

На правах рукописи



Виноградов Дмитрий Вячеславович

**Вероятностно-комбинаторный
формальный метод обучения, основанный
на теории решеток**

Специальность 05.13.17 —
«Теоретические основы информатики»

Автореферат
диссертации на соискание учёной степени
доктора физико-математических наук

Москва — 2018

Работа выполнена в Федеральном государственном учреждении «Федеральный исследовательский центр «Информатика и управление» Российской академии наук»

Научный консультант: Д.т.н., профессор, заслуж. деятель науки РФ
ФИНН Виктор Константинович

Официальные оппоненты: **Васильев Станислав Николаевич**
академик РАН, д. ф.-м. н., профессор,
главный научный сотрудник Института
Проблем Управления им. В.А. Трапезникова РАН

Еремеев Александр Павлович
д. т. н., профессор,
заведующий кафедрой «Прикладной математики»
НИУ «Московский Энергетический Институт»

Устинин Михаил Николаевич
д. ф.-м. н., доцент/с.н.с.,
заместитель директора по филиалу Института
Прикладной Математики им. М.В.Келдыша РАН

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

Зашита состоится «28» февраля 2019 г. в 13.00 на заседании диссертационного совета Д 002.073.05 на базе ФИЦ ИУ РАН по адресу: 119333, г.Москва, ул.Вавилова, д.40.

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН, а также на сайте <http://www.frccsc.ru/>.

Автореферат разослан « » декабря 2018.

Ученый секретарь
диссертационного
совета Д 002.073.05
д.ф.-м.н., профессор

В.В. Рязанов

Общая характеристика работы

Диссертационная работа посвящена исследованию новой модели машинного обучения, использующего современные методы теории решеток (анализа формальных понятий). Предлагается новый вероятностно-комбинаторный формальный подход к интеллектуальному анализу данных, обладающих хорошей структурированностью, позволяющей определить такую операцию сходства, которая выявит некоторые структурные фрагменты, отвечающие за исследуемые целевые свойства.

Актуальность темы. В различных областях человеческой деятельности (социологии, истории, медицине, фармакологии, экономике, лингвистике, и др.) повседневно возникает необходимость решения задач анализа, прогноза и диагностики, выявления скрытых зависимостей и поддержки принятия рациональных решений. Из-за бурного роста объема информации, развития технологий ее сбора и хранения в базах данных (описываемых термином Big Data) точные методы анализа информации и моделирования исследуемых объектов нуждаются в автоматизации поддержки эксперта средствами интеллектуального анализа данных, машинного обучения, распознавания образов и классификации¹.

В большинстве случаев эти подходы используют выборки прецедентов (наборы описаний-наблюдений объектов, предметов, ситуаций или процессов) в качестве исходной информации, при этом каждый прецедент записывается в виде вектора значений отдельных его свойств-признаков.

К самым первым работам анализа данных по прецедентам можно отнести появившиеся в 30-х годах прошлого столетия труды основоположников математической статистики, заложивших основы байесовской теории принятия решений (Дж. Нейман, Э. Пирсон²), клас-

¹**Финн В.К.** Об интеллектуальном анализе данных // *Новости искусственного интеллекта*. – 2004. – № 3. – С. 3–18

²**Neyman, Jerzy and Egon S. Pearson.** On the Problem of the Most Efficient Tests of Statistical Hypotheses // *Philosophical Transactions of the Royal Society of London. Series A*,

сификации с использованием разделяющих функций (Р. Фишер³), теории проверки статистических гипотез (А. Вальд⁴).

В 50-х годах появились первые нейросетевые модели машинного обучения (перцептрон Ф. Розенблата⁵).

К концу 60-х годов уже были разработаны и детально исследованы различные подходы для решения задач ИАД в рамках статистических, нейросетевых моделей, и моделей с пороговыми функциями. Итоги данных и последующих исследований были представлены в ряде монографий^{6 7 8 9 10 11 12 13 14 15}.

Большой вклад в развитие теории ИАД внесли советские и российские ученые: М.А. Айзerman, Э.М. Браверман, Л.И. Розоноэр (метод потенциальных функций), В.Н. Вапник, А.Я. Червоненкис (статистическая теория обучения), Ю.И. Журавлев (алгоритмы вычисления оценок и алгебраическая теория распознавания¹⁶), Н.Г. За-

Mathematical and Physical Sciences., Vol. 231(694–706). – 1933. – p. 289–337

³**Fisher, R.A.** The use of multiple measurements in taxonomic problems // *Annals of Eugenics*, Vol. 7, – Part 2. – 1936. – p. 179–188

⁴**Wald, A.** Contributions to the theory of statistical estimation and testing of hypotheses // *Annals of Mathematical Statistics*, Vol. 10. – 1939. – p. 299–326

⁵**Розенблатт Ф.** *Принципы нейродинамики. Перцептроны и теория механизмов мозга.* Пер. с англ. – М.: Мир, 1965. – 480 с.

⁶**Айзerman М.А., Браверман Э.М., Розоноэр Л.И.** *Метод потенциальных функций в теории обучения машин.* – М.: Наука. – 1970. – 384 с.

⁷**Бонгард М.М.** *Проблема узнавания.* – М.: Наука. – 1967. – 320 с.

⁸**Вапник В.Н., Червоненкис А.Я.** *Теория распознавания образов (статистические проблемы обучения).* – М.: Наука, Гл. ред. физ.–мат. лит. – 1974. – 416 с.

⁹**Журавлев Ю.И., Рязанов В.В., Сенько О.В.** *«Распознавание». Математические методы. Программная система. Практические применения.* – М.: Фазис – 2006. – 176 с.

¹⁰**Загоруйко Н.Г.** *Методы распознавания и их применение.* – М.: Сов.радио. – 1972. – 206 с.

¹¹**Загоруйко Н.Г.** *Прикладные методы анализа данных и знаний.* – Новосибирск: Изд-во Института математики. – 1999. – 270 с.

¹²**Лбов Г.С.** *Методы обработки разнотипных экспериментальных данных.* – Новосибирск: Наука. – 1981. – 160 с.

¹³**Метод комитетов в распознавании образов.** (ред.: **Мазуров Вл.Д.**) – Свердловск: ИММ УНЦ АН СССР, – 1974. – 165 с.

¹⁴**Минский М., Пейперт С.** *Перцептроны.* Пер. с англ. – М.: Мир, 1971. – 261 с.

¹⁵**Нильсон Н.** *Обучающиеся машины.* Пер. с англ. – М.: Мир, 1967. – 180 с.

¹⁶**Журавлев Ю.И.** Об алгебраическом подходе к решению задач распознавания или классификации // *Проблемы кибернетики*. Вып. 33. – М.:Наука, 1978. – С. 5–68

горуйко (алгоритмы таксономии), Г.С. Лбов (логические методы распознавания и поиска зависимостей), Вл.Д. Мазуров (метод комитетов), В.Л. Матросов (статистическое обоснование алгебраического подхода к распознаванию¹⁷), К.В. Рудаков (теория алгебраического синтеза корректных алгоритмов¹⁸).

Интенсивные исследования проводятся с начала 80-х годов в ВИНИТИ АН СССР (потом в ВИНИТИ РАН, в настоящее время - в ФИЦ ИУ РАН). С 1981 года¹⁹ группа исследователей под руководством проф. В.К. Финна создала и развивает логико-комбинаторный ДСМ-метод автоматического порождения гипотез²⁰, в котором formalизованы различные когнитивные процедуры, основанные на понятии сходства.

ДСМ-метод назван так в честь известного английского философа, экономиста и логика Джона Стьюарта Милля. Используя технику многозначных логик, В.К. Финну с коллегами^{21 22}, удалось поставить систему индуктивной логики Милля²³ на четкие логические основания. Ключевым компонентом этого подхода является бинарная операция сходства²⁴. Следует указать, что примерно в это же

¹⁷Матросов В.Л. Синтез оптимальных алгоритмов в алгебраических замыканиях моделей алгоритмов распознавания // *Распознавание, классификация, прогноз.* – М.:Наука, 1989. – С. 149–176

¹⁸Рудаков К.В. Об алгебраической теории универсальных и локальных ограничений для задач классификации // *Распознавание, классификация, прогноз.* – М.:Наука, 1989. – С. 176–201

¹⁹Финн В.К. Базы данных с неполной информацией и новый метод автоматического порождения гипотез // В кн.: *Диалоговые и фактографические системы информационного обеспечения.* – М., 1981. – С. 153–156

²⁰ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания. (ред.: Финн В.К., Аншаков О.М.) – М.: Эдиториал УРСС – 2009. – 432 с.

²¹Аншаков О.М., Скворцов Д.П., Финн В.К. Логические средства экспертных систем типа ДСМ // *Семиотика и информатика.* – Вып. 28. – 1986. – С. 65–102

²²Аншаков О.М., Скворцов Д.П., Финн В.К. О дедуктивной имитации некоторых вариантов ДСМ-метода автоматического порождения гипотез // *Семиотика и информатика.* – Вып. 33. – 1993. – С. 164–233

²³Милль Дж.Ст. *Система логики сyllogistической и индуктивной: Изложение принципов доказательства в связи с методами научного исследования.* Пер. с англ. Изд. 5. – М.: Эдиториал УРСС, 2011. – 832 с.

²⁴Гусакова С.М., Финн В.К. Сходства и правдоподобный вывод // *Известия АН СССР, Сер. «Техническая кибернетика».* – 1987. – № 5. – С. 42–63

самое время аналогичный подход (но основанный не на логике, а на теории решеток) был разработан группой зарубежных исследователей под руководством проф. Рудольфа Вилле под названием анализ формальных понятий (АФП)²⁵. Однако отечественный подход включил в рассмотрение контр-примеры, чего не имеется у зарубежных авторов.

Второй когнитивной процедурой стало доопределение по аналогии, что превратило ДСМ-метод в средство интеллектуального анализа данных, когда после анализа прецедентов стало возможным применить приобретенное знание (гипотезы о причинах) для прогнозирования целевых свойств у ранее неизученных примеров.

Наконец, третья когнитивная процедура - абдуктивное принятие гипотез - возникло в трудах В.К. Финна в результате осмыслиения наследия известного американского математика и логика Чарльза Сэндерса Пирса²⁶.

После выяснения сути указанных когнитивных процедур проф. В.К. Финн создал единую систему, объединяющую все эти процедуры в одно целое. Эта система и получила название ДСМ-метод²⁷.

Следует признать, что имеются некоторые особенности ДСМ-метода, которые выдвигают вопрос о реализации вычислений для интеллектуального анализа данных на его основе.

Во-первых, множество порождаемых ДСМ-гипотез может оказаться экспоненциально велико по сравнению с размером обучающей выборки.

Во-вторых, С.О. Кузнецовым²⁸, М.И. Забежайло и др. были доказаны пессимистические оценки сложности для многих ДСМ-процедур (NP -полнота и $\#P$ -полнота).

²⁵Ganter, Bernhard and Rudolf Wille. *Formal Concept Analysis*. Transl. from German. – Berlin: Springer-Verlag, 1999. – 284 pp.

²⁶Пирс Ч.С. *Рассуждение и логика вещей: Лекции для Кэмбриджских конференций 1898 года*. Пер. с англ. – М.: РГГУ, 2005. – 371 с.

²⁷Финн В.К. Синтез познавательных процедур и проблема индукции // *Научная и техническая информация*, Сер. 2. – 1999. – № 1–2. – С. 8–45

²⁸Кузнецов С.О. Интерпретация на графах и сложностные характеристики задач поиска закономерностей определенного вида // *Научная и техническая информация*, Сер. 2. – 1989. – № 1. – С. 23–28

В-третьих, автор сумел обнаружить эффект «переобучения»: порождение так называемых фантомных ДСМ-гипотез. Эти фантомные гипотезы возникают тогда, когда вычисляется сходство двух (или более) обучающих примеров, каждый из которых имеет свой собственный механизм порождения целевого свойства. Это сходство оказывается фрагментом (набором общих признаков), который не является причиной исследуемого целевого свойства. Если же допустить его в процедуру предсказания эффекта у нового примера, предъявленного на прогноз, то он будет мешать корректному предсказанию. Подобный эффект «переобучения» характерен для многих методов машинного обучения, когда максимальный учет информации из обучающей выборки приводит к модели, демонстрирующей плохую предсказательную способность.

Чтобы справиться с возникающими проблемами, автором предлагается новый вероятностно-комбинаторный подход. Так как некоторые ингредиенты заимствованы мной из анализа формальных понятий (АФП), я назвал его вероятностно-комбинаторный формальный метод, сокращенно ВКФ-метод.

Цель диссертационной работы. Целью данной работы является исследовать модель машинного обучения, основанного на методах теории решеток, разработать вероятностные алгоритмы интеллектуального анализа данных для этого метода и исследовать математические свойства предложенных алгоритмов.

Научная новизна. Вероятностный подход к машинному обучению, основанному на методах теории решеток, до сих пор не исследовался.

Известные ранее детерминированные алгоритмы основывались на полном переборе возникающих сходств. Теоретическая оценка в этом случае пессимистична: возможно получение $O(2^n)$ различных битовых строк длины n с помощью побитого умножения на $n \times n$ бинарных матрицах. На практике это проявлялось как «экспоненциальный взрыв», когда из обучающей выборки, содержащей несколько

сотен примеров, порождалось более миллиона гипотез, даже уже сокращенных проверками дополнительных логических условий. Некоторые из этих гипотез только вредят предсказанию (наблюдается эффект «переобучения»). Изучение феномена «переобучения» в главе 2 также является новым.

Методы исследования. Для исследования нового вероятностно-комбинаторного метода машинного обучения, основанного на теории решеток, пришлось привлечь технику цепей Маркова, особенно, спаривающих цепей Маркова, производящих функций распределений вероятностей, теорию представлений групп.

Применяемые в работе методы относятся к области дискретной математики на стыке с алгеброй и теорией вероятностей. Все комбинаторные результаты имеют наглядный вероятностный смысл.

Теоретическая значимость. Математические результаты данной работы могут служить фундаментом для дальнейшего изучения предложенных вероятностных моделей и алгоритмов.

Наиболее интересной темой для дальнейших исследований, на взгляд автора, является вопрос о возможности полностью избавиться от «переобучения» посредством последовательного расширения обучающих выборок. Анализ производящих функций, полученных в теореме 5, возможно, приведет к разрешению этого вопроса.

Все полученные вероятностные результаты имеют наглядный алгоритмический смысл и приводят к значительному ускорению вычислений (оценка эффективности ленивых вычислений в теореме 1, применение остановленной «спаривающей» цепи Маркова из теоремы 10) или определению ключевых параметров (достаточное число сходств в теореме 13).

Практическая значимость. Разработанные математические модели, методы и алгоритмы позволяют организовать интеллектуальный анализ данных, основываясь как на малых, так и на больших выборках сложно структурированных обучающих примеров.

Малыми можно считать такие выборки, для которых все множество сходств может быть проанализировано экспертом. Большие выборки обеспечивают достаточный объем, чтобы статистические выводы могли быть сделаны с заданной надежностью.

Хотя диссертационная работа носит теоретический характер, автор проверил свои идеи путем применения созданной им программной системы, реализующий синтез описываемых вероятностных алгоритмов, к двум массивам (SPECT Hearts и Mushrooms) из репозитория данных для тестирования алгоритмов машинного обучения (UCI Machine Learning Repository).

Успешное применение к массиву Mushrooms (8124 объекта) позволяет надеяться, что предложенный подход сможет конкурировать с другими методами интеллектуального анализа «больших данных».

Апробация работы. Результаты работы неоднократно рассказывались на научных семинарах ФИЦ ИУ РАН и на конференциях:

- XIII Всероссийская конференция по искусственному интеллекту КИИ-2012, Белгород, 2012 ([18])
- 35 European Conference on Information Retrieval, Moscow, 2013 ([14])
- VI Мультиконференция по проблемам управления МКПУ-2013, с. Дивноморское, 2013 ([19])
- XIV Всероссийская конференция по искусственному интеллекту КИИ-2014, Казань, 2014 ([20])
- Conference on Analysis of Images, Social networks, and Texts AIST-2014, Ekaterinburg, 2014 ([15])
- Всероссийская конференция «Гуманитарные чтения РГГУ – 2014», Москва, 2014 ([2])
- VIII Мультиконференция по проблемам управления МКПУ-2015, с. Дивноморское, 2015 ([21])

- International Workshop «Formal Concept Analysis for Knowledge Discovery», Moscow, 2017 ([16])
- X Мультиконференция по проблемам управления МКПУ-2017, с. Дивногорское, 2017 ([22])
- XVI Всероссийская конференция по искусственному интеллекту КИИ-2018, Москва, 2018 ([17])

Материалы настоящей работы используются при чтении курсов лекций «Теория сходства в интеллектуальных системах» и «Интеллектуальный анализ данных и машинное обучение», читаемых студентам старших курсов Отделения интеллектуальных систем в гуманитарной сфере Российского Государственного Гуманитарного Университета.

Публикации. Публикации по теме диссертации в изданиях из списка, рекомендованного ВАК: [1–17].

Другие публикации автора по теме: [18–22].

Отдельные результаты включались в отчеты по проектам РФФИ

- 11-07-00618а «Интеллектуальные системы для наук о жизни и социальном поведении и стратегии когнитивного анализа данных» 2011-2013
- 14-07-00856а «ДСМ-метод автоматического порождения гипотез как средство конструирования интеллектуальных систем» 2014-2016
- 17-07-00539а «Интеллектуальная система для обнаружения эмпирических закономерностей в последовательностях баз фактов» 2017

и по программам Президиума РАН П15 за 2012-2014 гг.

Личный вклад автора. В диссертационной работе представлены только результаты, полученные лично автором: исследование феномена переобучения для комбинаторных методов, основанных на операции сходства (вероятности возникновения фантомного сходства при наличии контр-примеров), вероятностные алгоритмы машинного обучения, основанного на прикладной теории решеток, и их свойства. Из совместных публикаций в диссертацию включены лишь результаты автора.

Структура и объем работы. Диссертационная работа состоит из Введения, 4 глав, Заключения, списка используемых сокращений, словаря терминов и библиографии. Общий объем работы – 131 страница. Список литературы содержит 80 наименований.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, сформулированы научная новизна и теоретическая и практическая значимость представляемой работы.

Первая глава описывает основы современного раздела теории решеток - анализа формальных понятий (АФП) - позволяющего эффективно работать со сходством.

Сходство - бинарная операция, задающая структуру нижней полурешетки с наименьшим элементом. Мы ограничимся побитовым умножением как сходством на множестве битовых строк и столбцов.

Формальный контекст можно понимать как бинарное отношение между элементами множества O , которые мы называем *именами объектов*, и элементами множества F , которые мы называем *признаками*. Если в строчке, соответствующей объекту $o \in O$, и столбце, соответствующим признаку $f \in F$, стоит единица, то мы говорим,

что *объект* o *обладает признаком* f , и обозначаем это через oIf . В противном случае, говорим, что *объект* o *не имеет признака* f .

Для подмножества $A \subseteq O$ объектов-родителей их *сходством* (или *общим фрагментом*) называется подмножество признаков $A' = \{f \in F : \forall o \in A [oIf]\} \subseteq F$. Полагаем $\emptyset' = F$.

Эту операцию можно реализовать последовательным вычислением побитового умножения строк, соответствующих отобранным во множество A объектов, с отбором всех признаков, соответствующих единицам в результирующей строке.

Для подмножества $B \subseteq F$ признаков (фрагмента) их *сходством* называется подмножество $B' = \{o \in O : \forall f \in B [oIf]\} \subseteq O$ обучающих примеров (*родителей фрагмента*). Полагаем $\emptyset' = O$.

Эта операция вычисляется побитовым умножением отобранных столбцов.

Определение 1. Пару $\langle A, B \rangle$ назовем **кандидатом**, если $A = B' \subseteq O$ и $B = A' \subseteq F$.

Определение 2. Операция **замыкай-по-одному-вниз** на кандидате $\langle A, B \rangle$ и объекте $o \in O$ порождает пару

$$CbODown(\langle A, B \rangle, o) = \langle (A \cup \{o\})'', (A \cup \{o\})' \rangle.$$

Операция **замыкай-по-одному-вверх** на кандидате $\langle A, B \rangle$ и признаком $f \in F$ порождает пару

$$CbOUp(\langle A, B \rangle, f) = \langle (A \cup \{f\})', (A \cup \{f\})'' \rangle.$$

Ускорение вычислений достигается с помощью ленивой схемы для $CbODown$ и $CbOUp$.

Операция $CbODown(\langle A, B \rangle, o)$ содержит быстро вычисляемый фрагмент $(A \cup \{o\})' = B \cap \{o\}'$ и вычислительно трудный список родителей $(A \cup \{o\})'' = (B \cap \{o\})'$. Аналогично, операция $CbOUp$ имеет в своем составе потребляющую много времени компоненту $(B \cup \{f\})'' = (A \cap f')'$.

Предлагается лениво откладывать вычисления замыкания (второй производной), пока последовательный выбор нескольких объектов не сменится выбором признака (для $CbODown$) или серия выборов признаков не оборвется выбором объекта (для $CbOUp$).

Степень экономии оценивается так:

Теорема 1. *В ленивой схеме вычислений на каждую пару применений операции замыкания (одной в $CbOUp$ и одной в $CbODown$) в среднем в классической схеме мы будем делать $\frac{(n+k)^2}{k \cdot n} \geq 4$ операций замыкания.*

В параграфе 1.3 описан алгоритм порождения формального контекста для обучающей выборки, объекты которой описываются признаками, имеющими на множестве V значений структуру нижней полурешетки (с добавленным наименьшим элементом, интерпретируемым как отсутствие сходства по этому признаку).

Основной идеей описанного в параграфе 1.3 алгоритма представления битовыми строками является процедура сокращения формального контекста $\geq \subseteq V \times V$ до множества F всех \cup -неразложимых элементов в качестве бинарных атрибутов.

Определение 3. *Элемент $x \in L$ назовем \cup -неразложимым, если $x \neq \emptyset$ и для любых $y, z \in L$ если $y < x$ и $z < x$, то $y \cup z < x$.*

Основная теорема этого параграфа доказывает корректность этого алгоритма.

Вторая глава исследует эффект «переобучения» - возникновения фантомных сходств при ограничении на число родителей и запрете контр-примеров. *Фантомными* называем сходства, возникающие лишь из-за одновременного наличия общего фрагмента в нескольких объектах-родителях, каждый из которых имеет свою собственную фрагмент-причину. *Контр-примером* называем объекты, не обладающие целевым свойством.

Пусть число n обозначает количество сопутствующих (не входящих ни в какую причину) признаков, которыми мы ограничиваемся.

Для каждого контр-примера или обучающего примера независимо образуем последовательность n испытаний Бернулли с одинаковой вероятностью успеха p , чтобы сформировать появление сопутствующих признаков.

Теорема 2. Для $p \geq (-\ln(1 - \varepsilon)/n)^{1/b}$ вероятность появления фантомного сходства b случайных примеров не меньше, чем $\varepsilon > 0$.

Теорема 3. При числе сопутствующих признаков $n \rightarrow \infty$ и вероятности появления этих признаков у контр-примеров и обучающих примеров, равной $p = \sqrt{\frac{a}{n}}$ ($a \leq 1$), вероятность возникновения фантомного сходства двух случайных примеров, не устраниенного никаким из $t = c \cdot \sqrt{n}$ случайных контр-примеров, будет стремиться к

$$1 - e^{-a} - a \cdot e^{-a} \cdot \left[1 - e^{-c \cdot \sqrt{a}}\right] > 0.$$

Оценим вероятность переобучения с запретом контр-примеров и ограничением на число родителей через производящие функции вероятностей выживания разного количества контр-примеров, если одновременное определение фантомного сходства и случайных контр-примеров осуществляется по признакам (столбцам).

Определение 4. Назовем *выжившими* на шаге t контр-примеры $\langle y_1^k, \dots, y_t^k, \dots, y_n^k \rangle$, для которых $\forall j \leq t [a_j = 1 \Rightarrow y_j^k = 1]$.

Будем следить за числом $X_t^{(m)}$ контр-примеров, выживших после одновременного нахождения t -ых признаков m контр-примеров и фантомного сходства. Ясно, что это число должно быть элементом множества $S = \{0, 1, \dots, m\}$.

Производящие многочлены для распределений

$$\mathbf{P} \left[X_t^{(m)} = s \right]$$

будем обозначать через

$$\varphi_t^{(m)}(z) = \sum_{j=0}^m \mathbf{P} \left[X_t^{(m)} = j \right] \cdot z^j.$$

Теорема 4. $\varphi_n^{(m)}(z) = \sum_{j=0}^m \binom{m}{j} \cdot \prod_{t=1}^n \left[p_t^{b+j} + (1 - p_t^b) \right] \cdot (z - 1)^j$

Определение 5. *Двойной производящей функцией для $P[X_n^{(m)} = s]$ назовем формальный ряд*

$$\varphi_n(z, u) = \sum_{m=0}^{\infty} \sum_{s=0}^m P[X_n^{(m)} = s] \cdot z^s \cdot u^m = \sum_{m=0}^{\infty} \varphi_n^{(m)}(z) \cdot u^m.$$

Теорема 5. $\varphi_n(0, u) = \sum_{j=0}^{\infty} \prod_{t=1}^n \left[p_t^{b+j} + (1 - p_t^b) \right] \cdot \frac{(-u)^j}{(1-u)^{j+1}}.$

Третья глава посвящена исследованию вероятностных алгоритмов поиска кандидатов.

Data: множество обучающих (+)-примеров; внешние функции $CbOUp(,)$ и $CbODown(,)$ операций «замыкай-по-одному»

Result: случайный кандидат $\langle A, B \rangle$

$O := (+)$ -примеры, $F :=$ признаки; $I \subseteq O \times F$ - формальный контекст для (+)-примеров;

$R := O \cup F$; $Min := \langle O, O' \rangle$; $Max := \langle F', F \rangle$;

while ($Min \neq Max$) **do**

Выбираем случайный элемент $r \in R$;

if ($r \in O$) **then**

$Min := CbODown(Min, r)$;

$Max := CbODown(Max, r)$;

end

else

$Min := CbOUp(Min, r)$; $Max := CbOUp(Max, r)$;

end

end

$\langle A, B \rangle := Min$;

Algorithm 1: Спаривающая цепь Маркова

Состоянием цепи Маркова является упорядоченная пара ВКФ-кандидатов $\langle A_1, B_1 \rangle \leq \langle A_2, B_2 \rangle$, что выполняется при $B_1 \subseteq B_2$.

Первоначально меньший кандидат совпадает с наименьшим кандидатом $Min := \langle O, O' \rangle$, а больший - с наибольшим $Max := \langle F', F \rangle$.

В цикле к обоим кандидатам применяется одна и та же операция $CbODown$ с выбранным объектом, или $CbOUp$ с выбранным признаком.

Процесс останавливается, когда меньший кандидат совпадет с большим. Тогда этот общий кандидат и выдается алгоритмом 1.

Теорема 6. Алгоритм 1 соответствует цепи Маркова.

Определение 6. Состояние вида $\langle A, B \rangle = \langle A, B \rangle$ спаривающей цепи Маркова для совпадающей пары кандидатов называется **эрго-дическим**. Состояние вида $\langle A_1, B_1 \rangle < \langle A_2, B_2 \rangle$ называется **невоз-вратным**.

Теперь мы имеем теорему о конечности траекторий с вероятностью единица, как следствие классической теоремы о невозвратных состояниях:

Теорема 7. Вероятность того, что состояние

$$\langle A_1(t), B_1(t) \rangle \leq \langle A_2(t), B_2(t) \rangle$$

спаривающей цепи Маркова окажется невозвратным, стремится к нулю, когда $t \rightarrow \infty$.

Теперь, соединяя вместе алгоритм 1 и теорему 7, видим, что **с вероятностью единица алгоритм спаривающей цепи Маркова остановится**.

Хотя вопрос о среднем времени работы алгоритма 1 остался открытым, получены теорема 8 о среднем времени склеивания спаривающей цепи Маркова и теорема 9 о сильной концентрации около этого среднего для случая Булеана.

Этот случай возникает, когда рассматривается контекст: $O = \{o_1, o_2, \dots, o_n\}$ - множество объектов, $F = \{f_1, f_2, \dots, f_n\}$ - множество признаков, а отношение между ними задается как

$$o_i If_j \Leftrightarrow i \neq j. \quad (1)$$

Другими словами,

O	F	f_1	f_2	\dots	f_n
o_1	0	1	\dots	1	
o_2	1	0	\dots	1	
\vdots	\vdots	\vdots	\ddots	\vdots	
o_n	1	1	\dots	0	

Во время проведения экспериментов с ВКФ-системой был обнаружен феномен очень быстрого нахождения очередного кандидата. Хотя мы не смогли получить оценку в общем виде, для случая Булеана имеются результаты о среднем времени склеивания и сильной концентрации этого времени около своего среднего.

Теорема 8. *Среднее время склеивания для n -мерного гиперкуба равно*

$$E\left[\sum_{j=1}^n T_j\right] = \sum_{j=1}^n \frac{n}{j} \approx n \cdot \ln(n) + n \cdot \gamma + \frac{1}{2}.$$

Теорема 9. $P\left[\sum_{j=1}^n T_j \geq (1 + \varepsilon) \cdot n \cdot \ln(n)\right] \rightarrow 0$ при $n \rightarrow \infty$ для любого $\varepsilon > 0$.

В качестве практического средства для устранения наиболее длинных траекторий возможно применение следующей техники остановки алгоритма 1 и запуска его заново:

Определение 7. *Если T_1, \dots, T_r – независимые целочисленные случайные величины, одинаково распределенные с распределением времени склеивания T , то верхняя граница склеивания по r испытаниям определяется как $\hat{T} = T_1 + \dots + T_r$.*

Мы устранием те траектории спаривающей цепи Маркова, которые оказываются длиннее, чем \hat{T} (запускаем цепь заново).

Оценим соотношения вероятностей $\mu(R)$ и $\mu_{\hat{T}}(R)$ попадания в некоторое подмножество R эргодических состояний для обычной и остановленной по верхней границе \hat{T} спаривающих цепей Маркова:

Теорема 10. *Для любого $R \subseteq U$ с $\mu(R) = \rho$ и $r > \log_2(\rho+1) - \log_2(\rho)$ имеем $\mu_{\hat{T}}(R) \geq \rho - \frac{1}{2^r - 1}$ для верхней границы \hat{T} склеивания по $r > 1$ испытаниям.*

Спаривающая цепь Маркова, в общем случае, не является обратимой и стационарное распределение не является равномерным.

После склеивания алгоритм 1 будет совпадать со следующим алгоритмом.

Data: множество обучающих (+)-примеров; внешние функции $CbOUp(,)$ и $CbODown(,)$ операций «замыкай-по-одному»

Result: кандидат $\langle A, B \rangle$

$O := (+)$ -примеры, $F :=$ признаки; $I \subseteq O \times F$ - формальный контекст для (+)-примеров;

$A := O$; $B = O'$; $R := O \cup F$;

for ($i := 0$; $i < T$; $i = i + 1$) **do**

 Выбираем случайный элемент $r \in R$;

if ($r \in O$) **then**

$\langle A, B \rangle := CbODown(\langle A, B \rangle, r)$;

end

else

$\langle A, B \rangle := CbOUp(\langle A, B \rangle, r)$;

end

end

Algorithm 2: Монотонная цепь Маркова

В случае Булеана алгоритм 2 с вероятностью $\frac{n}{2 \cdot n} = \frac{1}{2}$ не изменяет своего состояния, и с равной $\frac{1}{2 \cdot n}$ вероятностью переходит с текущего подмножества на одного из n его соседей, то есть представляет собой ленивое случайное блуждание по соответствующему гиперкубу.

Мы исследовали время перемешивания алгоритма 2 в случае Булеана.

Определение 8. Для заданного порога $\varepsilon > 0$ временем перемешивания называется такое минимальное целое число $T = \tau(\varepsilon)$, что для любого стартового состояния $j \in S$ выполняется

$$\forall t \geq T \left[\|\delta_j \cdot P^t - \pi\|_{TV} \leq \varepsilon \right],$$

где π является стационарным распределением, а $\|\cdot\|_{TV}$ - норма тотальной вариации (=половина l_1 -нормы).

Введем *единичные орты* в \mathbb{Z}_2^n стандартным способом:

$$e_1 = (1, 0, \dots, 0); \dots; e_n = (0, 0, \dots, 1).$$

Ясно, что для алгоритма 2 на Булеане выполнено $(P)_{i,j} = \mu((-i) + j)$, где μ - распределение вероятностей, задаваемое на $S = \mathbb{Z}_2^n$ равенствами

$$\begin{aligned} \mu(0) &= \frac{1}{2}; \\ \mu(e_1) &= \dots = \mu(e_n) = \frac{1}{2^n}. \end{aligned} \tag{2}$$

и $\mu(j) = 0$ всех остальных $j \in \mathbb{Z}_2^n$.

Теорема 11. Пусть μ определяется равенствами (2), а π - равномерное распределение. Для $t \geq \frac{1}{2} \cdot n \cdot (\log n + c)$ имеем

$$(\|\mu^{*t} - \pi\|_{TV})^2 \leq \frac{1}{4} \cdot (e^{e^{-c}} - 1).$$

Из этой теоремы получаем оценку сверху для времени перемешивания:

$$\tau(\varepsilon) \leq \frac{1}{2} \cdot n \cdot (\log n - \log \log(4\varepsilon^2 + 1)) \tag{3}$$

при $0 < \varepsilon < \frac{\sqrt{e-1}}{2}$.

Теорема 12. Пусть μ определяется равенствами (2), а π - равномерное распределение. Для $t = \frac{1}{2} \cdot n \cdot (\log n - 2 \cdot \log(2d))$ и достаточно больших n имеем

$$\|\mu^{*t} - \pi\|_{TV} \geq 1 - \frac{5}{d^2}.$$

Четвертая глава посвящена процедурам машинного обучения и их программной реализации.

Автор развел вероятностно-комбинаторный формальный метод (ВКФ-метод) машинного обучения. ВКФ-метод реализует вероятностную процедуру индуктивного обобщения обучающих примеров

и процедуру предсказания целевого свойства у тестовых примеров по аналогии с обучающими.

Первоначально была предложена процедура абдуктивного уточнения множества гипотез, но по результатам экспериментов на реальных данных от нее было решено отказаться в пользу увеличения числа порождаемых процедурой индукции гипотез.

Наконец, процедура предсказания целевого свойства по аналогии с обучающими примерами соответствует следующему алгоритму:

Data: расширенная выборка S^+ ВКФ-гипотез, файл (τ) -примеров

Result: предсказанные свойства (τ) -примеров

$X := (\tau)$ -примеры;

for ($o \in X$) **do**

$PredictPositively(o) := \text{false};$

for ($\langle A, B \rangle \in S^+$) **do**

if ($B \subseteq o'$) **then**

$PredictPositively(o) := \text{true};$

end

end

end

Algorithm 3: Процедура предсказания по аналогии

Формальное определение и полезная переформулировка таковы:

Определение 9. Объект o , описываемый фрагментом $o' \subseteq F$ (множеством признаков), **предсказывается положительным** с помощью ВКФ-гипотезы $\langle A, B \rangle$, если $B \subseteq o'$.

Определение 10. **Нижнее полупространство** $H_{1/2}^\downarrow(o)$, определяемое объектом o с фрагментом $o' \subseteq F$, задается линейным неравенством $x_{j_1} + \dots + x_{j_k} < \frac{1}{2}$, где $F \setminus o' = \{f_{j_1}, \dots, f_{j_k}\}$.

Лемма 1. Пример o предсказывается положительным тогда и только тогда, когда в любом его нижнем полупространстве содержится хотя одна ВКФ-гипотеза.

Индуктивное обобщение обучающих примеров осуществляется следующей процедурой:

Data: множество обучающих (+)- и (-)-примеров; число N порождаемых ВКФ-гипотез

Result: случайная выборка S ВКФ-гипотез

$O := (+)$ -примеры, $F :=$ признаки; $I \subseteq O \times F$ формальный контекст для (+)-примеров;

$C := (-)$ -примеры; $S := \emptyset$; $i := 0$;

while ($i < N$) **do**

 породить кандидата $\langle A, B \rangle$ с помощью цепи Маркова;

$hasObstacle := \text{false}$;

for ($c \in C$) **do**

if ($B \subseteq c'$) **then**

$hasObstacle := \text{true}$;

end

end

if ($hasObstacle = \text{false}$) **then**

$S := S \cup \{\langle A, B \rangle\}$;

$i := i + 1$;

end

end

Algorithm 4: Процедура индуктивного обобщения

Проверка условия ($B \subseteq c'$) в алгоритме 4 означает, что фрагмент B кандидата $\langle A, B \rangle$ вкладывается в фрагмент (множество признаков) контр-примера c . Любое такое вложение означает, что кандидат нарушает условие «запрета контр-примеров». Если кандидат преодолевает все такие проверки, то он становится ВКФ-гипотезой (о причине наличия целевого свойства).

Для выбора числа N запусков спаривающей цепи Маркова (алгоритма 1) полезно применение следующей теоремы (мы используем объекты, представленные для предсказания):

Зафиксируем $\varepsilon > 0$ - точность предсказания.

Определение 11. Объект o назовем **ε -важным**, если суммарная вероятность появления таких ВКФ-гипотез $\langle A, B \rangle$, что $B \in H_x^{\downarrow}(o)$

будет больше ε .

Теорема 13. Для n признаков и любых $\varepsilon > 0$ и $1 > \delta > 0$ достаточно породить

$$N \geq \frac{2 \cdot (n + 1) - 2 \cdot \log_2 \delta}{\varepsilon}$$

ВКФ-гипотез, чтобы вероятностью $> 1 - \delta$ все ε -важные объекты могли быть предсказаны положительно.

Вероятностные процедуры машинного обучения, описанные выше, были запрограммированы автором в программной системе, получившей название ВКФ-система:

- Программа реализована как консольное приложение с использованием библиотеки разделяемого доступа. Она была создана в среде Visual Studio Code с использованием библиотеки boost.
- Примеры (обучающие, контр- и представленные для предсказания целевого свойства) представляются объектами класса `boost :: dynamic_bitset`. Они сохраняются в контейнерах типа `std :: vector` стандартной библиотеки C++.
- Программа использует классы `std :: random` для датчиков случайных чисел. Это нужно для спаривающей цепи Маркова (алгоритм 1 или его ленивый вариант).
- Для реализации многопоточности используются классы `std :: thread`.
- Программа платформенно независима: она собиралась и запускалась и под Windows, и под Linux.

Укажем на некоторые достоинства ВКФ-системы:

- Так как каждая ВКФ-гипотеза порождается независимым запуском цепи Маркова, то ВКФ-программа использует несколько потоков для вычисления индуктивного обобщения.

- Предсказание свойств по аналогии осуществляется в один поток, так как вычислительная сложность этого шага мала в сравнении с шагом индукции.
- Загрузка ядер процессора замечательно балансируется по вычислительным потокам (превышает 90% на этапе индуктивного обобщения).

Программная ВКФ-система применялась к двум массивам из репозитория данных для проверки алгоритмов машинного обучения.

Первым массивом был SPECT Hearts (данные компьютерной томографии сердца).

- Обучающая выборка содержит 40 (+)- и 40 (-)-примеров.
- Тестовая выборка содержит 172 (+)- и 15 (-)-примеров.
- Каждый пример описывался 22 бинарными атрибутами.
- ВКФ-система добавила отрицания исходных признаков, чтобы отсутствие атрибута могло быть частью причины проявления свойства. Поэтому обучающая выборка - это матрица 40×44 .
- Точность предсказания простейшей ВКФ-системы достигла 86.1% (151 из 172 (+)-примеров и 10 из 15 (-)-примеров).
- Авторы массива SPECT достигли 84.0% точности своей программой CLIP3, которая реализует обучение покрытию средствами целочисленного программирования.

Второй массив Mushrooms - данные из определителя грибов Северной Америки²⁹, оцифрованные в файл agaricus-lepiota.data

- Исходные данные включают описания 8124 грибов, разделенные на две категории (съедобные и ядовитые). Мы случайным образом разделили их на обучающую и тестовую выборки.

²⁹**Lincoff, G.H.** *The Audubon Society Field Guide to North American Mushrooms.* – NY: Knopf, 1981. – 926 pp.

- Обучающая выборка содержит 4032 объекта, из которых 2088 (+)-объектов (съедобные грибы).
- Тестовая выборка содержит 2120 (+)- и 1972 (-)-примеров (ядовитые грибы).
- Каждый пример описывался 22 признаками, описывающими различные характеристики грибов (цвет, форма шляпки, места произрастания, частота встречаемости и т.п.). Эти признаки - номинальные, принимающие одно из нескольких значений.
- ВКФ-система закодировала эти признаки битовыми строками длины 110 бит.
- Точность предсказания ВКФ-системы достигла 100% для 80 ВКФ-гипотез о причинах ядовитости или 150 ВКФ-гипотез о причинах съедобности (без процедуры абдуктивного уточнения).

Заключение содержит перечисление математических результатов диссертационного исследования:

1. Оценка эффективности ленивых вычислений на шаге индукции в теореме 1.
2. Теорема 2 об оценке снизу вероятности возникновения фантомного сходства без учета контр-примеров.
3. Оценка (теорема 3) асимптотической вероятности появления фантомного сходства при наличии контр-примеров.
4. Явный вид производящих функций (теоремы 4 и 5) для вероятности возникновения фантомного сходства при фиксированном и произвольном числе контр-примеров.
5. Теорема 8 о среднем времени склеивания и теорема 9 о сильной концентрации времени склеивания около его среднего для случая Булеана (множества всех подмножеств признаков).

6. Теорема 10 об изменении вероятностей множеств эргодических состояний для спаривающей цепи Маркова, остановленной с верхней границей по r предварительным прогонам.
7. Верхняя оценка (3) (из теоремы 11) времени перемешивания и теорема 12 об асимптотической точности этой оценки для случая Булеана.
8. Теорема 13 о необходимом числе ВКФ-гипотез, чтобы с вероятностью, не ниже заданной, можно было предсказать положительно все ε -важные объекты.

выводы из проведенных исследований

1. При вычислении всех сходств обучающих примеров возможно порождение фантомных сходств, которые вредят корректному предсказанию целевого свойства у объектов, предъявленных для его прогнозирования.
2. Запрет на контр-примеры и ограничение на минимальное число родителей не позволяют полностью избавиться от эффекта переобучения (порождения фантомных сходств).
3. Механизм отбрасывания кандидатов с малым числом родителей может устранять и нужные причины целевого свойства.
4. Эффекты экспоненциального числа сходств в худшем случае и переобучения требуют создания нового метода обучения с использованием вероятностных алгоритмов, опирающихся на технику алгебраической теории решеток.
5. Алгоритм вычисления сходств объектов сводится побитовому умножению, что позволяет эффективно использовать архитектуру современных компьютеров.
6. Среди нескольких вероятностных алгоритмов поиска сходств имеются такие (спаривающие цепи Маркова), которые обеспечивают остановку вычислений с вероятностью единица.

7. Ленивая спаривающая цепь Маркова значительно эффективнее стандартного варианта (и теория находится в хорошем соответствии с практикой).
8. Имеется механизм удаления длинных траекторий спаривающих цепей Маркова с учетом времени работы предварительных запусков этой цепи.
9. Оценка среднего времени работы спаривающей цепи Маркова в частном случае Булеана демонстрирует вычислительную эффективность этого алгоритма.
10. Оценка необходимого числа ВКФ-гипотез для надежного предсказания важных объектов превращает ВКФ-метод в метод статистического машинного обучения.
11. Программная ВКФ-система продемонстрировала хорошую балансировку нагрузки по вычислительным узлам при многопоточной реализации.
12. Применение компьютерной системы к массивам данных продемонстрировало превосходство предложенного подхода над некоторыми другими алгоритмами комбинаторного машинного обучения и возможность его применения к массивам данных большого объема.

список открытых проблем, исследование которых представляется важным

- Получить оценку среднего времени склеивания для спаривающей цепи Маркова в случае произвольного контекста.
- Исследовать вопрос о времени перемешивания для монотонной цепи Маркова в случае произвольного контекста.
- Исследовать асимптотическую вероятность возникновения фантомного сходства, когда число контр-примеров растет, а число признаков сохраняется.

Публикации автора по теме диссертации

В изданиях, удовлетворяющих требованиям ВАК

1. **Виноградов Д.В.** Вероятностное порождения гипотез в ДСМ-методе с помощью простейших цепей Маркова // *Научная и техническая информация*, Сер. 2. – 2012. – № 9. – С. 20–27
2. **Виноградов Д.В.** Вероятностно-комбинаторный подход к автоматическому порождению гипотез // В кн.: *Гуманитарные чтения РГГУ – 2014*. – М.: РГГУ. – 2015. – С. 771-775
3. **Виноградов Д.В.** Вероятность порождения случайного ДСМ-сходства при наличии контр-примеров // *Научная и техническая информация*, Сер. 2. – 2015. – № 3. – С. 1–5
4. **Виноградов Д.В.** Предельная вероятность порождения случайного сходства при наличии контр-примеров // *Научная и техническая информация*, Сер. 2. – 2017. – № 2. – С. 17–19
5. **Виноградов Д.В.** Эффективность ленивых вычислений для поиска сходств в ВКФ-системе // *Научная и техническая информация*, Сер. 2. – 2017. – № 4. – С. 19–23
6. **Виноградов Д.В.** Анализ результатов применения ВКФ-

системы: успехи и открытая проблема // *Научная и техническая информация*, Сер. 2. – 2017. – № 5. – С. 1–4

7. **Виноградов Д.В.** ВКФ-метод интеллектуального анализа данных: обзор результатов и открытых проблем // *Искусственный интеллект и принятие решений*. – 2017. – № 2. – С. 9–16
8. **Виноградов Д.В.** Надежность предсказания по аналогии // *Научная и техническая информация*, Сер. 2. – 2017. – № 7. – С. 11–15
9. **Виноградов Д.В.** Скорость сходимости к пределу вероятности порождения случайного сходства при наличии контр-примеров // *Научная и техническая информация*, Сер. 2. – 2018. – № 2. – С. 21–24
10. **Виноградов Д.В.** Учет предварительных оценок скорости порождения сходств спаривающей цепью Маркова // *Информатика и ее применения*. – 2018. – № 1. – С. 50–55
11. **Виноградов Д.В.** О представлении объектов битовыми строками для ВКФ-метода // *Научная и техническая информация*, Сер. 2. – 2018. – № 5. – С. 1–4
12. **Панкратова Е.С., Виноградов Д.В.** Формальное описание настройки интеллектуальных ДСМ-систем на область клинической и лабораторной диагностики // *Научная и техническая информация*, Сер. 2. – 2011. – № 9. – С. 1–5
13. **Galitsky, B.A., S.O. Kuznetsov, and D.V. Vinogradov.** Applying hybrid reasoning to mine for associative features in biological data // *Journal of Biomedical Information*, Vol. 40, – Issue 3. – 2007. – p. 203–220
14. **Vinogradov, D.V.** A Markov chain approach to random generation of formal concepts // *Proceedings of the Workshop "Formal Concept Analysis Meets Information Retrieval (FCAIR 2013)": CEUR Workshop Proceedings*, Vol. 977. – 2013. – p. 127–133

15. **Vinogradov, D.V.** VKF-method of hypotheses generation // *Communications in Computer and Information Science*, Vol. 436. – 2014. – p. 237–248
16. **Vinogradov, D.V.** Accidental formal concepts in the presence of counterexamples // *Proceedings of International Workshop on Formal Concept Analysis for Knowledge Discovery (FCA4KD 2017): CEUR Workshop Proceedings*, Vol. 1921. – 2017. – p. 104–112
17. **Vinogradov, D.V.** Machine learning based on similarity operation // *Communications in Computer and Information Science*, Vol. 934. – 2018. – p. 46–59

Другие публикации

18. **Виноградов Д.В.** Автоматическое порождение гипотез в ДСМ-методе с помощью цепей Маркова // В кн.: *Труды 13 национальной конференции по искусственному интеллекту КИИ-2012*, Т. 2. – 2012. – С. 121–127
19. **Виноградов Д.В.** Качество вероятностно порожденных ДСМ-гипотез // В кн.: *Материалы 6 всероссийской мультиконференции по проблемам управления (МКПУ-2013)*, Т. 1. – 2013. – С. 14–16
20. **Виноградов Д.В.** ВКФ-метод порождения гипотез: программная реализация // В кн.: *Труды 14 национальной конференции по искусственному интеллекту КИИ-2014*, Т. 2. – 2014. – С. 252–258
21. **Виноградов Д.В.** Сильная концентрация времени работы алгоритма поиска сходств // В кн.: *Материалы 8 всероссийской мультиконференции по проблемам управления (МКПУ-2015)*, Т. 1. – 2015. – С. 42–45

22. **Виноградов Д.В.** О надежном предсказании ВКФ-гипотезами // В кн.: *Материалы 10 всероссийской мультиконференции по проблемам управления (МКПУ-2017)*, Т. 1. – 2017. – С. 48–50