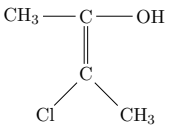
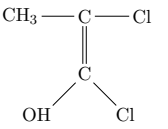
$c_{test} = \emptyset, r_{test} = \emptyset$
$c_{pos} = \frac{|c'_{train}|}{|G_{train}|}$
$\mathcal{S} = \{(A,d) : inf(d, c_{train}) \mid A \subseteq G_{train}, d \in D, A^\diamond = d, d^\diamond = A, |A| \geq min\_supp\} = CbO_{PS}(PS_{train}, min\_supp)$
$\mathcal{S} = sort(\mathcal{S}, inf)$
for $g_t \in G_{test}$ **do**
  $\{d_i\}_{i\in[1,n_{rules}]} = \{d \mid (A,d) \in \mathcal{S}, g_t^\diamond \sqsubseteq d\}$
  $c_i = \frac{|d_i^\diamond \cap c'_{train}|}{|d_i^\diamond|}$
  $r_{test}[i] = \{d_i \to_{c_i} +\}_{i\in[1,n_{rules}]}$
  $c_{test}[i] = [\frac{1}{n\_rules}\sum_{i=1}^{n\_rules} c_i \geq c_{pos}]$
end for

Пример 3. *В задаче предсказания наличия некоторого свойства P химических веществ дана обучающая выборка в виде упрощенной молекулярной структуры 4 положительных веществ и 3 отрицательных веществ. Про положительные объекты известно, что они обладают свойством P, про отрицательные известно, что нет. Для тестовых объектов необходимо сделать прогноз, обладают ли они свойством P.*

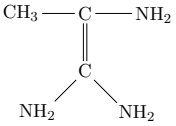
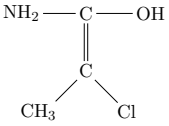
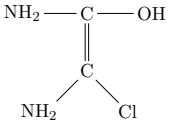
	Классифицирующее правило	Объекты	Средняя неопределенность Джини
1	$\{CH_3 - C = C, OH - C = C\} \xrightarrow[(0.8)]{} +$	1,2,3,4 | 6	0.22
2	$\{C = C - NH_2\} \xrightarrow[(0.4)]{} +$	1,2 | 5,6,7	0.34
3	$\{C = C - CH_3\} \xrightarrow[(0.67)]{} +$	1,2,3,4 | 5,6	0.38
4	$\{C = C - OH\} \xrightarrow[(0.67)]{} +$	1,2,3,4 | 6,7	0.38
5	$\{CH_3 - C = C - OH\} \xrightarrow[(0.75)]{} +$	2,3,4 | 6	0.4
6	$\{CH_3 - C = C - NH_2\} \xrightarrow[(0.5)]{} +$	1,2 | 5,6	0.47
7	$\{C = C\} \xrightarrow[(0.57)]{} +$	1,2,3,4 | 5,6,7	0.49

Таблица 6: Классифицирующие правила в Примере 3. Символом | отделены положительные объекты от отрицательных.

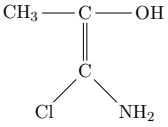
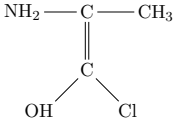
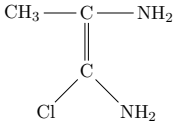
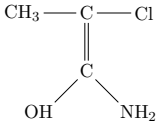
Положительные объекты:

1 :  2 :  3 :  4 : 

Отрицательные объекты:

5 :  6 :  7 : 

Тестовые объекты:

8 :  9 :  10 :  11 : 

Доля положительных объектов в обучающей выборке равна 0.57 (4 из 7).

Найдем все узорные понятия обучающей узорной структуры $PS_{train} = (G_{train}, ((D, \sqcap), c_{train}), \delta_{train})$ *с абсолютной поддержкой не менее 4* ($min\_supp = \frac{4}{7}$). *Здесь* $G_{train} = \{1, \ldots, 7\}$, D *– пространство всех помеченных графов,* $\sqcap$ *– полурешёточная операция для помеченных графов, функция* δ_{train} *задана выше, а* $c_{train} = \{+, +\,, +\,, +\,, -\,, -\,, -\}$. *Правила, построенные на основе найденных узорных понятий, указаны в Таблице 6. Если делать прогнозы с помощью 3 лучших правил* ($n\_rules = 3$), *то объекты 8,9,11 классифицируются положительно* ($\frac{1}{3}(0.8 + 0.4 + 0.66) \approx 0.62 > 0.57$), *а объект 10 – отрицательно* ($\frac{1}{3}(0.4 + 0.66 + 0.5) \approx 0.52 < 0.57$).

В **четвертой главе** описывается разработанный программный комплекс, реализующий алгоритмы, описанные в третьей главе, затем приводятся результаты вычислительных экспериментов с наборами данных репозитория UCI (UC Irvine Machine Learning Repository)[2] – крупнейшего репо-

[2] http://archive.ics.uci.edu/ml/

зитория реальных и модельных задач машинного обучения. Также приводятся результаты экспериментов в задачах прогнозирования свойств химических веществ.

Структура основных классов программного комплекса CoLiBRi, реализующего алгоритмы, описанные в Главе 3, представлена на Рис. 3. На схеме стрелки синего цвета соответствуют отношению “быть наследником класса”, а стрелки черного цвета – отношению “задействовать”.

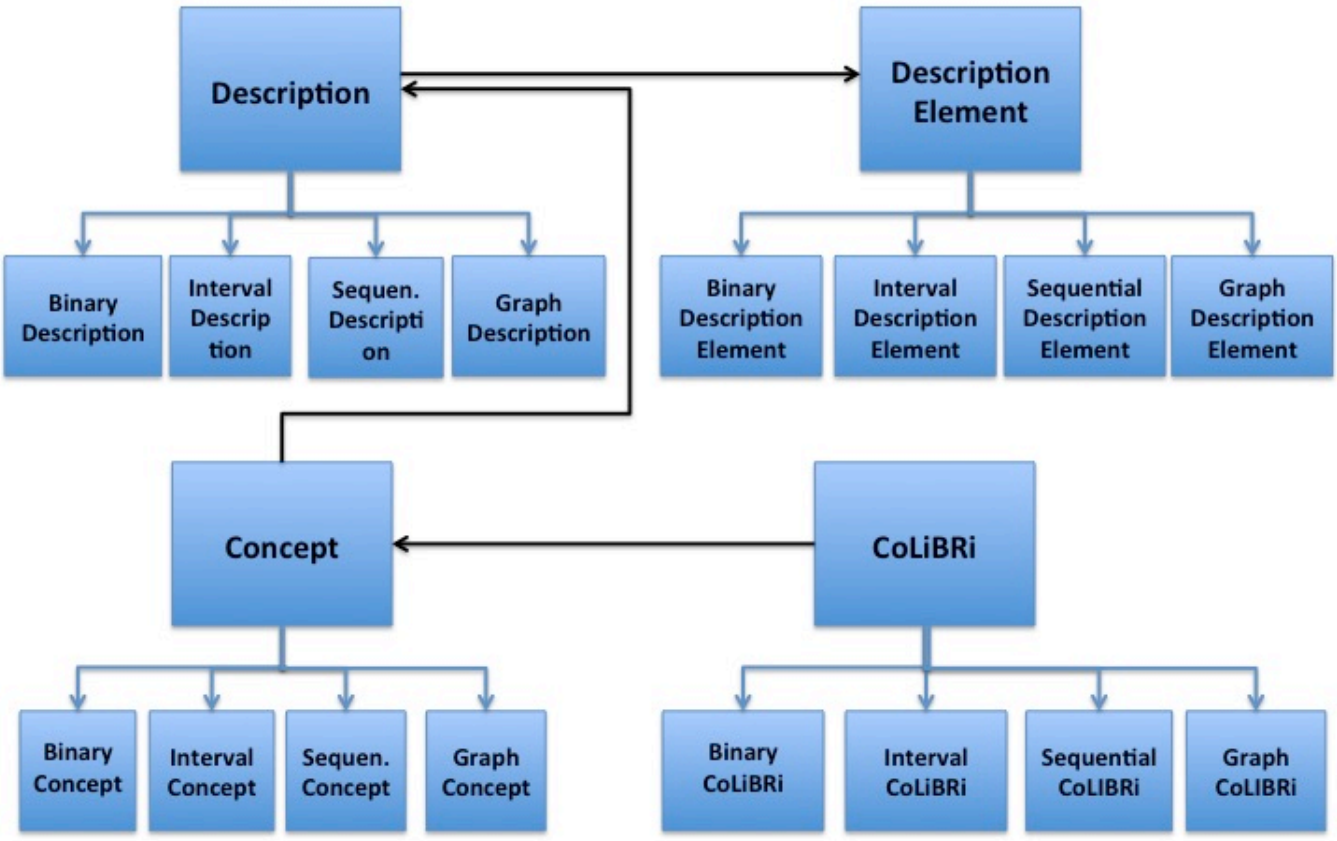
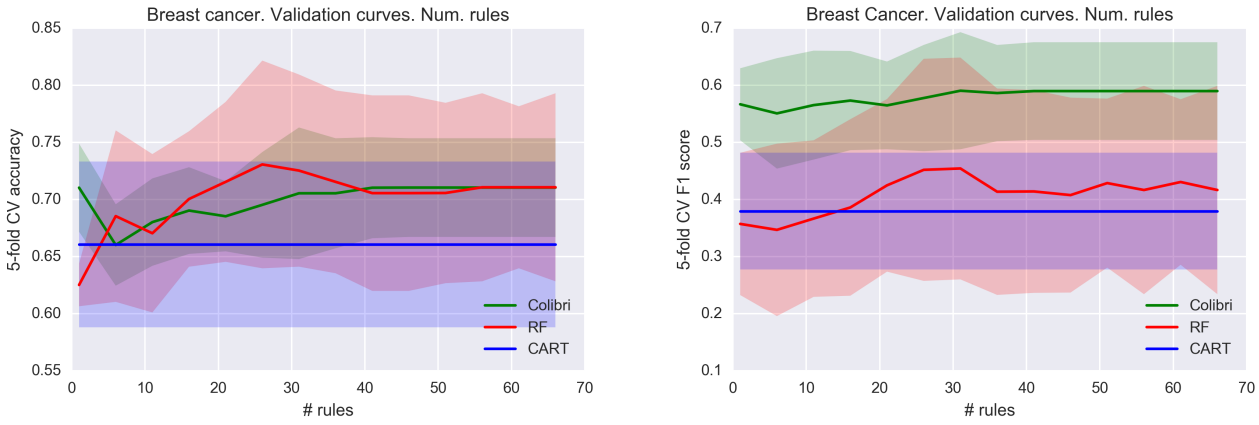


Рисунок 3: Структура основных классов программного комплекса CoLiBRi.

Имеются 4 абстрактных класса: DescriptionElement, Description, Concept и CoLiBRi. У каждого из них, в свою очередь, есть по 4 наследника. Технические детали программной реализации описаны в Разделе 4.2. диссертационной работы.

Версия алгоритма CoLiBRi (“Concept Lattice-Based Rule-learner”) для работы с бинарными признаками (Алгоритм 1) была протестирована на 13 наборах данных UCI[3]. Сравнение проводилось с реализациями Scikit-learn[4] алгоритмов построения деревьев решений CART (Breiman, 1984), случайного леса (Breiman, 2001), а также с методом ближайших соседей. Для каждого набора данных решалась задача бинарной классификации, где выделялись самый частый класс и все остальные. Категориальные признаки были преобразованы в бинарные методом One Hot Encoding. Отслежива-

[3] http://repository.seasr.org/Datasets/UCI/csv/
[4] http://scikit-learn.org/stable/index.html



(a) Доля верных ответов (b) F1

Рисунок 4: Кривые валидации по числу правил (для CoLiBRi) или деревьев (для случайного леса) в сравнении с деревом решений CART. 5-кратная стратифицированная кросс-валидация для набора данных Breast Cancer репозитория UCI.

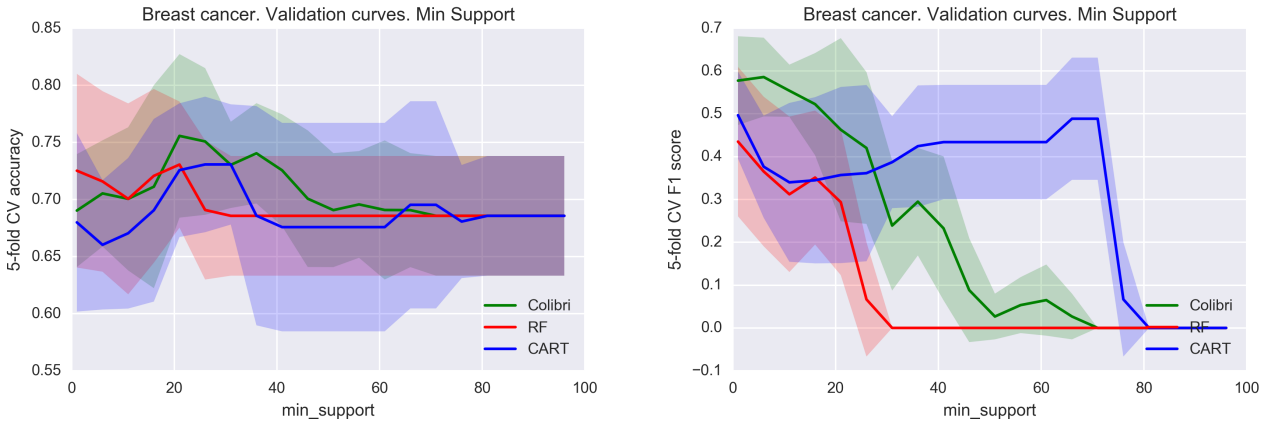
лись значения доли правильных ответов и F1-метрики при 5-кратной кросс-валидации. Результаты представлены в Таблице 7.

Данные	DT acc	RF acc	kNN acc	CoLiBRi acc	DT F1	RF F1	kNN F1	CoLiBRi F1
audiology	0.75	**0.8**	0.63	0.79*	0.71	**0.74**	0.58	**0.74**
breast-cancer	0.63	0.66	**0.76**	0.65	0.58	0.63	**0.75**	0.61
breast-wisc	0.7	0.74	0.73	**0.76**	**0.45**	0.42	0.38	0.44*
car	0.75	0.78*	0.71	**0.79**	0.75	**0.76**	0.71	**0.76**
hayses-roth	0.84*	0.83*	0.49	**0.86**	0.84*	0.82	0.49	**0.85**
lymph	0.8	0.83	**0.86**	0.83	0.77	**0.85**	0.84*	0.84*
mol-bio-prom	0.78	**0.83**	**0.83**	0.82*	0.78	**0.84**	0.8	0.83*
nursery	0.64	0.65	**0.72**	0.65	0.62	0.62	**0.7**	0.62
primary-tumor	0.41	**0.46**	0.41	0.45*	0.37	**0.41**	0.37	0.4*
solar-flare	0.7*	0.7*	0.63	**0.72**	0.67	0.69*	0.6	**0.71**
soybean	0.91*	0.91*	**0.92**	0.91*	0.91*	**0.93**	0.92*	0.91*
spect-train	0.61	0.69	0.68	0.7	0.34	0.36	0.23	0.38
tic-tac-toe	0.79	0.79	**0.85**	0.78	0.82	0.86	**0.89**	0.85

Таблица 7: Значения доли правильных ответов и F1-метрики для 13 наборов данных репозитория UCI. “DT acc” и “DT F1” означают средние по 5 запускам доли правильных ответов и F1-метрики алгоритма CART при 5-кратной кросс-валидации , . . . , “CoLiBRi F1” означает среднее по 5 запускам значение F1-метрики алгоритма CoLiBRi при 5-кратной кросс-валидации. Жирным выделены лучшие значения метрик, звездочками отмечены значения, которые не являются статистически значимо уступающими лучшим.

Также изучалась зависимость качества алгоритмов от значений параметров. Для этого были построены кривые валидации по числу правил, минимальной поддержке и максимальной мощности посылки правил для наборов данных репозитория UCI. Для набора данных Breast Cancer кривые валидации по числу правил представлены на Рис. 4a (доля правильных ответов) и 4b (F1-метрика), а по минимальной поддержке – на Рис. 5a (доля правильных ответов) и 5b (F1-метрика).

Распределения мощностей посылок правил (“длин” правил), которыми определялись метки тестовых объектов для 3 наборов данных UCI и для 3 алгоритмов (CART, RF и CoLiBRi) показаны в



(a) Доля верных ответов (b) F1

Рисунок 5: Кривые валидации по минимальной поддержке для CoLiBRi, случайного леса и дерева решений CART. 5-кратная стратифицированная кросс-валидация для набора данных Breast Cancer репозитория UCI.

виде “ящиков с усами” (boxplots) на Рис. 6, 7 и 8. Средние мощности посылок правил для 13 наборов данных UCI и 3 алгоритмов показаны на Рисунке 9.

Видно, что в целом правила, получаемые с CoLiBRi сравнимы с теми, что порождаются алгоритмом CART (хотя у дерева решений это одно правило для одного объекта, а у CoLiBRi – несколько), но короче, чем у случайного леса. Это делает алгоритм CoLiBRi более интерпретируемым, чем случайный лес, если речь идет об интерпретации отнесения конкретного тестового объекта к одному из классов. Заметим, что длину правил CoLiBRi можно еще сильнее понизить, если для посылки каждого правила считать соответствующий минимальный генератор.

Версия алгоритма CoLiBRi (“Concept Lattice-Based Rule-learner”) для работы с описаниями в виде последовательностей (Алгоритм 2) была протестирована в серии экспериментов с данными в виде последовательностей.

Рассматривались 7 наборов данных. Подробно эти задачи описаны в (Moerchen, Fradkin, 2010). Далее в Таблице 8 приведены средние доли правильных ответов при 10-кратной кросс-валидации для 7 алгоритмов и 7 задач классификации. Описания алгоритмов даны на следующих ресурсах: http://misere.co.nf/, http://adrem.ua.ac.be/scii. Результаты позволяют утверждать, что качество классификации метода SequentialCoLiBRi достаточно высокое в сравнении с прочими алгоритмами классификации последовательностей.

Также проводились вычислительные эксперименты еще с 4 алгоритмами и 5 наборами данных, представленных графами.

Наборы данных IMDB, MUTAG, NCI, NCI109 и PROTEINS[5] известны тем, что в задачах классификации с этими данными часто проверяются алгоритмы графовой классификации.

Краткое описание задач:

– IMDB – граф отношения совместной съемки в фильме для актеров; фильмы поделены на 2 жанра: романтические и боевики;

[5] https://ls11-www.cs.uni-dortmund.de/staff/morris/graphkerneldatasets

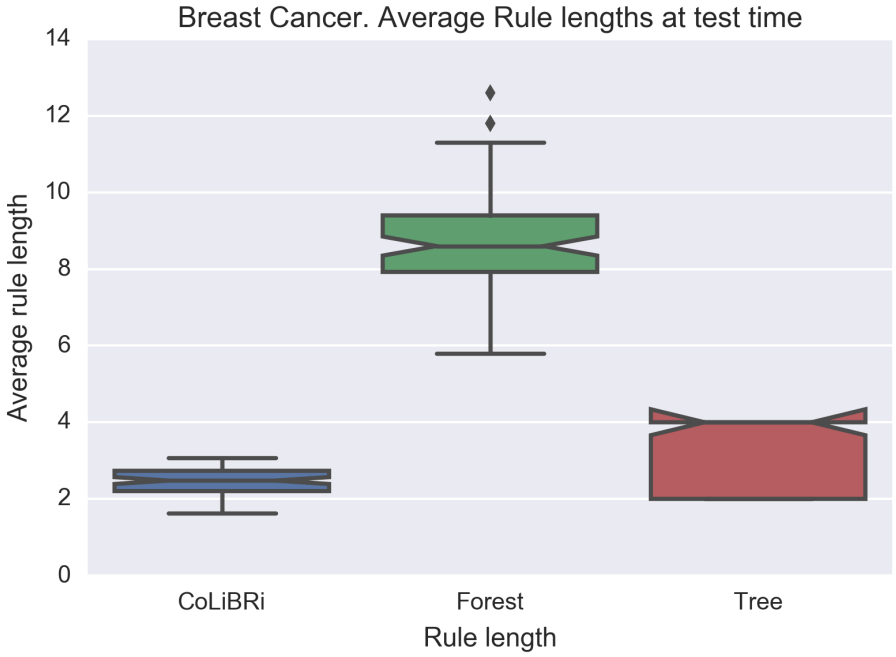


Рисунок 6: Средние мощности посылок правил, которыми были классифицированы тестовые объекты набора данных Breast Cancer репозитория UCI, для 3 алгоритмов.

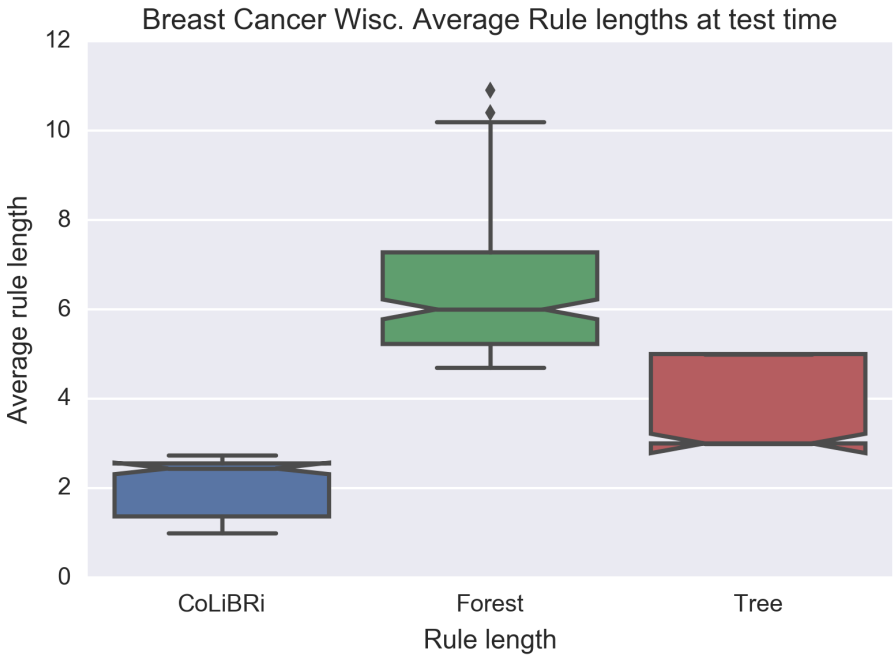


Рисунок 7: Средние мощности посылок правил, которыми были классифицированы тестовые объекты набора данных Breast Cancer Wisconsin репозитория UCI, для 3 алгоритмов.

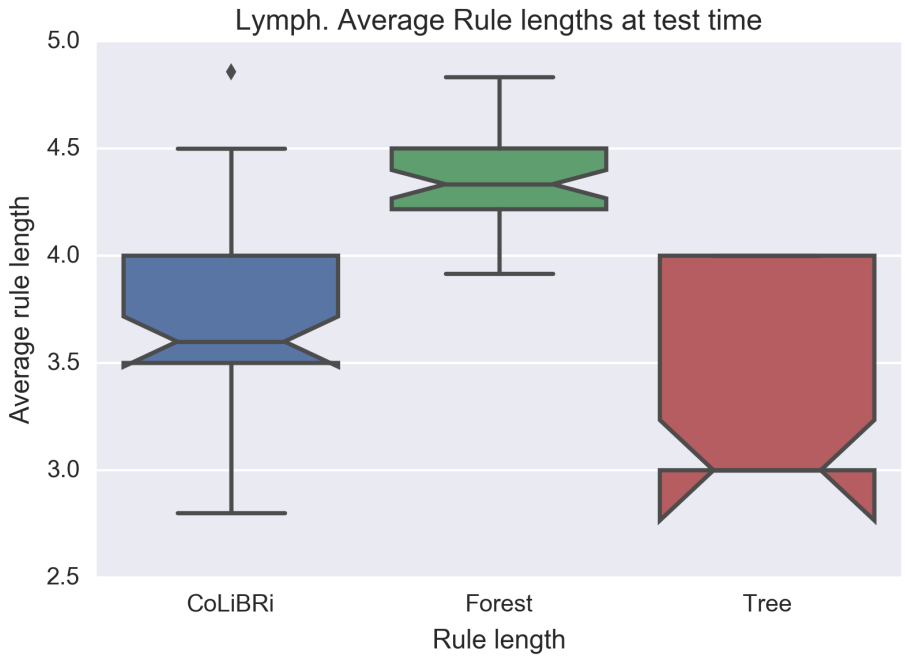


Рисунок 8: Средние мощности посылок правил, которыми были классифицированы тестовые объекты набора данных Lymph репозитория UCI, для 3 алгоритмов.

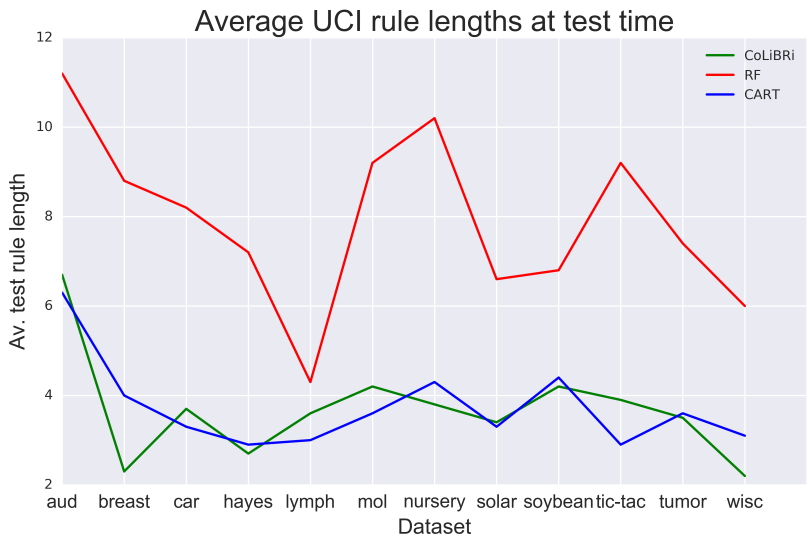


Рисунок 9: Средние мощности посылок правил, которыми были классифицированы тестовые объекты, для 3 алгоритмов и 13 наборов данных репозитория UCI (лучше смотреть в цвете).

	CBS	BayesFM	SCII Match	SCII CBS	MiSeRe	Binary CoLiBRi	Sequential CoLiBRi
aslbu	0.43	**0.7**	0.57	0.56	**0.7**	0.48	0.62
aslgt	0.23	0.738	0.04	0.04	**0.77**	0.32	0.71
auslan	0.32	0.34	0.04	0.03	0.34	0.33	**0.35**
blocks	**1**	**1**	0.08	0.08	**1**	0.99	**1**
context	0.58	0.896	0.32	0.33	**0.9**	0.74	**0.9**
pioneer	0.79	0.96	0.97	0.95	**1**	0.77	0.97
skater	0.55	**0.87**	0.18	0.18	0.86	0.69	**0.87**

Таблица 8: Доля верных ответов при 10-кратной кросс-валидации в задачах классификации последовательностей.

	CBA	DT	SVM graphlet	CoLiBRi
IMDB	60.1	55.6	**62.1**	59.3
MUTAG	72.1	68.4	**77.4**	74.6
NCI1	55.1	52.1	**59.6**	58.3
NCI109	56.6	52.8	**59.7**	58.8
PROTEINS	60.5	60.2	66.3	**68.9**

Таблица 9: Доли правильных ответов 4 алгоритмов на 5 графовых наборах данных.

– MUTAG – 188 структур химических веществ, поделенных на 2 класса по мутагенному эффекту, производимому на бактерии;

– NCI, NCI109 – два сбалансированных подмножества наборов данных химических соединений, у которых измерена, соответственно, активность борьбы против немелкоклеточного рака легких и раковых клеток яичников;

– PROTEINS – предсказание функциональных классов принадлежности ферментов.

Для всех графов были построены бинарные признаки по включению подграфов до 6 вершин. Проверялись 4 алгоритма: CBA – классификация на основе ассоциативных правил (реализация LUCS-KDD[6]), DT – дерево решений (sklearn[7]), SVM graphlet – метод опорных векторов (sklearn), CoLiBRi – предлагаемый алгоритм.

Данные были поделены в пропорции 7/3 на обучающую и проверочную выборку. В Таблице 9 указаны доли правильных ответов 4 алгоритмов проверенных на 5 графовых наборах данных. Можно заметить, что в целом SVM справляется лучше остальных алгоритмов, зато остальные алгоритмы — интерпретируемые, на выходе можно получить набор классифицирующих правил для каждого тестового примера.

В Таблице 10 представлены средние мощности посылок правил, участвовавших в классификации тестовых примеров в задачах классификации, результаты которых представлены в Таблице 9. Можно сделать вывод, что в данных задачах алгоритм CoLiBRi демонстрирует качество классификации выше, чем CBA и DT, при этом сохраняется интерпретируемость алгоритма (в отличие от

[6] http://cgi.csc.liv.ac.uk/~frans/KDD/Software/CBA/cba.html
[7] http://scikit-learn.org

	CBA	DT	CoLiBRi
IMDB	5.1	5.2	5.5
MUTAG	6.8	7.8	7.2
NCI1	8.3	10.5	12.7
NCI109	8.5	11.3	10.5
PROTEINS	7.6	12.2	8.6

Таблица 10: Средние мощности посылок правил, участвовавших в классификации тестовых примеров.

случая применения SVM) – мощности посылок правил, участвовавших в классификации тестовых примеров в случае CoLiBRi примерно такие же, как и в случае CBA и DT.

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

1. Предложен универсальный подход к классификации данных со сложной структурой на основе решеток замкнутых описаний;

2. В рамках этого подхода предложены алгоритмы для классификации данных, представленных последовательностями и графами, а также числовыми и интервальными признаками;

3. Алгоритмы апробированы в задачах классификации последовательностей и графов и показали высокие значения доли правильных ответов. При этом классификация проводилась с помощью коротких классифицирующих правил;

4. На данных Predictive Toxicology Challenge показаны метрики качества выше, чем у SVM с графлет-ядром, и сравнимые с лучшими из результатов участников соревнования;

5. В вычислительных экспериментах с данными репозитория UCI получены значения метрик качества классификации на кросс-валидации, статистически значимо более высокие, чем у алгоритмов построения деревьев решений;

6. При этом показано, что интерпретируемость полученных правил, понимаемая как средняя мощность посылок правил, которыми определялись метки тестовых объектов, у предлагаемого алгоритма лучше, чем у случайного леса;

7. Методы классификации, основанные на правилах, в том числе деревья решений, представлены с помощью проекций интервальных узорных структур;

8. Предложены и исследованы дискретизирующие проекции для интервальных узорных структур. На их основе предложен способ выбора правил на основе множеств формальных понятий, гарантирующий нахождение правил не хуже, чем построенные деревом решений, по выбранному критерию информативности;

9. Разработан программный комплекс, позволяющий анализировать сложно структурированные данные и решать для них задачи классификации с помощью интерпретируемых наборов правил, подходящих для дальнейшего экспертного анализа.

Публикации автора по теме диссертации

Публикации по теме диссертации в изданиях, входящих в **перечень ВАК**:

1. Кашницкий Ю. С., Игнатов Д. И. Ансамблевый метод машинного обучения, основанный на рекомендации классификаторов // Интеллектуальные системы. Теория и приложения. 2015. Т. 19. № 4. С. 37–55;

Прочие публикации, индексируемые в базе данных научного цитирования Scopus:

2. Masyutin A. and Kashnitsky Y. (2017). Query-Based Versus Tree-Based Classification: Application to Banking Data. Foundations of Intelligent Systems. LNAI 10352, pages 664–673;

3. Kashnitsky Y. and Kuznetsov S. O. (2016). Global Optimization in Learning with Important Data: an FCA-Based Approach, in: CLA 2016: Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications. CEUR Workshop Proceedings / Ed. by M. Huchard, S. O. Kuznetsov. Vol. 1624. Higher School of Economics, National Research University, Ch. 19, pages 189–202.

4. Kashnitsky, Y. and Kuznetsov, S. O. (2016). Interval Pattern Concept Lattice as a Classifier Ensemble. Proceedings of the 5^{th} Workshop "What FCA can do for Artificial Intelligence" collocated with ECAI 2016, CEUR Workshop Proceedings, volume 1703, pages 105–112;

5. Kashnitsky, Y. (2016). Lazy Learning of Succinct Classification Rules for Complex Structure Data. Supplementary Proceedings of the 5th International Conference on Analysis of Images, Social Networks and Texts (AIST-SUP 2016), Yekaterinburg, Russia, April 7-9, 2016, CEUR Workshop Proceedings, volume 1710, pages 73–84;

6. Kashnitsky, Y. and Kuznetsov, S. O. (2015). Lazy associative graph classification. Proceedings of the 4^{th} Workshop "What FCA can do for Artificial Intelligence" collocated with IJCAI 2016, CEUR Workshop Proceedings, volume 1430, pages 63—74;

7. Masyutin, A., Kashnitsky, Y., and Kuznetsov, S. O. (2015). Lazy classification with interval pattern structures: Application to credit scoring. In CEUR Workshop Proceedings, volume 1430, pages 43–54;

8. Kashnitsky, Y. and Ignatov, D. (2014). Can FCA-based recommender system suggest a proper classifier? In CEUR Workshop Proceedings, volume 1257, pages 17–26;