ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ
«ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
«ИНФОРМАТИКА И УПРАВЛЕНИЕ» РОССИЙСКОЙ
АКАДЕМИИ НАУК»

На правах рукописи

УДК 004.855

Виноградов Дмитрий Вячеславович

Вероятностно-комбинаторный формальный метод обучения, основанный на теории решеток

Специальность 05.13.17 —
«Теоретические основы информатики»

Диссертация на соискание учёной степени
доктора физико-математических наук

Научный консультант:
д.т.н., профессор,
заслуженный деятель науки РФ
В.К. Финн

Москва — 2018

Оглавление

Введение

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 классификации [58].

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.

Выборки признаковых описаний являются представлениями ис-

ходных данных, которые возникают в различных предметных областях в процессе сбора однотипной информации, и которые могут быть использованы для решения следующих задач:

- классификация ситуаций, явлений, объектов или процессов;
- выявление существенных и несущественных признаков (снижение размерности);
- исследование структуры данных;
- нахождение эмпирических закономерностей различного вида;
- нахождение выбросов, пропущенных значений и устранение их влияния;
- формирование эталонных описаний.

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. Пирсон [73]), классификации с использованием разделяющих функций (Р. Фишер [65]), теории проверки статистических гипотез (А. Вальд [80]).

В 50-х годах появились первые нейросетевые модели машинного обучения (перцептрон Ф. Розенблата [51]).

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 [1, 5, 7, 33–35, 41, 42, 45, 47].

Большой вклад в развитие теории ИАД внесли советские и российские ученые: М.А. Айзерман, Э.М. Браверман, Л.И. Розоноэр (метод потенциальных функций [1]), В.Н. Вапник, А.Я. Червоненкис (статистическая теория обучения, метод «обобщенный портрет» [7]), Ю.И. Журавлев (алгоритмы вычисления оценок и алгебраическая теория распознавания [32]), Н.Г. Загоруйко (алгоритмы таксономии

[34,35]), Г.С. Лбов (логические методы распознавания и поиска зависимостей [41]), Вл.Д. Мазуров (метод комитетов [42]), В.Л. Матросов (статистическое обоснование алгебраического подхода к распознаванию [43]), К.В. Рудаков (теория алгебраического синтеза корректных алгоритмов [52]).

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 года [56] группа исследователей под руководством проф. В.К. Финна создала и развивает логико-комбинаторный ДСМ-метод автоматического порождения гипотез [31], в котором формализованы 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 [2, 3] удалось поставить систему индуктивной логики Милля [44] на четкие логические основания. Ключевым компонентом этого подхода является бинарная операция сходства [30]. Следует указать, что примерно в это же 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) [67]. 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 [58], 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 осмысления наследия известного американского математика и логика Чарльза Сэндерса Пирса [50].

После выяснения сути указанных когнитивных процедур проф. В.К. Финн создал единую систему, объединяющую все эти процеду-

ры в одно целое. Эта система и получила название ДСМ-метод [57].

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. Кузнецовым [39], М.И. Забежайло и др. были доказаны пессимистические оценки сложности для многих ДСМ-процедур (NP-полнота и $\#P$-полнота).

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 2.4, возможно, приведет к разрешению этого вопроса.

Все полученные вероятностные результаты имеют наглядный алгоритмический смысл и приводят к значительному ускорению вы-

числений (оценка эффективности ленивых вычислений в теореме 1.2, применение остановленной «спаривающей» цепи Маркова из теоремы 3.3) или определению ключевых параметров (достаточное число сходств в теореме 4.1).

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».

Область исследования. По паспорту специальности 05.13.17 — «Теоретические основы информатики» областями исследования являются:

- разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных (п.5)
- моделирование формирования эмпирического знания (п.7)
- разработка методов обеспечения высоконадежной обработки информации (п.11)

Согласно формуле специальности «Теоретические основы информации» к ней относятся, в числе прочего, «... исследования методов преобразования информации в данные и знания; создание и исследование ... методов машинного обучения и обнаружения новых знаний». Таким образом, исследование вероятностной модели машинного обучения, основанного на теории решеток, соответствует данной специальности.

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 ([10])
- 35 European Conference on Information Retrieval, Moscow, 2013 ([76])
- VI Мультиконференция по проблемам управления МКПУ-2013, с. Дивноморское, 2013 ([11])
- XIV Всероссийская конференция по искусственному интеллекту КИИ-2014, Казань, 2014 ([12])
- Conference on Analysis of Images, Social networks, and Texts AIST-2014, Ekaterinburg, 2014 ([77])
- Всероссийская конференция «Гуманитарные чтения РГГУ – 2014», Москва, 2014 ([13])
- VIII Мультиконференция по проблемам управления МКПУ-2015, с. Дивноморское, 2015 ([15])
- International Workshop «Formal Concept Analysis for Knowledge Discovery», Moscow, 2017 ([78])
- X Мультиконференция по проблемам управления МКПУ-2017, с. Дивноморское, 2017 ([21])

- XVI Всероссийская конференция по искусственному интеллекту КИИ-2018, г. Москва, 2018 ([79])

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: [9, 13, 14, 16–20, 22–24, 49, 66, 76–79].

Другие публикации автора по теме: [10–12, 15, 21].

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 названий.

Краткое содержание работы по главам. В главе 1 определяются решетки сходства и напоминаются основные факты анализа формальных понятий (АФП) и вводятся ключевые операции «замыкай-по-одному», для которых оценивается алгоритмическая эффективность «ленивой» схемы их вычисления (теорема 1.1). Используя технику АФП, удалось сформулировать и доказать (теорема 1.2) корректность алгоритма 1 кодирования объектов битовыми строками, при котором операция сходства заменяется побитовым умножением, что позволяет эффективно использовать архитектуру современных ЭВМ.

В главе 2 изложены результаты автора о «переобучении» при индуктивном обобщении обучающих примеров - возникновении «фантомных» сходств. Для устранения таких сходств имеется несколько механизмов. Наиболее важные среди них - ограничение на число родителей и запрет на контр-примеры. Доказаны теоремы 2.1 и 2.2 о том, что ни один из этих механизмов не могут полностью устранить феномен «переобучения». Параграф 2.3 содержит вывод явных формул производящих функций для вероятности возникновения «фантомных» сходств при наличии фиксированного и произвольного числа контр-примеров (теоремы 2.3 и 2.4).

В главе 3 предложены и исследованы вероятностные алгоритмы нахождения кандидатов в гипотезы о причинах появления целевого свойства. Сначала описываются несколько алгоритмов из работы автора [9], используемых для вероятностного порождения сходств. Для одного из них (спаривающая цепь Маркова в алгоритме 4) имеется естественный момент остановки, который является конечным с вероятностью единица. Для этой цепи Маркова удалось доказать теорему 3.3 об изменении вероятностей эргодических состояний, если мы отбросим траектории, длина которых превосходит сумму длин тра-

екторий во время заданного числа предварительных прогонов. Параграф 3.3 содержит теоремы о времени склеивания спаривающей цепи Маркова для случая Булевой алгебры. Завершается эта глава выводом верхней оценки (3.15) времени перемешивания монотонной цепи Маркова и доказательством теоремы 3.9 об асимптотической точности этой оценки (снова лишь для случая Булевой алгебры).

Глава 4 посвящена процедурам машинного обучения, основанного на методах теории решеток, для порождения причинно-следственных зависимостей. Здесь дано их формальное описание и приведены доказательства их свойств. Для доопределения по аналогии установлен ключевой результат о надежности (теорема 4.1). Описание программной реализации ВКФ-метода содержится в параграфе 4.3. Параграф 4.4 описывает апробацию разработанного подхода на массивах SPECT Hearts и Mushrooms из репозитория данных для тестирования алгоритмов машинного обучения.

Благодарности. Автор признателен своему учителю д.т.н. профессору Финну Виктору Константиновичу за советы и поддержку, д.ф.-м.н. профессору Бениаминову Евгению Михайловичу за советы и полезные идеи, д.ф.-м.н. профессору Шабату Георгию Борисовичу и д.ф.-м.н. профессору Павловскому Владимиру Евгеньевичу за полезные обсуждения, д.филол.н. профессору Гиляревскому Руджеро Сергеевичу за поддержку и информационную помощь.

Автор посвящает предложенный им метод (ВКФ-метод) своему учителю Виктору Константиновичу Финну.

Автор благодарит своих коллег д.ф.-м.н. О.М. Аншакова, к.х.н. В.Г. Блинову, Т.А. Волкову, к.ф.-м.н. С.М. Гусакову, к.т.н. Д.А. Добрынина, к.ф.-м.н. Е.А. Ефимову, д.ф.-м.н. М.И. Забежайло, д.ф.-м.н. проф. С.О. Кузнецова, д.т.н. М.А. Михеенкову, к.т.н. Е.С. Панкратову, к.ф.-м.н. Д.П. Скворцова, к.т.н. Е.Ф. Фабрикантову, к.т.н. Л.О. Шашкина за поддержку и идеи, которыми они щедро делились с автором.

Глава 1

Прикладная теория решеток

В первом параграфе этой главы мы обсудим свойства бинарной операции сходства, определяющей нижнюю полурешетку с наименьшим элементом. Классический метод превращения конечной нижней полурешетки (с добавлением наибольшего элемента, если его нет) в решетку ставит задачу о нахождении супремума. Анализ формальных понятий (АФП) [67] позволяет предложить эффективный алгоритм.

В параграфе 1.2 вводим понятия операций «замыкай-по-одному» и устанавливаем их базисные свойства (корректность и монотонность). С использованием идей АФП предлагается «ленивая» схема вычисления операций «замыкай-по-одному». Теорема 1.1 оценивает алгоритмическую эффективность этого подхода.

Наконец, параграф 1.3 описывает алгоритм 1 кодирования сложных структур признаков битовыми строками с операцией побитового умножения в качестве операции сходства. То, что такое представление всегда возможно составляет содержание фундаментальной теоремы АФП. Мы докажем корректность этого алгоритма (теорема 1.2), опираясь на результаты анализа формальных понятий [62].

1.1 Основные определения

Сходство является бинарной операцией на множестве X, объемлющем множество объектов, то есть представляет собой отображение $\cap : X \times X \to X$. Элементы множества X мы будем называть *фрагментами*. Терминология происходит из фармакологических исследований, где изучаются причины того или другого биологического действия химических соединений. Такие причины (фармакофоры) ищутся среди общих частей некоторой группы биологическиактивных соединений путем нахождения их общего фрагмента (возможно, несвязного). При вычислении сходства объектов, перечисленным в некотором порядке, применяется операция сходства. Промежуточные результаты - фрагменты - тоже могут выступать аргументами операции сходства.

Для независимости результата нахождения сходства нескольких объектов от порядка вычисления операция сходства должна удовлетворять аксиомам *нижней полурешетки*:

$$x \cap x = x \tag{1.1}$$

$$x \cap y = y \cap x \tag{1.2}$$

$$(x \cap y) \cap z = x \cap (y \cap x) \tag{1.3}$$

Для выражения тривиальности сходства добавляется специальный *пустой фрагмент* $\emptyset$ со свойством наименьшего элемента:

$$x \cap \emptyset = \emptyset \tag{1.4}$$

Важнейшим примером для нас будет нижняя полурешетка, состоящая из битовых строк фиксированной длины с побитовым умножением в качестве операции сходства. Пустым фрагментом будет являться строка, состоящая из одних нулей. Каждый бит может быть отождествлен с бинарным признаком. Тогда битовая строка соответствует множеству признаков, в которых встречаются единицы. При этом операция сходства соответствует пересечению множеств признаков, а пустой фрагмент - пустому множеству признаков.

Ясно, что в этом примере строка из одних единиц будет соответствовать наибольшему элементу F со свойством:

$$x \cap F = x \tag{1.5}$$

Легко доказать известный результат [62] о том, что **любую конечную нижнюю полурешетку с наибольшим элементом можно превратить в решетку**.

Операция $x \cup y$ задается как последовательное сходство (в произвольном порядке) множества $\{z_1, \ldots, z_k\}$ всех *общих верхних граней* для x и y - элементов полурешетки со свойствами $z_j \cap x = x$ и $z_j \cap y = y$. Сходством одноэлементного множества является фрагмент того элемента, который в нем содержится. Сходством пустого множества является наибольший элемент (существующий по условию).

Тогда можно проверить свойства решетки:

$$x \cup x = x \tag{1.1'}$$

$$x \cup y = y \cup x \tag{1.2'}$$

$$(x \cup y) \cup z = x \cup (y \cup x) \tag{1.3'}$$

$$x \cup F = F \tag{1.4'}$$

$$x \cup \emptyset = x \tag{1.5'}$$

$$x \cup (x \cap y) = x \qquad\qquad x \cap (x \cup y) = x \tag{1.6}$$

Важность этого примера объясняется двумя фактами:

1. Операция побитового умножения допускает эффективную реализацию на современных ЭВМ. Существуют специальные классы объектов (например, *boost* :: *dynamic_bitset* в C++), реализующие удобное оперирование битовыми строками.

2. Анализ формальных понятий позволяет эффективно определить полурешетку битовых строк с операцией побитового умножения, изоморфную заданной нижней полурешетке.

К сожалению, вариант вложения нижней полурешетки (с добавлением наибольшего элемента, если его первоначально не было), описанный в этом параграфе, может рассматриваться неудовлетворительным по двум причинам:

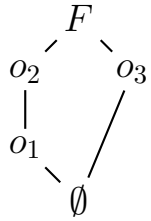
1. Обычно операция супремума $\cup$ не совпадает с побитовой дизъюнкцией, которая тоже эффективно вычисляется на современных ЭВМ;

2. Вычисление же супремума как сходства всех общих верхних граней не является эффективным (может потребовать побитово перемножить почти все объекты).

Для обоснования первого утверждения достаточно рассмотреть нижнюю полурешетку битовых строк с операцией побитового умножения:

$O \mid F$	f_1	f_2	f_3
o_1	1	0	0
o_2	1	1	0
o_3	0	0	1
$\emptyset$	0	0	0

Легко проверить, что добавление максимального элемента F породит решетку-«пентагон» N_5



Эта решетка не является *дистрибутивной*, то есть не удовлетворяет условиям:

$$x \cup (y \cap z) = (x \cup y) \cap (x \cup z) \quad x \cap (y \cup z) = (x \cap y) \cup (x \cap z) \quad (1.7)$$

Первое равенство опровергается при $x = o_1, y = o_2, z = o_3$, так как тогда $o_1 \cup (o_2 \cap o_3) = o_1 \cup \emptyset = o_1$, но $(o_1 \cup o_2) \cap (o_1 \cup o_3) = o_2 \cap F = o_2$.

То, что решетка $\langle \{0,1\}^n, \cap, \cup, \emptyset = 0^n, 1^n \rangle$ (т.е. множество всех строк с побитовыми операциями конъюнкции и дизъюнкции) является дистрибутивой, следует из того, что на каждой компоненте мы имеем дистрибутивную решетку $\langle \{0,1\}, \wedge, \vee, 0, 1 \rangle$, для которых образуется их Декартово произведение.

Теперь напомним определение *гомоморфизма* $h : \langle L_1, \cap, \cup, 0, 1 \rangle \to \langle L_2, \cap, \cup, 0, 1 \rangle$ решеток как такого отображения $h : L_1 \to L_2$, что $h(0) = 0, h(1) = 1$, и для всех $x, y \in L_1$ выполняются равенства:

$$h(x \cup y) = h(x) \cup h(y) \qquad h(x \cap y) = h(x) \cap h(y) \tag{1.8}$$

Если гомоморфизм является инъективным отображением $h : L_1 \to L_2$, то называется *мономорфизмом*, а решетка $\langle L_1, \cap, \cup, 0, 1 \rangle$ (точнее, ее изоморфный образ) является *подрешеткой* решетки $\langle L_2, \cap, \cup, 0, 1 \rangle$.

Из равенств (1.8) легко выводится, что из дистрибутивности решетки $\langle L_2, \cap, \cup, 0, 1 \rangle$ следует дистрибутивность любой ее подрешетки $\langle L_1, \cap, \cup, 0, 1 \rangle$.

Поэтому для вышеприведенного формального контекста, порождающего недистрибутивную решетку N_5, невозможен мономорфизм ни в какую Булеву алгебру вида $\langle \{0,1\}^n, \cap, \cup, \emptyset, 1^n \rangle$, то есть невозможно реализовать операцию супремума $\cup$ побитовой дизъюнкцией.

Однако анализ формальных понятий, к изложению которого мы переходим, позволяет обойти второе препятствие.

Собирая вместе битовые строки, представляющие объекты, мы получаем прямоугольную таблицу I, которую мы будем называть *формальным контекстом* [67]. Формальный контекст можно понимать как бинарное отношение между элементами множества O, которые мы называем *именами объектов* (или даже объектами), 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 его *сходством* называется подмножество $A' = \{f \in F : \forall o \in A\,[oIf]\} \subseteq F$. Полагаем $\emptyset' = F$.

На самом деле, это определение совпадает с последовательным вычислением побитового умножения строк, соответствующих отобранным во множество A объектов, как это определялось в предыдущем параграфе.

Для подмножества $B \subseteq F$ признаков его *сходством* называется подмножество $B' = \{o \in O : \forall f \in B\,[oIf]\} \subseteq O$. Полагаем $\emptyset' = O$.

Понятия сходства, определенные выше, задают операции $' : 2^O \to 2^F$ и $' : 2^F \to 2^O$, называемые *полярами*.

Сформулируем простую лемму, прямо выводящуюся из определения, которая будет широко применяться в последующем изложении:

Лемма 1.1. *Для $A_1 \subseteq O$ и $A_2 \subseteq O$ выполняется $(A_1 \cup A_2)' = A_1' \cap A_2'$. Для $B_1 \subseteq F$ и $B_2 \subseteq F$ выполняется $(B_1 \cup B_2)' = B_1' \cap B_2'$.*

Особенно часто мы будем использовать такие варианты:

$$(A \cup \{o\})' = A' \cap \{o\}' \tag{1.9}$$

для любых $A \subseteq O$ и $o \in O$, и

$$(B \cup \{f\})' = B' \cap \{f\}' \tag{1.10}$$

для любых $B \subseteq F$ и $f \in F$.

Легко проверить следующие свойства соответствий Галуа для сходства [67]:

$$\forall A\,[A \subseteq A''] \qquad \forall B\,[B \subseteq B''] \tag{1.11}$$

$$\forall A_1 \forall A_2\,[A_1 \subseteq A_2 \Rightarrow A_1' \supseteq A_2'] \quad \forall B_1 \forall B_2\,[B_1 \subseteq B_2 \Rightarrow B_1' \supseteq B_2'] \tag{1.12}$$

$$\forall A\,[A' = A'''] \qquad \forall B\,[B' = B'''] \tag{1.13}$$

Определение 1.1. *Пару $\langle A, B\rangle$ назовем **кандидатом**, если $A = B' \subseteq O$ и $B = A' \subseteq F$.*

В АФП такие пары называют *формальными понятиями*, но мы предпочитаем сменить название, так как термин «понятие» имеет другой смысл для специалистов в области машинного обучения.

Наглядно, кандидаты в формальном контексте соответствуют максимальным подматрицам, заполненным единицами:

$O \times F$	f_1	$\dots$	f_{j_1}	f_{j_1+1}	$\dots$	$f_{j_{m-1}}$	$\dots$	f_{j_m}	$\dots$	f_n
o_1	0	$\dots$	0	0	$\dots$	1	$\dots$	0	$\dots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
o_{i_1}	0	$\dots$	**1**	1	$\dots$	**1**	$\dots$	**1**	$\dots$	1
o_{i_1+1}	0	$\dots$	0	0	$\dots$	1	$\dots$	1	$\dots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$o_{i_{l-1}}$	1	$\dots$	**1**	0	$\dots$	**1**	$\dots$	**1**	$\dots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
o_{i_l}	1	$\dots$	**1**	1	$\dots$	**1**	$\dots$	**1**	$\dots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
o_k	0	$\dots$	1	0	$\dots$	0	$\dots$	0	$\dots$	0

Здесь подмножество объектов (строк) $A = \{o_{i_1}, \dots, o_{i_{l-1}}, o_{i_l}\}$ называется *списком родителей*, а подмножество признаков (столбцов) $B = \{f_{j_1}, \dots, f_{j_{m-1}}, f_{j_m}\}$ называется *фрагментом*. Максимальность означает, что нельзя добавить ни одну строку, ни одного столбца так, чтобы расширенная подматрица состояла лишь из одних единиц.

Равенство $A = B'$ из определения 1.1 означает невозможность добавить еше одну строку и называется «принципом исчерпываемости родителей». Равенство $B = A'$ соответствует невозможности добавить еще один столбец и говорит, что фрагмент B - общая часть всех примеров-родителей из A.

Одна из главных проблем предыдущих подходов к анализу данных, основанных на операции сходства, при которых предварительно вычислялись всевозможные сходства подмножеств обучающих примеров - высокая вычислительная сложность. На практике этому соответствует «комбинаторный взрыв». В теории имеем экспоненциальную оценку на число сходств в худшем случае.

Такой худший случай (Булеан) возникает, когда из n обучающих примеров с n признаками можно получить 2^n различных сходств.

Рассмотрим формальный контекст для Булеана:

Пусть $O = \{o_1, o_2, \ldots, o_n\}$ будет множеством объектов, каждый из которых признаками из списка $F = \{f_1, f_2, \ldots, f_n\}$, и

$$o_i I f_j \Leftrightarrow i \neq j. \tag{1.14}$$

$O \mid F$	f_1	f_2	$\ldots$	f_n
o_1	0	1	$\ldots$	1
o_2	1	0	$\ldots$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_n	1	1	$\ldots$	0

Ясно, что $\bigcap \{o_{j_1}, \ldots, o_{j_l}\} = F \setminus \{f_{j_1}, \ldots, f_{j_l}\}$, так как добавление в сходство примера o_k с номером k удаляет из фрагмента признак f_k с тем же самым номером k.

Очевидно, что таким образом может быть получено любое подмножество n-элементного множества F.

Нужно сказать, что такой «комбинаторный взрыв», хотя и оказывает влияние на скорость анализа данных, тем не менее гораздо безобиднее фантомных сходств (определяемых в главе 2): если есть некоторая «настоящая» причина, и порождается также много ее надмножеств, то, с точки зрения предсказания по аналогии (составляющий алгоритм 6), любое надмножество будет правильно доопределять объекты, так как при его включении «настоящая» причина тоже будет содержаться в предсказываемых объектах, и, значит, будет вынуждать целевое свойство. При этом все такие надмножества, когда они сработают (породят положительный прогноз), дадут одинаковый результат.

Определение 1.2. ***Порядок*** *на кандидатах:* $\langle A_1, B_1 \rangle \leq \langle A_2, B_2 \rangle$, *если* $B_1 \subseteq B_2$.

Это двойственное (с точки зрения анализа формальных понятий) определение приводится в настоящем виде для согласованности с традицией отечественной школы.

Легко проверить (доказательство есть, например, в [62]), что **множество всех кандидатов** (для фиксированного формального контекста $I \subseteq O \times F$) **образует решетку $\boldsymbol{L}$ относительно операций**

$$\langle A_1, B_1\rangle \cap \langle A_2, B_2\rangle = \langle (A_1 \cup A_2)'', B_1 \cap B_2\rangle \tag{1.15}$$

$$\langle A_1, B_1\rangle \cup \langle A_2, B_2\rangle = \langle A_1 \cap A_2, (B_1 \cup B_2)''\rangle \tag{1.16}$$

1.2 Операции «Замыкай-по-одному»

Для понимания смысла главных рабочих операций ВКФ-метода - операций «замыкай-по-одному» - полезны отображения $g : O \to L$ и $h : F \to L$ объектов и признаков, соответственно, в решетку $\boldsymbol{L}$ всех кандидатов

$$g(o) = \langle \{o\}'', \{o\}'\rangle \tag{1.17}$$

$$h(f) = \langle \{f\}', \{f\}''\rangle \tag{1.18}$$

Теперь операции «замыкай-по-одному» можно определить как:

Определение 1.3. *Операция* ***замыкай-по-одному-вниз*** *на кандидате $\langle A, B\rangle$ и объекте $o \in O$ порождает пару*

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

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

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

Операция $CbODown$ соответствует шагу алгоритма «Замыкай-по-одному», который был предложен С.О.Кузнецовым [40] для вычисления всех кандидатов перебором сверху-вниз. Операции $CbOUp$ и $CbODown$ были предложены автором безымянно [9] и под этим именем в [76] для процедур случайного блуждания по решетке всех кандидатов, варианты которого описаны в главе 3.

Легко проверяется с использованием формул (1.15), (1.16), (1.17), (1.18)что

$$CbODown(\langle A, B\rangle, o) = \langle A, B\rangle \cap g(o) \tag{1.19}$$

$$CbOUp(\langle A, B\rangle, f) = \langle A, B\rangle \cup h(f) \tag{1.20}$$

Теперь легко установить

Лемма 1.2. *Для любого кандидата $\langle A, B\rangle$ и любого объекта $o \in O$ пара $CbODown(\langle A, B\rangle, o)$ является кандидатом.*

Аналогично, для любого кандидата $\langle A, B\rangle$ и любого признака $f \in F$ пара $CbOUp(\langle A, B\rangle, f)$ является кандидатом.

Лемма 1.3. *Для всякой упорядоченной пары кандидатов*

$$\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$$

и любого $o \in O$ имеем

$$CbODown(\langle A_1, B_1\rangle, o) \leq CbODown(\langle A_2, B_2\rangle, o).$$

Для всякой упорядоченной пары кандидатов

$$\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$$

и любого $f \in F$ имеем

$$CbOUp(\langle A_1, B_1\rangle, f) \leq CbOUp(\langle A_2, B_2\rangle, f).$$

Это утверждение легко проверяется с использованием формул (1.12) и определения 1.2. □

Сформулированные выше леммы 1.2 и 1.3 будут использованы нами при описании вероятностных алгоритмов поиска сходств в параграфе 3.2.

Для дальнейшего нам будет полезен явный вид операций «замыкай-по-одному» в решетке-Булеане $o_i I f_j \Leftrightarrow i \neq j$ для множества объектов $O = \{o_1, o_2, \ldots, o_n\}$, каждый из которых описывается признаками из списка $F = \{f_1, f_2, \ldots, f_n\}$.

Ясно, что в этом случае

$$CbODown(\langle A, B\rangle, o_k) = \begin{cases} \langle A \cup \{o_k\}, B \setminus \{f_k\}\rangle, & \text{если } o_k \notin A \\ \langle A, B\rangle, & \text{иначе,} \end{cases} \tag{1.21}$$

так как добавление в сходство объекта o_k с номером k удаляет из фрагмента признак f_k с тем же самым номером k.

Аналогично,

$$CbODown(\langle A, B\rangle, f_k) = \begin{cases} \langle A \setminus \{o_k\}, B \cup \{f_k\}\rangle, & \text{если } f_k \notin B \\ \langle A, B\rangle, & \text{иначе,} \end{cases} \tag{1.22}$$

так как добавление в сходство признака f_k с номером k удаляет из родителей сходства объект o_k с тем же самым номером k.

Эти выражения помогут нам понять, что в этом частном случае вероятностные алгоритмы 2 и 3 поиска сходств, описанные в параграфе 3.1, совпадают с классическими случайными блужданиями на гиперкубе всех подмножеств.

В базисном алгоритме 4 операции $CbODown$ и $CbOUp$ применяются в зависимости от того, выпадает объект или признак. Используя возможность появления длинных серий из одних объектов или из одних признаков, можно добиться существенного ускорения, если реализовывать их с использованием ленивых вычислений. Вопросу о вычислительной эффективности этого подхода посвящена основная теорема настоящего параграфа.

Согласно определению 1.3

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

Если вычисление пересечения $B \cap \{o\}'$ фрагмента текущего кандидата с фрагментом выбранного объекта o соответствует побитовому умножению соответствующих строк, то операция $(A \cup \{o\})'' = (B \cap \{o\}')'$ (равенство следует из формулы (1.10)) формирования нового списка родителей может потребовать побитово перемножить с

полученным ранее пересечением почти все объекты, чтобы проверить, обладает ли еще какой-нибудь объект полученным пересечением.

Для улучшения ситуации предлагается (лениво) откладывать вычисления замыкания (двух последовательных поляр), пока последовательный выбор нескольких объектов для $CbODown$ не сменится выбором признака с переходом к операции $CbOUp$.

Аналогично, операция $CbOUp$ имеет в своем составе потребляющую много времени компоненту $(B \cup \{f\})'' = (A \cap \{f\}')'$ (равенство следует из формулы (1.11)). Здесь тоже можно лениво откладывать вычисления этой части до тех пор, пока выбор нескольких признаков для $CbOUp$ не сменится выбором объекта с переходом к операции $CbODown$.

Возникает вопрос о степени экономии, достигаемой такой процедурой. Впервые эта задача была решена в работе автора [17]. Здесь мы проведем более подробный анализ и используем более понятные обозначения.

Рассмотрим последовательность типов (объект или признак) элементов, выбираемых в ходе работы алгоритма 4. Ясно, что это - последовательность (вообще говоря, бесконечная) испытаний Бернулли $\langle \sigma_1, \ldots, \sigma_j, \ldots \rangle$ с вероятностью успеха (например, выбора признака), равной $p = \frac{n}{n+k}$, где n - число признаков, а k - число обучающих примеров.

Прежде всего зафиксируем два события $\{\sigma_1 = 0\}$ и $\{\sigma_1 = 1\}$.

В случае $\sigma_1 = 0$ нас интересует длина события $\{\sigma_1 = \ldots = \sigma_i = 0, \sigma_{i+1} = \ldots = \sigma_j = 1, \sigma_{j+1} = 0\}$, а в случае $\sigma_1 = 1$ интересна длина $\{\sigma_1 = \ldots = \sigma_i = 1, \sigma_{i+1} = \ldots = \sigma_j = 0, \sigma_{j+1} = 1\}$.

Рассмотрим случайную величину T суммы длин двух переходов от объектов к признакам и снова к объектам (при $\sigma = 0$) и суммы длин двух переходов от признаков к объектам (при $\sigma = 1$)

Воспользуемся формулой полной вероятности (для средних), чтобы получить

$$\boldsymbol{E}[T] = \boldsymbol{E}[T|\sigma_1 = 0] \cdot \boldsymbol{P}[\sigma_1 = 0] + \boldsymbol{E}[T|\sigma_1 = 1] \cdot \boldsymbol{P}[\sigma_1 = 1] \qquad (1.23)$$

Легко видеть, что

$$\boldsymbol{P}[T=j|\sigma_1=0]=\frac{1}{1-p}\cdot\sum_{i=1}^{j-1}(1-p)^{i+1}\cdot p^{j-i}=$$

$$=p^{j-1}\cdot(1-p)\cdot\sum_{r=1}^{j-2}\left(\frac{1-p}{p}\right)^r=p^{j-1}\cdot(1-p)\cdot\frac{\left(\frac{1-p}{p}\right)^{j-1}-1}{\frac{1-p}{p}-1}=$$

$$=p\cdot(1-p)\cdot\frac{(1-p)^{j-1}-p^{j-1}}{1-2p}. \quad (1.24)$$

$$\boldsymbol{P}[T=j|\sigma_1=1]=p\cdot(1-p)\cdot\frac{(1-p)^{j-1}-p^{j-1}}{1-2p}. \quad (1.25)$$

Теперь применим технику производящих функций [55], т.е. рассмотрим функции (для $\sigma=0$ и $\sigma=1$)

$$\psi_\sigma(z)=\sum_{j=2}^{\infty}\boldsymbol{P}(T=j|\sigma_1=\sigma)\cdot z^j.$$

Суммирование геометрической прогрессии доказывает

Лемма 1.4.

$$\psi_0(z)=\psi_1(z)=\frac{p\cdot(1-p)}{1-2p}\cdot z\cdot\Big[(1-(1-p)z)^{-1}-(1-pz)^{-1}\Big].$$

Среднее значение вычисляется как значение первой производной от $\psi_\sigma(z)$ в единице: Поэтому

$$\boldsymbol{E}[T=j|\sigma_1=0]=\psi_0'(1)=\frac{p\cdot(1-p)}{1-2p}\cdot$$

$$\cdot\left(\left[\frac{1}{1-(1-p)\cdot1}-\frac{1}{1-p\cdot1}\right]+1\cdot\left[\frac{1-p}{(1-(1-p)\cdot1)^2}-\frac{p}{(1-p\cdot1)^2}\right]\right)=$$

$$=\frac{p\cdot(1-p)}{1-2p}\cdot\left(\frac{1}{p^2}-\frac{1}{(1-p)^2}\right)=\frac{1}{1-2p}\cdot\left(\frac{1-p}{p}-\frac{p}{1-p}\right)=$$

$$=\frac{(1-p)^2-p^2}{p\cdot(1-p)\cdot(1-2p)}=\frac{1}{p\cdot(1-p)}. \quad (1.26)$$

$$\boldsymbol{E}[T = j|\sigma_1 = 0] = \psi_1'(1) = \frac{1}{p \cdot (1-p)}. \tag{1.27}$$

Теорема 1.1. *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* $\frac{(n+k)^2}{k \cdot n}$ *операций замыкания.*

Доказательство. По формулам (1.23), (1.26) и (1.27) имеем $\boldsymbol{E}[T] = \frac{1}{p \cdot (1-p)} \cdot (1-p) + \frac{1}{p \cdot (1-p)} \cdot p = \frac{1}{p \cdot (1-p)}$ Так как $p = \frac{k}{n+k}$ и $1 - p = \frac{n}{n+k}$, то выигрыш от введения ленивых вычислений в среднем составляет

$$\frac{1}{p \cdot (1-p)} = \frac{(n+k)^2}{k \cdot n} \tag{1.28}$$

раз. □

Ясно, что выигрыш тем больше, чем больше разница между k и n, где k - число обучающих примеров, а n - число признаков, используемых для описания объектов, так как $\frac{(n+k)^2}{k \cdot n} = 4 + \frac{(n-k)^2}{k \cdot n}$. Даже в худшем случае $k = n$ это сокращение вызовов трудоемкой операции не меньше двух раз, потому что $\frac{(2k)^2}{k \cdot k} = 4$. В работе [17] автора приводится формула $\frac{k}{n} + \frac{n}{k}$, которая получается из выведенной в теореме 1.1 вычитанием числа двух обязательных операций замыкания, так как $\frac{(n+k)^2}{k \cdot n} - 2 = \frac{k}{n} + \frac{n}{k}$.

1.3 Кодирование битовыми строками

Фундаментальная теорема анализа формальных понятий [62, 67] гласит: **Любая конечная решетка изоморфна решетке всех кандидатов для подходяще выбранного формального контекста.**

Сначала мы напомним некоторые понятия и воспроизведем некоторые результаты из анализа формальных понятий, так как это позволит нам использовать их для представления объектов, описываемых признаками со сложной структурой значений, с помощью битовых строк (породить формальный контекст).

Прежде всего мы назовем подмножество элементов $S \subseteq L$ решетки $\boldsymbol{L}$ $\cap$-*плотным*, если для любого элемента $x \in L$ найдется такое подмножество $X \subseteq S$, что $x = \cap X$. Подмножество элементов $S \subseteq L$ решетки $\boldsymbol{L}$ называется $\cup$-*плотным*, если для любого элемента $x \in L$ найдется такое подмножество $X \subseteq S$, что $x = \cup X$.

Очевидно, что все L является как $\cap$-, так и $\cup$-плотным. Более полезный пример доставляется следующим утверждением:

Лемма 1.5. *Для решетки кандидатов $\boldsymbol{L}$, порождаемой формальным контекстом $I \subseteq O \times F$, образ $g(O) = \{g(o) : o \in O\}$ отображения $g : O \to L$, задаваемого правилом $g(o) = \langle \{o\}'', \{o\}' \rangle$, является $\cap$-плотным подмножеством. Образ $h(F) = \{h(f) : f \in F\}$ отображения $h : F \to L$, задаваемого правилом $h(f) = \langle \{f\}', \{f\}'' \rangle$, является $\cup$-плотным подмножеством.*

Легко установить и обратный результат (доказательство имеется, например, в [62]):

Лемма 1.6. *Пусть для любой конечной решетки $\boldsymbol{L}$ найдутся такие два множества O и F с отображениями $g : O \to L$ и $h : F \to L$, что $g(O) \subseteq L$ - $\cap$-плотное подмножество, а $h(F) \subseteq L$ - $\cup$-плотное подмножество. Тогда, полагая $oIf \Leftrightarrow g(o) \geq h(f)$, мы получим формальный контекст, решетка кандидатов которого будет изоморфна исходной решетке $\boldsymbol{L}$.*

Ясно, что тождественные отображения $g = id : L \to L$ и $h = id : L \to L$ удовлетворяют условию леммы 1.6. Именно так и проходило первоначальное доказательство фундаментальной теоремы АФП.

К сожалению, этот вариант обычно порождает слишком большой формальный контекст. Множество объектов обычно бывает задано извне. Чтобы выбрать минимальное подмножество $F \subseteq L$, нам необходимо ввести понятие $\cup$-неразложимых элементов.

Определение 1.4. *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 $y \cup z < x$.*

Простые вычисления доказывают следующий результат:

Лемма 1.7. *Для любой конечной решетки* $\boldsymbol{L}$ *любое надмножество всех* $\cup$*-неразложимых элементов образуем* $\cup$*-плотное подмножество.*

Если предполагать, что объекты описываются признаками, имеющими на их значениях структуру нижней полурешетки (с добавленным наименьшим элементом, интерпретируемым как отсутствие сходства по этому признаку), а операция сходства вычисляется покомпонентно, то достаточно рассмотреть каждый признак отдельно. Кодирование же целого объекта составляет конкатенацию (соединение) кодирований значений каждого признака.

Сосредоточимся на кодировании множества V значений одного признака. Если нижняя полурешетка V не имеет максимального элемента, добавим новым элемент, объявляя его максимальным. Эта процедура уже встречалась нам ранее. Очевидно, что при таком добавлении этот максимальный элемент будет $\cup$-разложимым в расширенной решетке $\boldsymbol{L}$. Поэтому мы можем игнорировать его и рассматривать контекст $\geq \subseteq V \times V$.

Требование, чтобы значения каждого признака образовывали нижнюю полурешетку, необходимо для того, чтобы не возникало ситуации, когда сходством нескольких объектов в этом признаке порождается новое значение, которое не имеет имени.

Так как операция сходства порождает порядок ($x \leq y \equiv x \cap y = x$), то отношение накрытия ($x \prec y \equiv x < y \& \neg\exists z[x < z < y]$) задает ациклический ориентированный граф.

Будем считать заданными только перечисление вершин графа (с их именами, используемыми для наглядного представления порождаемых сходств) и отношение накрытия на них. Поэтому предварительно из этих данных восстанавливается порядок $\geq \subseteq V \times V$ (как транзитивное замыкание отношения накрытия). В алгоритме 1 это делается с помощью предварительной топологической сортировки множества вершин:

Определение 1.5. *Линейный порядок* $V[0] < V[1] < \ldots < V[n-1]$ *назовем топологической сортировкой, если* $\forall i, j[V[i] \prec V[j] \Rightarrow i < j]$.

Основной идеей описанного ниже алгоритма кодирования является процедура сокращения формального контекста $\geq \subseteq V \times V$ до множества F всех $\cup$-неразложимых элементов.

Data: множество $V = [0, 1, \ldots, n-1]$ значений текущего признака
Result: матрица B такая, что $B[j]$ - битовая строка для кодирования значения j

```
V := topological_sort(V); // топологическая сортировка
∀i∀j[T[i][j] := false]; // матрица порядка
for (index = 0; index < n; ++index) do
    T[V[index]][V[index]] := true;
    for (indx = 0; indx < index; ++indx) do
        if (V[indx] ≺ V[index]) then
            for (ndx = 0; ndx < n; ++ndx) do
                T[V[index]][ndx] |= T[V[indx]][ndx];
            end
        end
    end
end
∀i[Del[i] = false]; // удаляемые столбцы
for (index = 2; index < n; ++index) do
    for (indx = 1; indx < index; ++indx) do
        for (ndx = 0; ndx < indx; ++ndx) do
            if (T[ ][V[index]] == T[ ][V[indx]]&T[ ][V[ndx]])
             then
                Del[V[index]] := true;
            end
        end
    end
end
¬Del[indx] ⇒ B[indx][index] := T[index][indx];
```

Algorithm 1: Кодирование битовыми строками

Предлагаемый алгоритм кодирования состоит из четырех частей. Сначала осуществляется топологическая сортировка. Известно,

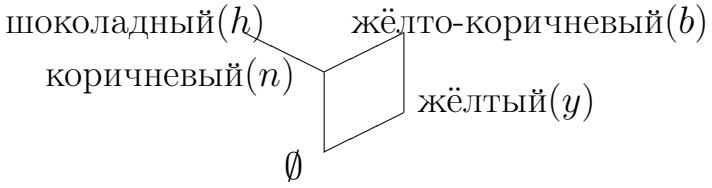
что топологическую сортировку можно выполнить за $O(|V| + |V|^2)$ шагов (например, как описано в [38]).

Во второй части строится матрица T порядка как транзитивное и рефлексивное замыкание отношения накрытия. Временная сложность этой части алгоритма пропорциональна $|V|^3$.

Главная часть - третья - обнаружение лишних столбцов. Временная сложность этой части равна $O(|V|^4)$. Если структура для хранения временной матрицы T представляет собой $std :: list < boost :: dynamic\_bitset <>>$ (список столбцов), то операция в самом внутреннем цикле хорошо векторно распараллеливается современными компиляторами, чем сильно уменьшается реальное время работы.

Наконец, последняя часть алгоритма составляет кодировочную матрицу B из оставшихся столбцов (=бинарных признаков). Ясно, что временная сложность этой части равна $O(|V|^2)$.

Теперь мы продемонстрируем работу этого алгоритма на примере (несколько цветов спор для массива Mushrooms из репозитория данных для тестирования алгоритмов машинного обучения Университета Калифорнии в г. Ирвайн):



На первом и втором шагах создается матрица отношения порядка. В матрицу пишется 1, если метка строки совпадает или расположена выше (более специфична), чем метка столбца

$$\begin{array}{c|cccc} & y & n & h & \boldsymbol{b} \\ y & 1 & 0 & 0 & \mathbf{0} \\ n & 0 & 1 & 0 & \mathbf{0} \\ h & 0 & 1 & 1 & \mathbf{0} \\ b & 1 & 1 & 0 & \mathbf{1} \end{array} \Rightarrow \begin{array}{ccc} y & n & h \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{array}$$

На третьем шаге в матрице помечаются те столбцы, которые являются побитовым умножением двух каких-то других столбцов.

Наконец, на четвертом шаге все отмеченные столбцы удаляются, а кодирование соответствует сокращенным битовым строкам.

В ключевой теореме 1.2) используется следующая лемма.

Лемма 1.8. *Для любой конечной решетки* $\boldsymbol{L}$ *в формальном контексте* $\geq\subseteq L \times L$ *для любого кандидата* $\langle A, B\rangle$ *найдется такой элемент* $x \in \boldsymbol{L}$*, что* $A = \{y \mid y \geq x\} = \{x\}'$ *и* $B = \{y \mid y \leq x\} = \{x\}''$.

Доказательство корректности этого алгоритма составляет утверждение следующей теоремы:

Теорема 1.2. *Для решетки кандидатов* $\boldsymbol{L}$*, порождаемой формальным контекстом* $\geq\subseteq L \times L$ *образ признака* $h(f) = \langle\{f\}', \{f\}''\rangle$ *является* $\cup$*-разложимым элементом, если и только если найдутся такие два признака* $f_1 < f$ *и* $f_2 < f$*, что* $\{f\}' = \{f_1\}' \cap \{f_2\}'$.

Доказательство. По определению 1.4 для $\cup$-разложимого элемента $\langle\{f\}', \{f\}''\rangle$ должна найтись такая пара $\langle A_1, B_1\rangle$ и $\langle A_2, B_2\rangle$, что

$$\langle\{f\}', \{f\}''\rangle = \langle A_1, B_1\rangle \cup \langle A_2, B_2\rangle.$$

По лемме 1.8 $\{f_1\}' = A_1$ и $\{f_2\}' = A_2$. С использованием уравнения (1.16)) и леммы 1.1 получаем

$$\langle\{f\}', \{f\}''\rangle = \langle A_1 \cap A_2, (A_1 \cap A_2)'\rangle = \langle\{f_1\}' \cap \{f_2\}', (\{f_1\}' \cap \{f_2\}')'\rangle.$$

Согласно определениям 1.2 и 1.4 из уравнений (1.12) и (1.13) получаем $f_1 \in \{f_1\}'' \subset \{f\}''$, т.е. $f_1 < f$. Аналогично, $f_2 < f$.

□

В работе Е.С.Панкратовой и автора [49] описаны представления битовыми строками сложных структур значений признаков для медицинских данных. Там рассматривалось несколько типов нижних полурешеток на значениях признаков, широко используемых в реальных экспериментальных исследованиях. Алгоритм 1 автоматизирует кодирование битовыми строками и расширяет его на более общий случай дискретных значений признаков, задающих произвольную нижнюю полурешетку. Замена операции сходства побитовым умножением позволяет эффективно использовать архитектуру современных ЭВМ.

Основные выводы

1. Некоторые особенности ДСМ-метода, требующие специального изучения вероятностных вариантов его познавательных процедур (индукции, абдукции и аналогии), могут быть продемонстрированы на формальном контексте для Булеана 1.21, в котором число сходств экспоненциально велико относительно размера контекста.

2. Задача вычисления сходства всегда сводится к случаю побитового умножения строк из 0 и 1 (по фундаментальной теореме анализа формальных понятий), что позволяет эффективно использовать архитектуру современных компьютеров.

3. Базовые операции «замыкай-по-одному» допускают ленивую организацию вычислений, для которой теорема 1.1 устанавливает степень увеличения эффективности не менее 2-х раз.

4. Можно предложить алгоритм 1 эффективного кодирования признаков, на значениях которых заданы нижние полурешетки, битовыми строками так, чтобы операция сходства совпадала с побитовым умножением. Корректность этого алгоритма устанавливается в теореме 1.2.

Глава 2

Переобучение при вычислении сходств

Во второй главе изложены результаты автора [16] и [14] о вероятности возникновения эффекта «переобучения» - порождении таких сходств обучающих примеров, которые не соответствуют никакой причине, а возникают лишь из-за одновременного наличия соответствующего общего фрагмента в нескольких объектах, при этом каждый из примеров-родителей имеет свой фрагмент-причину, отличную от этого фрагмента. Мы называем такие сходства «фантомными», так как они обнаруживаются тем же самым механизмом, что и реальные причины, но их использование для прогнозирования целевого свойства у тестовых примеров может приводить к ошибочному предсказанию.

Мы специально приводим в первом параграфе пример возникновения фантомного сходства, наглядно демонстрирующий нерелевантность таких сходств для предсказания целевого свойства.

Возникновение таких сходств следует рассматривать как мешающий фактор для проведения интеллектуального анализа данных с помощью операции сходства, подобно тому, как «переобучение» мешает применению многих методов машинного обучения. Для преодоления этой трудности можно предложить два подхода.

Первый - увеличить нижнюю границу на число объектов, порож-

дающих данное сходство. Такие объекты называются его *родителями*.

Следует заметить, что грубое применение правила отбрасывания сходств с малым числом родителей может привести к неправомерному отбрасыванию причин, для которых в обучающей выборке случайно оказалось недостаточно примеров. Так что, пытаясь избавиться от одной напасти (переобучения), мы впадаем в другую крайность - «недообучение».

Второй способ устранения фантомных сходств - это использовать контр-примеры. *Контр-примером* называется объект, представленный битовой строкой, который не обладает целевым свойством. Он может *устранить* сходство, если множество признаков сходства является подмножеством множества признаков контр-примера. Эта процедура называется *запретом контр-примеров*.

В параграфе 2.1 мы опишем вероятностную модель, основанную на случайном задании «сопутствующих» признаков, комбинации которых могут образовывать фантомные сходства, с помощью независимых серий испытаний Бернулли. Теорема 2.1 сформулирует оценку на вероятность успеха, чтобы фантомные сходства с заданным числом родителей возникали.

Следует отметить, что А.С. Опарышева в своей выпускной квалификационной работе [48] бакалавра, выполненной под руководством автора, продемонстрировала, что имеется значительная (от 10% по 22%) доля ДСМ-гипотез, подозрительных на «фантомность» в экспериментах по фармакологии. Запрет контр-примеров снижает степень переобучения (долю подозрительных гипотез), но не устраняет их полностью. Это показывает, вероятностная модель, положенная в основу всех математических результатов настоящей главы, является достаточно адекватным приближением к реальности.

В параграфе 2.2 мы представим негативный результат автора о положительности предельной (при числе признаков, стремящемся к бесконечности) вероятности возникновения фантомного сходства, неустранимого контр-примерами, даже тогда, когда вероятность появления каждого признака стремится к нулю, а число контр-примеров возрастает до бесконечности.

Параграф 2.3 описывает неасимптотические результаты автора о вероятности появления фантомного сходства при наличии контр-примеров. Мы получили производящие функции для вероятностей неустранения контр-примерами фантомного сходства при фиксированном и произвольном числе контр-примеров. Хочется надеяться, что асимптотический анализ этих производящих функций позволит оценить вероятность возникновения фантомного сходства, неустранимого контр-примерами, при расширении обучающей выборки (см. 3 открытый вопрос списка из Заключения).

2.1 Модель для переобучения

Мы начнем с наглядного примера:
Пусть

$$O = \{o_1 = B737, o_2 = SSJ100, o_3 = IL76, o_4 = A320\}$$

будет множеством самолетов, находящихся на ремонте, каждый из которых описывается проблемами из списка

$$F = \{f_1 = \text{оперение}, f_2 = \text{двигатель}, f_3 = \text{ругательство}\} :$$

$O \mid F$	f_1	f_2	f_3
o_1	1	0	0
o_2	1	0	1
o_3	0	1	1
o_4	0	1	0

Перечислим всех кандидатов для этого формального контекста в таблице:

i	A	B
1	$\emptyset$	$\{f_1, f_2, f_3\}$
2	$\{o_2\}$	$\{f_1, f_3\}$
3	$\{o_3\}$	$\{f_2, f_3\}$
4	$\{o_1, o_2\}$	$\{f_1\}$
5	$\{o_2, o_3\}$	$\{f_3\}$
6	$\{o_3, o_4\}$	$\{f_2\}$
7	$\{o_1, o_2, o_3, o_4\}$	$\emptyset$

Если рассмотреть непустые сходства не менее двух объектов, то мы получим две «настоящие» причины: $\langle\{o_1, o_2\}, \{f_1\}\rangle$ «самолет с поврежденным оперением не летает» и $\langle\{o_3, o_4\}, \{f_2\}\rangle$ «самолет с поврежденным двигателем не летает», и одно «фантомное» сходство $\langle\{o_2, o_3\}, \{f_3\}\rangle$ «самолет, на котором написано ругательство, не летает». Последний кандидат возник из-за случайного совпадения подмножества признаков $\{f_3\}$ у двух примеров o_2 и o_3, каждый из которых имеет свою отличную от других «настоящую» причину.

Этот наглядный пример, очевидно, может вызвать полемику: почему мы заранее не можем устранить признаки, оставив только существенные; почему мы используем нашу интуицию при отнесении третьего сходства к «фантомным»; почему все фрагменты состоят из единственного признака?

По поводу разделения признаков по степеням существенности имеются некоторые идеи у коллег автора, но он не будет здесь обсуждать чужие идеи. Отметим лишь, что признак f_3 может быть с таким же результатом называться «быть отечественного производства», что уже в значительно меньшей степени вызывает вопрос, почему вообще этот признак появляется при описании самолетов. Более того, необходимость его исключения тоже становится неочевидной, так как он может выступать частью реальной причины.

Однопризнаковость фрагментов в этом примере вызвана лишь желанием сократить несущественные детали. Понятно, что разрушение оперения и повреждение двигателя обычно описываются комбинациями многих признаков.

Ясно, что подобные фантомные сходства нежелательны, так как

они приводят в неправильному предсказанию целевых свойств у неисследованных примеров (предоставленных для прогноза).

В настоящей главе будут исследованы вероятности их появления. Но сначала мы должны определить вероятностную модель возникновения таких сходств.

Вопрос о том, является ли вероятностная модель, предложенная автором и описанная ниже, достаточно адекватным приближением к реальности, исследовался А.С. Опарышевой в ее выпускной квалификационной работе бакалавра, выполненной под руководством автора.

А.С. Опарышева разработала программу [48] для обнаружения подозрительных на «фантомность» ДСМ-гипотез, порожденных системой «ДСМ-решатель по фармакологии», созданной с.н.с. ФИЦ ИУ РАН, к.т.н. Д.А. Добрыниным.

В ДСМ-экспериментах исследовались причины мутагенной активности у замещенных нитробензолов. Обучающая выборка была подготовлена фармакологами Ливерпульского Университета (Великобритания). Биологически активными было 166 соединений.

Использовались кодирование фрагментарным кодом суперпозиций подструктур (ФКСП), разработанным к.х.н. В.Г. Блиновой и к.т.н. Д.А. Добрыниным, и MNA, предоставленным нам сотрудниками ИБМХ им. В.Н. Ореховича РАН.

Результаты исследований А.С. Опарышевой [48] суммированы в следующей таблице:

кодировка	контр-примеры	ДСМ-гипотез	подозрительных
ФКСП	нет	130	13
ФКСП	есть	247	32
MNA	есть	407	105

Очевидно, что имеется значительная (от 10% по 22%) доля ДСМ-гипотез, подозрительных на «фантомность». Запрет контр-примеров снижает степень переобучения (долю подозрительных гипотез), но не устраняет их полностью.

Зафиксируем все признаки (называемые *существенными*), участвующие в фрагментах, соответствующих «настоящим» причинам. Оставшиеся (называемые *сопутствующими*) признаки будут порождаться с помощью последовательностей испытаний Бернулли.

Последовательность n испытаний Бернулли — это распределение вероятностей на $\{0,1\}^n$ с

$$\mathbf{P}(x_1 = \delta_1, \ldots, x_n = \delta_n) = \prod_{j=1}^{n} p_j^{\delta_j} \cdot (1 - p_j)^{1-\delta_j},$$

где $n \in \mathbf{N}$ и $0 < p_j < 1$. Число p_j называется *вероятностью успеха* $x_j = 1$ в j-ом испытании.

Элементы формального контекста для вычисления сходств (мы их будем называть *обучающими примерами*) порождаются путем присоединения независимых реализаций этого вероятностного пространства (определяющих, какие сопутствующие признаки имеются у соответствующего примера) к заранее заданным «настоящим» причинам - битовым строкам, задающим соответствующие фрагменты.

Лемма 2.1. *Фантомное сходство b обучающих примеров является последовательностью $\langle a_1, \ldots, a_n \rangle$ испытаний Бернулли с вероятностью успеха $a_j = 1$ равной p_j^b, присоединенной к некоторой строке для существенным признаков, отличной от «настоящих» причин.*

Доказательство следует из условия независимости серий Бернулли и определения сходства как побитового умножения. Несовпадение «настоящих» причин обеспечивает отличие получаемой строки на месте существенных признаков от «настоящих» причин. □

В дальнейшем мы будем игнорировать признаки, необходимые для описания «настоящих» причин, сосредоточившись только на сопутствующих признаках, так как только комбинации таковых образуют фрагменты, соответствующие фантомным сходствам.

В этой главе, кроме последнего параграфа, нас будет интересовать случай, когда все сопутствующие признаки равновероятны: $\forall j\, [p_j = p]$. Такие обучающие примеры мы будем называть *случайными p-примерами*. Это требование носит технический характер, и в теореме 2.1 может быть заменена на минимум из p_j.

Лемма 2.2. *С вероятностью, равной* $1-(1-p^b)^n$*, сходство* b *независимых случайных* p*-примеров* $\langle x_1^1, \ldots, x_n^1\rangle, \ldots, \langle x_1^b, \ldots, x_n^b\rangle$ *является нетривиальным.*

Доказательство легко следует из леммы 2.1. □

Попробуем избежать возникновения фантомных сходств увеличением порога b на число родителей (наименьшее число обучающих примеров, сходства которых допускаются).

Теорема 2.1. *Для* $p \geq (-\ln(1-\varepsilon)/n)^{1/b}$ *вероятность появления фантомного сходства* b *случайных* p*-примеров не меньше, чем* $\varepsilon > 0$*.*

Доказательство. Из-за выпуклости e^{-u} имеем неравенство

$$(1-u) \leq e^{-u}$$

(e^{-u} лежит над касательной $(1-u)$ к ней в нуле). Применяем лемму 2.1 вместе с этим неравенством и получаем

$$1-(1-p^b)^n \geq 1-e^{-p^b \cdot n} \geq \varepsilon.$$

□

Из этой теоремы вытекает, что увеличение порога b на число родителей не сработает: когда число сопутствующих признаков велико, то даже при достаточно малой вероятности появления каждого признака вероятность возникновения фантомного сходства положительна.

Автор хотел бы подчеркнуть, что грубое применение правила отбрасывания сходств с малым числом родителей может привести к неправомерному отбрасыванию причин, для которых в обучающей выборке оказалось недостаточно примеров. Пытаясь избавиться от одной напасти (переобучения), мы впадаем в другую крайность - «недообучение».

Как уже говорилось в начале настоящей главы, остается надежда на контр-примеры. Напомним, что *контр-примером* называется объект, представленный битовой строкой, который не обладает целевым свойством.

Контр-примеры для устранения фантомных сходств получаются путем присоединения независимых реализаций последовательностей испытаний Бернулли к нулевым строкам, соответствующим существенным признакам, используемых для «настоящих» причин. В дальнейшем мы будем игнорировать существенные признаки.

Определение 2.1. *Контр-пример* $\langle y_1, \ldots, y_n \rangle$ ***устраняет*** *фантомное сходство* $\langle a_1, \ldots, a_n \rangle$*, если*

$$\forall j \, [a_j = 1 \Rightarrow y_j = 1] \, .$$

Другими словами, если отождествить битовую строку со множеством признаков, на местах соответствующих которым стоит единица, то контр-пример устраняет фантомное сходство, если множество признаков в сходстве является подмножеством множества признаков контр-примера.

В следующем параграфе мы получим асимптотическую оценку на вероятность возникновения фантомного сходства при наличии контр-примеров. Этот отрицательный результат указывает, что подход к интеллектуальному анализу данных, основанный на вычислении всех сходств среди обучающих примеров, демонстрирует эффект «переобучения», когда построенная исчерпывающая модель будет иметь плохую предсказательную силу.

2.2 Предельная вероятность переобучения

Рассмотрим модель формирования множества контр-примеров и обучающих примеров, основанную на независимых последовательностях испытаний Бернулли, описанную в предыдущем параграфе.

В этом параграфе число n обозначает количество сопутствующих (не входящих ни в какую «настоящую» причину) признаков, которыми мы ограничиваемся. Для каждого контр-примера или обучающего примера образуем последовательность n испытаний Бернулли с одинаковой вероятностью успеха p, причем последовательности для разных объектов независимы. Число m будет равно числу контр-примеров.

Мы ограничимся в этом параграфе случаем двух родителей фантомного сходства (это является стандартным условием, применяемым в ДСМ-методе). Из теоремы 2.1 вытекает, что вероятность успеха должна быть не меньше $\sqrt{\frac{-\ln(1-\varepsilon)}{n}}$, чтобы возникло хотя бы одно фантомное сходство. Мы обозначим $a = -\ln(1-\varepsilon)$ и предположим, что $a \leq 1$, т.е. $0 < \varepsilon \leq 1 - e^{-1}$.

Лемма 2.3. *С вероятностью*

$$\sum_{j=0}^{m} \binom{m}{j} \cdot (-1)^j \cdot (1 - p^2 + p^{2+j}))^n$$

возникнет фантомное сходство, которое не устранится ни одним из m случайных контр-примеров.

Доказательство. По лемме 2.1 фантомному сходству двух обучающих примеров соответствует последовательность n испытаний Бернулли с вероятностью успеха p^2.

Зафиксируем мощность (= количество единиц = число сопутствующих признаков) l фантомного сходства. Тогда искомая вероятность равна

$$\sum_{l=0}^{n} \binom{n}{l} \cdot \left(p^2\right)^l \cdot \left(1 - p^2\right)^{n-l} \cdot \left(1 - p^l\right)^m,$$

так как p^l равна вероятности устранения этого сходства фиксированным контр-примером, поэтому $\left(1 - p^l\right)^m$ – вероятность того фантомное сходство не устранится ни одним из m случайных контр-примеров, а $\binom{n}{l} \cdot \left(p^2\right)^l \cdot \left(1 - p^2\right)^{n-l}$ – вероятность порождения фан-

томного сходства мощности l. Но

$$\sum_{l=0}^{n} \binom{n}{l} \cdot \left(p^2\right)^l \cdot \left(1-p^2\right)^{n-l} \cdot \left(1-p^l\right)^m =$$

$$= \sum_{l=0}^{n} \binom{n}{l} \cdot \left(p^2\right)^l \cdot \left(1-p^2\right)^{n-l} \cdot \left(\sum_{j=0}^{m} \binom{m}{j} \cdot (-1)^j \cdot p^{lj}\right) =$$

$$= \sum_{j=0}^{m} \binom{m}{j} \cdot (-1)^j \cdot \left(\sum_{l=0}^{n} \binom{n}{l} \cdot \left(p^{2+j}\right)^l \cdot \left(1-p^2\right)^{n-l}\right) =$$

$$= \sum_{j=0}^{m} \binom{m}{j} \cdot (-1)^j \cdot (1-p^2+p^{2+j}))^n.$$

□

С помощью принципа «включение-исключение» можно провести прямое доказательство леммы 2.3. Мы получим этот же самый результат в следующем параграфе как следствие теоремы 2.3.

Сформулируем теперь несколько вспомогательных лемм:

Лемма 2.4. *Для любой константы c верно*

$$\sum_{j=0}^{m} \binom{m}{j} \cdot (-1)^j \cdot c = 0.$$

Требуемое равенство следует из формулы бинома Ньютона. □

Лемма 2.5. *Для натурального $j > 0$ и $a \leq 1$ при $n \to \infty$ имеем*

$$\left(1 - \frac{a}{n} + \left(\frac{a}{n}\right)^{1+j/2}\right)^n - \left(1-\frac{a}{n}\right)^n =$$

$$= a \cdot e^{-a} \cdot \left(\frac{a}{n}\right)^{j/2} + \frac{a^2}{2} \cdot e^{-a} \cdot \left(\frac{a}{n}\right)^j + r_n,$$

где остаточный член оценивается сверху

$$r_{n,j} \leq a \cdot e^{-a} \cdot \left(\frac{a}{n}\right)^{j/2} \cdot \left(1-e^{1/n}\right) + \frac{a^2}{2} \cdot e^{-a} \cdot \left(\frac{a}{n}\right)^j \cdot \left(1-e^{2/n}\right) + \frac{e^a}{n^{3j/2}}$$

и снизу $-\frac{a^{2+j}}{2n^{1+j}} \cdot e^{-a} \le r_{n,j}$.

Доказательство. Применим формулу бинома Ньютона

$$\left[1-\frac{a}{n}+\left(\frac{a}{n}\right)^{1+j/2}\right]^n - \left(1-\frac{a}{n}\right)^n = n\cdot\left(1-\frac{a}{n}\right)^{n-1}\cdot\left(\frac{a}{n}\right)^{1+j/2} +$$

$$+\frac{n(n-1)}{2}\cdot\left(1-\frac{a}{n}\right)^{n-2}\cdot\left(\frac{a}{n}\right)^{2+j}+\sum_{k=3}^{n}\binom{n}{k}\cdot\left(1-\frac{a}{n}\right)^{n-k}\cdot\left(\frac{a}{n}\right)^{(1+j/2)k} \le$$

$$\le \left(1-\frac{a}{n}\right)^{n-1}\cdot\frac{a^{1+j/2}}{n^{j/2}} + \left(1-\frac{a}{n}\right)^{n-2}\cdot\frac{a^{2+j}}{2n^j} + \frac{1}{n^{3j/2}}\cdot\sum_{k=0}^{\infty}\frac{a^k}{k!} \le$$

$$\le \left(1-\frac{a}{n}\right)^{n-1}\cdot\frac{a^{1+j/2}}{n^{j/2}} + \left(1-\frac{a}{n}\right)^{n-2}\cdot\frac{a^{2+j}}{2n^j} + \frac{e^a}{n^{3j/2}}.$$

Используем оценку (при $n \ge m$ для $m = 1, 2$)

$$\left(1-\frac{a}{n}\right)^{n-m} = \exp\{(n-m)\cdot\ln\left(1-\frac{a}{n}\right)\} =$$

$$= \exp\{-(n-m)\cdot\sum_{k=1}^{\infty}\frac{a^k}{k\cdot n^k}\} \le \exp\{-a+\frac{m\cdot a}{n}\} = e^{-a}\cdot e^{\frac{am}{n}}.$$

Первые два слагаемых оцениваются так

$$\left(1-\frac{a}{n}\right)^{n-1}\cdot\frac{a^{1+j/2}}{n^{j/2}} + \left(1-\frac{a}{n}\right)^{n-2}\cdot\frac{a^{2+j}}{2n^j} \le$$

$$\le e^{-a}\cdot\frac{a^{1+j/2}}{n^{j/2}}\cdot e^{\frac{a}{n}} + e^{-a}\cdot\frac{a^{2+j}}{2n^j}\cdot e^{\frac{2a}{n}}.$$

Оценка снизу получается тривиально.

□

Лемма 2.6. *При* $n \to \infty$ *имеем разложение*

$$\left(1-\frac{a}{n}\right)^n = e^{-a} - e^{-a}\cdot\frac{a^2}{2n} + e^{-a}\cdot\frac{a^3(3a-8)}{24n^2} + O\left(n^{-3}\right).$$

Доказательство аналогично предыдущему и общеизвестно. □

Теперь все готово, чтобы доказать основную теорему этого параграфа и статьи автора [16]:

Теорема 2.2. *При числе сопутствующих признаков $n \to \infty$ и вероятности появления этих признаков у контр-примеров и обучающих примеров, равной $p = \sqrt{\frac{a}{n}}$, вероятность возникновения фантомного сходства двух обучающих примеров, не устраненного никаким из $m = c \cdot \sqrt{n}$ контр-примеров, будет стремиться к*

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

Доказательство. Оценим $\sum_{j=0}^{m} \binom{m}{j} \cdot (-1)^j \cdot (1 - p^2 + p^{2+j}))^n$ и сошлемся на лемму 2.3.

Подставляя $p = \sqrt{\frac{a}{n}}$, по лемме 2.4 получим

$$\sum_{j=0}^{m} \binom{m}{j} \cdot (-1)^j \cdot \left(1 - \frac{a}{n} + \left(\frac{a}{n}\right)^{1+j/2}\right)^n =$$

$$= \sum_{j=0}^{m} \binom{m}{j} \cdot (-1)^j \cdot \left[\left(1 - \frac{a}{n} + \left(\frac{a}{n}\right)^{1+j/2}\right)^n - \left(1 - \frac{a}{n}\right)^n\right].$$

Далее применяем леммы 2.5 и 2.6

$$\sum_{j=0}^{m} \binom{m}{j} \cdot (-1)^j \cdot \left[\left(1 - \frac{a}{n} + \left(\frac{a}{n}\right)^{1+j/2}\right)^n - \left(1 - \frac{a}{n}\right)^n\right] =$$

$$= \left[1 - \left(1 - \frac{a}{n}\right)^n\right] +$$

$$+ \sum_{j=1}^{m} \binom{m}{j} \cdot (-1)^j \cdot \left[\left(1 - \frac{a}{n} + \left(\frac{a}{n}\right)^{1+j/2}\right)^n - \left(1 - \frac{a}{n}\right)^n\right] =$$

$$= \left[1 - e^{-a} + O(n^{-1})\right] +$$

$$+ \sum_{j=1}^{m} \binom{m}{j} \cdot (-1)^j \cdot \left[a \cdot e^{-a} \cdot \left(\frac{a}{n}\right)^{j/2} + \frac{a^2}{2} \cdot e^{-a} \cdot \left(\frac{a}{n}\right)^j + r_{n,j}\right].$$

Так как $m = c \cdot \sqrt{n} = O(n^{1/2})$, то

$$\sum_{j=1}^{m} \binom{m}{j} \cdot (-1)^j \cdot r_{n,j} = O(n^{-1}).$$

Во втором слагаемом главный вклад у первого элемента

$$\sum_{j=1}^{m} \binom{m}{j} \cdot (-1)^j \cdot \left[\frac{a^2}{2} \cdot e^{-a} \cdot \left(\frac{a}{n}\right)^j\right] =$$

$$= -m \cdot \frac{a^3}{2n} \cdot e^{-a} + O(n^{-1}) = -e^{-a} \cdot \frac{c \cdot a^3}{2\sqrt{n}} + O(n^{-1}).$$

Наконец, оценка

$$\sum_{j=1}^{m} \binom{m}{j} \cdot (-1)^j \cdot \left[a \cdot e^{-a} \cdot \left(\frac{a}{n}\right)^{j/2}\right] =$$

$$= \sum_{j=1}^{m} \binom{m}{j} \cdot (-1)^j \cdot \left[a \cdot e^{-a} \cdot \left(\frac{c \cdot \sqrt{a}}{m}\right)^j\right] =$$

$$= a \cdot e^{-a} \cdot \left[\left(1 - \frac{c \cdot \sqrt{a}}{m}\right)^m - 1\right] =$$

$$= a \cdot e^{-a} \cdot \left[e^{-c \cdot \sqrt{a}} - 1 - e^{-c\sqrt{a}} \cdot \frac{c \cdot a}{2\sqrt{n}}\right] + O(n^{-1}),$$

где последнее равенство следует из леммы 2.6, завершает доказательство. □

Легко увидеть, что

$$1 - e^{-a} - a \cdot e^{-a} \cdot \left[1 - e^{-c \cdot \sqrt{a}}\right] > 0 \tag{2.1}$$

для любых положительных a и c.

Достаточно выражение в скобках заменить единицей (вероятность только уменьшится), а затем заметить, что результат совпадет с вероятностью для случайной величины Пуассона (со средним a) принять значение больше единицы. Эта вероятность, очевидно, строго положительна.

Неравенство 2.1 означает, что при большом числе сопутствующих признаков, даже если вероятность появления каждого такого признака мала, а число контр-примеров не слишком велико, существует ненулевая вероятность появления фантомного сходства.

Конечно, наша модель случайных контр-примеров будет неадекватной, если их число m сравнимо с 2^n, так как тогда многие контр-примеры будут повторяться. Впрочем, с практической точки зрения наличие экспоненциального числа контр-примеров тоже нереалистично (уж очень быстро растет экспонента).

Интересно исследовать случай, когда число m контр-примеров растет, а число n признаков постоянно. В следующем параграфе мы найдем производящие функции для вероятностей появления фантомного сходства при фиксированном и переменном числе контр-примеров. Можно надеяться, что в этом интересном случае асимптотический анализ этих производящих функций позволит получить стремящуюся к нулю вероятность возникновения фантомного сходства, неустранимого контр-примерами (см. 3 открытую проблему из списка в Заключении).

2.3 Производящие функции для переобучения

Настоящий параграф посвящен выводу производящих функций для вероятностей появления фантомного сходства при фиксированном и переменном числе контр-примеров. Первоначально эти производящие функции были опубликованы в работе автора [14].

Напоминаем, что мы используем вероятностную модель возникновения фантомного сходства и случайных контр-примеров из параграфа 2.1. Через n мы обозначим число сопутствующих признаков, которыми мы и ограничиваемся. В этом параграфе ситуация рассматривается в максимальной общности: число b - граница на число родителей - минимальное число обучающих примеров, участвующих в порождении фантомного сходства - может быть любой. Более того, вероятность появления j-ого признака - вероятность успеха в j-ом

испытании серии Бернулли - равна p_j, то есть может изменяться в зависимости от признака.

Сначала рассматривается случай фиксированного числа m контр-примеров.

Для вывода производящих функций мы воспользуемся техникой конечных цепей Маркова [37], производящими функциями [55] (многочленами) для конечных распределений вероятностей и операторами перехода [28], действующими на многочленах.

Рассмотрим процесс одновременного задания t-ых признаков

$$\langle a_1, \ldots, a_t \ldots, a_n \rangle$$

фантомного сходства и всех контр-примеров

$$\left\langle y_1^1, \ldots, y_t^1 \ldots, y_n^1 \right\rangle, \ldots, \left\langle y_1^m, \ldots, y_t^m \ldots, y_n^m \right\rangle.$$

Ясно, что это возможно из-за независимости n испытаний Бернулли для фантомного сходства и контр-примеров.

Определение 2.2. *Назовем* ***выжившими*** *на шаге t контр-примеры $\left\langle y_1^k, \ldots, y_t^k \ldots, y_n^k \right\rangle$, для которых $\forall j \leq t \left[a_j = 1 \Rightarrow y_j^k = 1 \right]$.*

В момент времени t состоянием цепи Маркова [37] будет число $X_t^{(m)}$ контр-примеров, выживших после одновременного нахождения t-ых признаков m контр-примеров и фантомного сходства. Ясно, что это число должно быть элементом множества $S = \{0, 1, \ldots, m\}$.

Определение 2.3. ***Цепью Маркова*** *на множестве $S = \{0, 1, \ldots, m\}$ состояний назовем $(m+1) \times (m+1)$-матрицу $P = (p_{i,j})$ из неотрицательных чисел, удовлетворяющую условию, что все суммы по строкам равны 1: $\sum_{j=0}^{m} p_{i,j} = 1$. Матрицы такого вида называются* ***стохастическими****. Элемент $p_{i,j} = \boldsymbol{P}\left[X_{t+1}^{(m)} = j \mid X_t^{(m)} = i \right]$ этой матрицы — это вероятность перехода в момент времени $t+1$ в состояние $X_{t+1}^{(m)} = j$, стартуя в момент времени t из состояния $X_t^{(m)} = i$.*

Так как первоначально было m контр-примеров (на это указывает верхний индекс), то $X_0^{(m)} = m$ с вероятностью 1. Нас интересует $\mathbf{P}\left[X_n^{(m)} = 0\right]$ — вероятность того, что после определения всех n признаков ни один из m контр-примеров не будет выжившим.

Лемма 2.7. *Матрица перехода цепи Маркова имеет элементы:*

$$p_{s+r,s} = \begin{cases} \left(1 - p_{t+1}^b\right) + p_{t+1}^b \cdot p_{t+1}^s, & \text{если } r = 0 \\ p_{t+1}^b \cdot \binom{s+r}{r} \cdot \left(1 - p_{t+1}\right)^r \cdot p_{t+1}^s, & \text{если } 0 < r \leq m - s \end{cases}$$

Доказательство. Если после определения t-ых признаков оставалось $s + r$ $(0 < r \leq m - s)$ выживших контр-примеров, а после определения $(t+1)$-ых признаков их осталось s, то t-й признак a_{t+1} сходства обязан быть единицей $a_{t+1} = 1$, что происходит с вероятностью p_{t+1}^b, и какие-то r из $s + r$ выживших контр-примеров должны получить нули в $(t+1)$-ой позиции, а остальные контр-примеры должны получить единицы в $(t+1)$-ой позиции, вероятность чего равна $\binom{s+r}{r} \cdot \left(1 - p_{t+1}\right)^r \cdot p_{t+1}^s$.

Верхняя строка (когда число выживших контр-примеров не уменьшается) вычисляется разбором случаев, когда $(t+1)$-й признак сходства равен нулю $a_{t+1} = 0$, что происходит с вероятностью $1 - p_{t+1}^b$, и когда $a_{t+1} = 1$ и все s выживших контр-примеров получают единицы в $(t+1)$-ой позиции, вероятность чего равна $p_{t+1}^b \cdot p_{t+1}^s$. □

При вычислении вероятности p_{t+1}^b того, что $(t+1)$-ый признак фантомного сходства равен единице $(a_{t+1} = 1)$, мы использовали лемму 2.1.

Определение 2.4. ***Производящей функцией*** *для конечного распределения вероятностей $p : S \to [0, 1]$ на множестве $S = \{0, 1, \ldots, m\}$ состояний назовем многочлен $\varphi(z) = \sum_{j=0}^m p(j) \cdot z^j$.*

Производящие функции (многочлены) для распределений

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

будем обозначать через

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

Очевидно, что $\varphi_0^{(m)}(z) = z^m$. При этом нас интересует число

$$\varphi_n^{(m)}(0) = \mathbf{P}\left[X_n^{(m)} = 0\right].$$

Лемма 2.8. *Производящие многочлены $\varphi_t^{(m)}(z)$ связаны следующим образом:*

$$\varphi_{t+1}^{(m)}(z) = \left(1 - p_{t+1}^b\right) \cdot \varphi_t^{(m)}(z) + p_{t+1}^b \cdot \varphi_t^{(m)}\left(p_{t+1} \cdot z + (1 - p_{t+1})\right).$$

Доказательство. Пусть

$$\varphi_t^{(m)}(z) = \sum_{s=0}^{m} w_s \cdot z^s, \varphi_{t+1}^{(m)}(z) = \sum_{s=0}^{m} v_s \cdot z^s,$$

где $w_s = \mathbf{P}\left[X_t^{(m)} = s\right]$ и $v_s = \mathbf{P}\left[X_{t+1}^{(m)} = s\right]$. Тогда

$$\begin{aligned} v_s = \mathbf{P}\left[X_{t+1}^{(m)} = s\right] = \\ = \sum_{r=0}^{m-s} \mathbf{P}\left[X_{t+1}^{(m)} = s \mid X_t^{(m)} = s + r\right] \cdot \mathbf{P}\left[X_t^{(m)} = s + r\right] = \\ = \sum_{r=0}^{m-s} \mathbf{P}\left[X_{t+1}^{(m)} = s \mid X_t^{(m)} = s + r\right] \cdot w_{s+r} = \\ \left(1 - p_{t+1}^b\right) \cdot w_s + \sum_{r=0}^{m-s} p_{t+1}^b \cdot \binom{s+r}{r} \cdot (1 - p_{t+1})^r \cdot p_{t+1}^s \cdot w_{s+r}. \end{aligned}$$

по лемме 2.7. Поэтому получаем

$$\varphi_{t+1}^{(m)}(z) = \sum_{s=0}^{m} v_s \cdot z^s = \left(1 - p_{t+1}^b\right) \cdot \sum_{s=0}^{m} w_s \cdot z^s + \\ + p_{t+1}^b \cdot \sum_{s=0}^{m} \sum_{r=0}^{m-s} \binom{s+r}{r} \cdot (1 - p_{t+1})^r \cdot w_{s+r} \cdot (p_{t+1} \cdot z)^s .$$

Переставляя суммирования и вводя обозначение $k = s + r$, получим

$$\varphi_{t+1}^{(m)}(z) = \left(1 - p_{t+1}^b\right) \cdot \varphi_t^{(m)}(z) + \\ + p_{t+1}^b \cdot \sum_{k=0}^{m} \sum_{r=0}^{k} \binom{k}{r} \cdot (1 - p_{t+1})^r \cdot w_k \cdot (p_{t+1} \cdot z)^{k-r} = \\ = \left(1 - p_{t+1}^b\right) \cdot \varphi_t^{(m)}(z) + \\ + p_{t+1}^b \cdot \sum_{k=0}^{m} w_k \cdot \left[\sum_{r=0}^{k} \binom{k}{r} \cdot (1 - p_{t+1})^r \cdot (p_{t+1} \cdot z)^{k-r}\right] = \\ = \left(1 - p_{t+1}^b\right) \cdot \varphi_t^{(m)}(z) + p_{t+1}^b \cdot \sum_{k=0}^{m} w_k \cdot (p_{t+1} \cdot z + (1 - p_{t+1}))^k = \\ = \left(1 - p_{t+1}^b\right) \cdot \varphi_t^{(m)}(z) + p_{t+1}^b \cdot \varphi_t^{(m)}\left(p_{t+1} \cdot z + (1 - p_{t+1})\right).$$

Сворачивание выражения в квадратных скобках происходит по формуле бинома Ньютона, что завершает доказательство. □

Преобразование

$$q(z) \mapsto \left(1 - p_{t+1}^b\right) \cdot q(z) + p_{t+1}^b \cdot q\left(p_{t+1} \cdot z + (1 - p_{t+1})\right)$$

задает линейный оператор L_{t+1} на многочленах.

Применяя лемму 2.8 n раз, мы получим, что

$$\varphi_n^{(m)}(z) = L_n\left[L_{n-1}\left[\ldots\left[L_1[z^m]\right]\ldots\right]\right], \tag{2.2}$$

так как $\varphi_0^{(m)}(z) = z^m$. Как уже говорилось ранее, нам нужно вычислить значение этого многочлена в нуле: $\varphi_n^{(m)}(0) = \mathbf{P}\left[X_n^{(m)} = 0\right]$

Для вычислений окажется полезным общий собственный базис для всех операторов $L_1, \ldots, L_n$ [28].

Как обычно, *собственным базисом* линейного оператора L, действующего в векторном пространстве V размерности d, называется базис из *собственных векторов*, то есть таких элементов

$$\{v_1, \ldots, v_d\} \subseteq V,$$

что найдутся такие числа $\{\lambda_1, \ldots, \lambda_d\}$, для которых выполняются равенства

$$L[v_j] = \lambda_j \cdot v_j$$

для всех $j = 1, \ldots, d$.

Лемма 2.9. *Для каждого оператора L_t, действующего в $(m+1)$-мерном пространстве многочленов степени $\leq m$ собственный базис образуют многочлены $(z-1)^j$ с собственными значениями $p_t^{b+j} + \left(1 - p_t^b\right)$.* □

Имея собственный базис, остается только разложить по нему исходный полином $\varphi_0^{(m)}(z) = z^m$

Лемма 2.10. $\varphi_0^{(m)}(z) = \sum_{j=0}^{m} \binom{m}{j} \cdot (z-1)^j$.

Доказательство. По формуле бинома Ньютона

$$\varphi_0^{(m)}(z) = z^m = (1 + (z-1))^m = \sum_{j=0}^{m} \binom{m}{j} \cdot (z-1)^j .$$

□

Теорема 2.3. $\varphi_n^{(m)}(z) = \sum_{j=0}^{m} \binom{m}{j} \cdot \prod_{t=1}^{n} \left[p_t^{b+j} + \left(1 - p_t^b\right)\right] \cdot (z-1)^j$

Доказательство. Собираем вместе утверждения лемм 2.9 и 2.10. □

Подставляя в результат теоремы 2.3 нуль вместо z, получаем при $b = 2$ другое доказательство леммы 2.3.

Переходя к случаю переменного числа контр-примеров, заметим, что каждый оператор L_t не зависит от числа m контр-примеров, и их собственные базисы совпадают.

Определение 2.5. ***Производящей функцией*** *для последовательности вероятностей* $\boldsymbol{P}\left[X_n^{(m)} = 0\right]$, *где* $m = 0, 1, \ldots$, *назовем формальный степенной ряд*

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

Двойной производящей функцией *для* $\boldsymbol{P}\left[X_n^{(m)} = s\right]$ *назовем формальный ряд*

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

Лемма 2.11. $\varphi_0(z, u) = \sum_{j=0}^{m} \frac{u^j}{(1-u)^{j+1}} \cdot (z-1)^j$ *- разложение по собственному базису.*

Доказательство.

$$\varphi_0(z, u) = \sum_{m=0}^{\infty} z^m \cdot u^m = \frac{1}{1 - z \cdot u} = \frac{1}{1-u} \cdot \frac{1}{1 - \frac{u \cdot (z-1)}{1-u}} =$$
$$= \sum_{j=0}^{m} \frac{u^j}{(1-u)^{j+1}} \cdot (z-1)^j.$$

□

Теорема 2.4. $\varphi_n(0, u) = \sum_{j=0}^{\infty} \prod_{t=1}^{n} \left[p_t^{b+j} + \left(1 - p_t^b\right)\right] \cdot \frac{(-u)^j}{(1-u)^{j+1}}$.

Доказательство. Аналогично доказательству теоремы 2.3, но вместо леммы 2.10 нужно использовать лемму 2.11 совместно с леммой 2.9, а потом подставить нуль вместо z. □

Основные выводы

1. Возникновение фантомных сходств (общих фрагментов обучающих примеров, каждый из которых имеет отличную от этого фрагмента структурную причину) мешает правильному предсказанию целевых свойств (наблюдается эффект переобучения).

2. Теоремы 2.1 и 2.2) утверждают о недостаточности двух механизмов устранения сходств (запрет контр-примеров и ограничения на число родителей) для полного устранения эффекта переобучения.

3. Ограничение на число родителей может привести к эффекту «недообучения» - неправомерному отбрасыванию причин, для которых не нашлось достаточного числа обучающих примеров.

4. Явный вид производящих функций (теоремы 2.3 и 2.4) для вероятности переобучения при наличии контр-примеров позволит получать приближенную оценку на число фантомных сходств на основе характеристик обучающей выборки и границы на число родителей.

Глава 3

Вероятностный поиск кандидатов

Отвечая на вызовы, описанные во введении и предыдущих главах, мы предлагаем использовать вероятностный подход к интеллектуальному анализу данных с использованием прикладной теории решеток. Если некоторые сходства заведомо плохи (фантомные), а огромное большинство сходств предсказывает по аналогии примеры одинаковым образом, то нет никакой необходимости вычислять их все, достаточно найти случайное подмножество сходств.

В этой главе мы будем обсуждать вероятностные алгоритмы для нахождения сходств и их свойства.

Первый параграф опишет несколько алгоритмов, первоначально предложенных в работе автора [9], для спаривающих вариантов которых удалось доказать останавливаемость с вероятностью единица (следствие из теоремы 3.2). Для спаривающей цепи Маркова доказана теорема 3.3 об изменении вероятностей эргодических состояний для спаривающей цепи Маркова, остановленной по верхней оценке на основе предварительных прогонов, по сравнению с исходной цепью Маркова.

Параграф 3.2 содержит описание других цепей Маркова, подходящих для вероятностного вычисления сходств (немонотонная и монотонная цепи Маркова), для которых, однако, имеется проблема

момента остановки. В этом параграфе на примерах демонстрируются некоторые простейшие свойства предложенных цепей Маркова.

Параграф 3.3 содержит теорему 3.4 об оценке среднего времени склеивания спаривающей цепи Маркова и теорему 3.5 о сильной концентрации времени склеивания около его среднего в частном случае Булевой алгебры всех подмножеств признаков (Булеана).

В параграфе 3.4 приводятся ключевые понятия и результаты о времени перемешивания произвольной цепи Маркова. Вопрос о времени перемешивания существенен даже для случая спаривающей цепи Маркова, так как после склеивания она совпадает с монотонной цепью Маркова.

Для случая Булеана в параграфе 3.5 получена верхняя оценка (3.15) времени перемешивания монотонной цепи Маркова и доказана теорема 3.9 об асимптотической точности этой оценки.

3.1 Цепи Маркова для поиска сходств

В работе автора [9] были рассмотрены три алгоритма для вероятностного поиска сходств: немонотонный, монотонный и спаривающий. В случае решетки всех подмножеств признаков - Булеана - первые два оказались классическими алгоритмами случайного блуждания и ленивого случайного блуждания, соответственно.

Все эти алгоритмы используют операции «замыкай-по-одному», введенные в определении 1.3 главы 1:

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

для кандидата $\langle A, B\rangle$ и объекта $o \in O$ и

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

для кандидата $\langle A, B\rangle$ и признака $f \in F$.

Ниже мы представим формальное представление этих алгорит-

мов:

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;

$A := O$; $B = O'$;

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

  $R := (O \setminus A) \cup (F \setminus B)$;

  Выбираем случайный элемент $r \in R$;

  if $(r \in O \setminus A)$ **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: Немонотонная цепь Маркова

Очевидно, что решетке-Булеане алгоритм 2 никогда не использует вторые случаи в уравнениях (1.22) и (1.23), а с равной $\frac{1}{n}$ вероятностью переходит с текущего подмножества на одного из n его соседей, то есть представляет собой случайное блуждание по соответствую-

щему гиперкубу.

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 3: Монотонная цепь Маркова

В случае Булеана алгоритм 3 с вероятностью $\frac{n}{2\cdot n} = \frac{1}{2}$ попадает на вторые случаи в уравнениях (1.22) и (1.23), 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.

Заметим, что корректность алгоритмов 2 и 3 (т.е. то, что в результате их работы мы обязательно получим кандидат) следует из леммы 1.2 главы 1.

Основная проблема с алгоритмами 2 и 3, применяемыми к произвольным формальным контекстам, состоит в том, что неизвестна никакая оценка на «время остановки» T, которая обеспечивала бы хорошую «перемешиваемость» соответствующей цепи Маркова. Подробнее о скорости перемешивания можно прочитать в параграфе 3.4. Таким образом, вопрос о выборе параметра T (фактически, о длине цикла) в этих алгоритмах остается открытым. Этот вопрос составляет открытую проблему 2 списка направлений дальнейших

исследований из Заключения.

Дополнительные характеристики цепей Маркова, задаваемых алгоритмами 2 и 3, будут представлены в параграфе 3.2. Там же мы обсудим причины, по которым эти алгоритмы получили свое название.

Гораздо более интересными, чем вышеописанные алгоритмы, являются те или иные варианты спаривающих цепей Маркова. Состоянием таких алгоритмов является упорядоченная пара кандидатов. Эти цепи Маркова обладают естественным моментом остановки - склеиванием - первым шагом, на котором пары кандидатов совпадают.

Ниже мы приводим классический вариант спаривающей цепи:

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 4: Спаривающая цепь Маркова

Состоянием изменяемых переменных в цикле (= состоянием цепи Маркова) является упорядоченная пара кандидатов $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$.

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 4.

Следующая теорема из статьи автора [77] объясняет, откуда здесь возникает цепь Маркова:

Теорема 3.1. *Алгоритм 4 соответствует цепи Маркова.*

Доказательство. Операция $CbODown$ относительно фиксированного объекта определяет детерминистскую функцию (не являющуюся биекцией, в общем случае) из множества кандидатов в себя. Поэтому каждая строка соответствующей матрицы перехода имеет единицу в одной из ячеек и нули в остальных местах. Очевидно, что такая матрица является стохастической.

Аналогичный факт верен для операции $CbOUp$ относительно фиксированного признака.

Так как преобразования кандидатов определяются случайным (равномерным) выбором или объекта, или признака, то матрица перехода является взвешенной (равномерно) суммой соответствующих матриц для каждого признака и каждого объекта. Но взвешенная сумма стохастических матриц сама является стохастической матрицей.

Одновременное применение операций «замыкай-по-одному» относительно или некоторого признака, или некоторого объекта к упорядоченной паре кандидатов соответствует тензорному (Кронекеровскому) произведению матрицы преобразований саму на себя с ограничением на инвариантное подпространство таких пар.

То, что это подпространство инвариантно, следует из леммы 1.3 в главе 1. Тензорное произведение двух стохастических матриц является стохастической матрицей. Ограничение стохастической матрицы на инвариантное подпространство оставляет матрицу стохастиче-

ской. Поэтому матрица преобразования, соответствующая алгоритму 4 задает конечную цепь Маркова. □

Следует отметить, что первая часть доказательства предыдущей теоремы обосновывает, почему алгоритмы 2 и 3 задают цепи Маркова на пространстве кандидатов. Теперь мы переходим к вопросу о склеивании спаривающей цепи Маркова.

Определение 3.1. *Состояние вида* $\langle A, B\rangle = \langle A, B\rangle$ *спаривающей цепи Маркова для совпадающей пары кандидатов называется* ***эргодическим****. Состояние вида* $\langle A_1, B_1\rangle < \langle A_2, B_2\rangle$ *называется* ***невозвратным****.*

Теперь мы имеем классическую теорему о сходимости с вероятностью единица:

Теорема 3.2. *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 $t \to \infty$*.*

□

В статье автора [9] приводится доказательство этого результата, но это - классический результат в теории цепей Маркова [36], поэтому здесь мы не будем приводить его доказательство.

Теперь, соединяя вместе алгоритм 4 и теорему 3.2, видим, что с вероятностью единица алгоритм спаривающей цепи Маркова остановится.

Хотя вопрос о среднем времени работы алгоритма 4 остался открытым, в параграфе 3.3 получены теорема 3.4 о среднем времени склеивания спаривающей цепи Маркова и теорема 3.5 о сильной концентрации около этого среднего для случая Булеана.

В качестве практического средства для устранения наиболее длинных траекторий возможно применение следующей техники остановки алгоритма 4 (или его ленивого варианта в алгоритме 8) и запуска его заново:

Определение 3.2. *Если $T_1, \ldots, T_r$ – независимые целочисленные случайные величины, имеющие распределение времени склеивания T, то* ***верхняя граница склеивания*** *по r испытаниям определяется как $\hat{T} = T_1 + \cdots + T_r$.*

На практике предлагается сделать r прогонов спаривающей цепи Маркова и взять оценку $t_1 + \ldots + t_r$ верхней границы склеивания.

Оценим, как изменяются вероятности попадания в эргодические состояния при остановке спаривающей цепи Маркова по r прогонам.

Определение 3.3. *Для целочисленной случайной величины $\hat{T}$, независимой от целочисленной случайной величины T,* ***условное распределение*** *состояний относительно события $B = \{T \leq \hat{T}\}$ есть распределение*

$$\mu_{\hat{T},i} = \frac{\boldsymbol{P}[X_T = i, T \leq \hat{T}]}{\boldsymbol{P}[T \leq \hat{T}]}$$

для любого эргодического состояния i.

Определение 3.4. *Расстояние* ***тотального изменения*** *между распределениями вероятностей $\mu = (\mu_i)_{i \in U}$ и $\nu = (\nu_i)_{i \in U}$ на конечном пространстве U определяется правилом: $\|\mu - \nu\|_{TV} = \frac{1}{2} \cdot \sum_{i \in U} |\mu_i - \nu_i|$.*

Это расстояние является половиной метрики l_1, следовательно, само является метрикой (в частности, симметрично).

Известна классическая и доказываемая прямо из определения 3.4

Лемма 3.1. $\|\mu - \nu\|_{TV} = \boldsymbol{max}_{R \subseteq U} |\mu(R) - \nu(R)|$.

□

В лемме 3.1 подмножество R, на котором достигается максимум, определяется так: $R = \{i \in U \mid \mu_i > \nu_i\}$.

Следующая лемма является технической:

Лемма 3.2. $\|\mu - \mu_{\hat{T}}\|_{TV} \leq \frac{\boldsymbol{P}[T > \hat{T}]}{1 - \boldsymbol{P}[T > \hat{T}]}$, *где $\mu_{\hat{T}}$ - распределение остановленной на верхней границе $\hat{T}$ склеивания по $r > 1$ испытаниям, а μ - распределение выдачи неостановленной цепи.*

Доказательство. По определению 3.3 $\mu_{\hat{T},i} = \frac{\boldsymbol{P}[X_T=i, T\leq\hat{T}]}{\boldsymbol{P}[T\leq\hat{T}]}$. Тогда

$$\boldsymbol{P}[T \leq \hat{T}] \cdot \left(\mu_{\hat{T},i} - \mu_i\right) = \boldsymbol{P}[X_T = i, T \leq \hat{T}] - \boldsymbol{P}[T \leq \hat{T}] \cdot \mu_i =$$
$$= \boldsymbol{P}[T > \hat{T}] \cdot \mu_i - \boldsymbol{P}[X_T = i, T > \hat{T}] \leq \boldsymbol{P}[T > \hat{T}] \cdot \mu_i.$$

Суммируя по множеству $R = \{i \in U \mid \mu_i > \mu_{\hat{T},i}\}$, получим

$$\boldsymbol{P}[T \leq \hat{T}] \cdot \|\mu - \mu_{\hat{T}}\|_{TV} \leq \boldsymbol{P}[T > \hat{T}],$$

что и приводит к утверждению леммы. □

Теперь докажем основную лемму:

Лемма 3.3. $\|\mu - \mu_{\hat{T}}\|_{TV} \leq \frac{1}{2^r-1}$, *где* $\mu_{\hat{T}}$ - *распределение остановленной на верхней границе склеивания по* $r > 1$ *испытаниям, а* μ - *распределение выдачи неостановленной цепи.*

Доказательство. Из-за леммы 3.2 достаточно доказать, что

$$\boldsymbol{P}[T > \hat{T}] \leq 2^{-r}.$$

Из определения $T, T_1, \ldots, T_r$ как независимых одинаково распределенных случайных величин, следует, что $\boldsymbol{P}[T > T_j] \leq \frac{1}{2}$ для всех $1 \leq j \leq r$.

Докажем субмультипликативность:

$$\boldsymbol{P}[T > \sum_{j=1}^{k} T_j] \leq \boldsymbol{P}[T > \sum_{j=1}^{k-1} T_j] \cdot \boldsymbol{P}[T > T_k]$$

для всех $1 < k \leq r$.

Но это следует из формулы условной вероятности, так как если

$$T > \sum_{j=1}^{k-1} T_j,$$

то

$$\langle A_1(\sum_{j=1}^{k-1} T_j), B_1(\sum_{j=1}^{k-1} T_j)\rangle < \langle A_2(\sum_{j=1}^{k-1} T_j), B_2(\sum_{j=1}^{k-1} T_j)\rangle.$$

Поэтому, применяя ко всем четырем кандидатам

$$Min \le \langle A_1(0+\sum_{j=1}^{k-1} T_j), B_1(0+\sum_{j=1}^{k-1} T_j)\rangle < \\ < \langle A_2(0+\sum_{j=1}^{k-1} T_j), B_2(0+\sum_{j=1}^{k-1} T_j)\rangle \le Max$$

одинаковые операции $CbODown$ и $CbOUp$, имеем, что если

$$\langle A_1(t+\sum_{j=1}^{k-1} T_j), B_1(t+\sum_{j=1}^{k-1} T_j)\rangle < \langle A_2(t+\sum_{j=1}^{k-1} T_j), B_2(t+\sum_{j=1}^{k-1} T_j)\rangle$$

склеивается позднее момента $T_k + \sum_{j=1}^{k-1} T_j = \sum_{j=1}^{k} T_j$, то и склеивание $Min < Max$ (из-за транзитивности порядка) совершается позднее момента T_k, то есть

$$\boldsymbol{P}\left[T > \sum_{j=1}^{k} T_j | T > \sum_{j=1}^{k-1} T_j\right] \le \boldsymbol{P}[T > T_k].$$

Теперь результат леммы следует по индукции. □

Соединяя результаты лемм 3.1 и 3.3, получим

Теорема 3.3. *Для любого $R \subseteq U$ с $\mu(R) = \rho$ и $r > log_2(\rho+1) - log_2(\rho)$ имеем $\mu_{\hat{T}}(R) \ge \rho - \frac{1}{2^r-1}$ для верхней границы $\hat{T}$ склеивания по $r > 1$ испытаниям.*

Доказательство.

$$\rho - \frac{1}{2^r-1} \le \mu(R) - \|\mu - \mu_{\hat{T}}\|_{TV} = \\ = \mu(R) - \mathbf{max}_{Q\subseteq U}|\mu(Q) - \mu_{\hat{T}}(Q)| \le \\ \le \mu(R) - |\mu(R) - \mu_{\hat{T}}(R)| \le \mu_{\hat{T}}(R).$$

□

3.2 Свойства цепей Маркова

Этот параграф воспроизводит наблюдения и результаты из статьи автора [9]. Многие свойства цепей Маркова, получаемых из алгоритмов 2, 3 и 4, станут более понятными.

Рассмотрим формальный контекст:

$O \mid F$	f_1	f_2	f_3	f_4
o_1	1	1	0	0
o_2	1	0	1	0
o_3	0	0	0	1

Решетка кандидатов содержит 6 элементов с фрагментами

$$\emptyset, \{f_1\}, \{f_4\}, \{f_1, f_2\}, \{f_1, f_3\}, \{f_1, f_2, f_3, f_4\},$$

упорядоченными по включению.

Выпишем стохастическую матрицу для алгоритма 2, примененного к данному формальному контексту, (порядок строк и столбцов соотвествует перечислению кандидатов из предыдущего абзаца):

$$\begin{Vmatrix} 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 \\ 1/4 & 0 & 0 & 1/4 & 1/4 & 1/4 \\ 2/5 & 0 & 0 & 0 & 0 & 3/5 \\ 1/4 & 1/4 & 0 & 0 & 0 & 2/4 \\ 1/4 & 1/4 & 0 & 0 & 0 & 2/4 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 \end{Vmatrix}$$

Заметим, что последний элемент (1/4) второй строки соответствует переходу из кандидата $\langle\{o_1, o_2\}, \{f_1\}\rangle$ в кандидата $\langle\emptyset, \{f_1, f_2, f_3, f_4\}\rangle$, получающемуся при выборе признака f_4. Обратный переход из кандидата $\langle\emptyset, \{f_1, f_2, f_3, f_4\}\rangle$ в $\langle\{o_1, o_2\}, \{f_1\}\rangle$ невозможен (0 на втором месте последней строки), так как при выборе признака мы остаемся на месте, а при выборе объекта добавится только один этот объект, тогда как у $\langle\{o_1, o_2\}, \{f_1\}\rangle$ имеется 2 родителя.

Такая ситуация препятствует выполнению *тождества баланса*:

$$\pi_i \cdot (P)_{i,j} = \pi_j \cdot (P)_{j,i}, \tag{3.1}$$

так как иначе для $i = \langle \emptyset, \{f_1, f_2, f_3, f_4\} \rangle$ и $j = \langle \{o_1, o_2\}, \{f_1\} \rangle$ получаем, что должно быть $\pi_j = 0$, так как $(P)_{i,j} = 0$ и $(P)_{j,i} \neq 0$.

Поэтому цепь Маркова из алгоритма 2 не является *обратимой*.

Обсудим теперь название алгоритма 2. Нам потребуется два определения:

Определение 3.5. *Подмножество $J \subseteq S$ упорядоченного множества $\langle S, \leq \rangle$ называется* ***порядковым идеалом****, если выполняется следующие условие*

$$\forall l, k \in S \left[l \leq k \wedge k \in J \Rightarrow l \in J \right].$$

Примером порядкового идеала для решетки кандидатов является $J = \{\langle \{o_1, o_2, o_3\}, \emptyset \rangle, \langle \{o_3\}, \{f_4\} \rangle\}$.

Определение 3.6. *Цепь Маркова с упорядоченного пространством состояний $\langle S, \leq \rangle$ называется* ***монотонной****, если для любых $i \leq j \in S$ и любого порядкового идеала $J \subseteq S$ выполняется неравенство*

$$\boldsymbol{P}\left[X_{t+1} \in J \mid X_t = i\right] \geq \boldsymbol{P}\left[X_{t+1} \in J \mid X_t = j\right].$$

Используя формулу $\mathbf{P}\left[X_{t+1} \in J \mid X_t = i\right] = \sum_{k \in J} (P)_{i,k}$, для $i = \langle \{o_1\}, \{f_1, f_2\} \rangle$ и порядкового идеала $J = \{\langle \{o_1, o_2, o_3\}, \emptyset \rangle, \langle \{o_3\}, \{f_4\} \rangle\}$ получаем для алгоритма 2 $\mathbf{P}\left[X_{t+1} \in J \mid X_t = i\right] = \frac{1}{4}$. Аналогично, для $j = \langle \emptyset, \{f_1, f_2, f_3, f_4\} \rangle$ имеем в алгоритме 2 $\mathbf{P}\left[X_{t+1} \in J \mid X_t = j\right] = \frac{1}{3}$. Но это нарушает условие из определения 3.6, поэтому алгоритм 2 задает *немонотонную* цепь Маркова.

Выпишем стохастическую матрицу для алгоритма 3 (порядок строк и столбцов определяется перечислением кандидатов с фрагментами $\emptyset, \{f_1\}, \{f_4\}, \{f_1, f_2\}, \{f_1, f_3\}, \{f_1, f_2, f_3, f_4\}$):

$$\begin{Vmatrix} 3/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 3/7 & 0 & 1/7 & 1/7 & 1/7 \\ 2/7 & 0 & 2/7 & 0 & 0 & 3/7 \\ 1/7 & 1/7 & 0 & 3/7 & 0 & 2/7 \\ 1/7 & 1/7 & 0 & 0 & 3/7 & 2/7 \\ 0 & 0 & 1/7 & 1/7 & 1/7 & 4/7 \end{Vmatrix}$$

Заметим, что в этом случае переход из кандидата $\langle\{o_1, o_2\}, \{f_1\}\rangle$ в кандидат $\langle\emptyset, \{f_1, f_2, f_3, f_4\}\rangle$ возможен, а обратный переход нет.

Поэтому цепь Маркова из алгоритма 3 тоже не является обратимой.

Напомним, что *стационарным* называется такое распределение вероятностей π на пространстве состояний S цепи Маркова, что выполняется соотношение:

$$\forall i \in S \left[\pi_i = \sum_{j \in S} \pi_j \cdot (P)_{j,i} \right]. \tag{3.2}$$

Отождествляя распределение вероятностей со строкой чисел $\pi = (\pi_i)_{i \in S}$, можно записать это в матричном виде: $\pi \cdot P = \pi$, где $(P)_{i,j} = \mathbf{P}\left[X_{t+1} = j \mid X_t = i\right]$ - компоненты матрицы переходов цепи Маркова.

Вычисление собственного вектора транспонированной матрицы переходов для цепи Маркова из алгоритма 3 с собственным значением 1 дает стационарное распределение $(\frac{78}{512}, \frac{58}{512}, \frac{50}{512}, \frac{77}{512}, \frac{77}{512}, \frac{172}{512})$, которое не является равномерным.

Установим теперь монотонность цепи Маркова, определяемого алгоритмом 3, для произвольного формального контекста:

Лемма 3.4. *Пусть J - порядковый идеал в решетке кандидатов, и пусть $i = \langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle = j$ - возрастающая пара кандидатов. Тогда $\boldsymbol{P}\left[X_{t+1} \in J \mid X_t = i\right] \geq \boldsymbol{P}\left[X_{t+1} \in J \mid X_t = j\right]$ для цепи Маркова из алгоритма 3.*

Доказательство. В алгоритме 3 равновероятно выбирается элемент $O \cup F$. Если выбран $o \in O$, то по лемме 1.3 имеем

$$CbODown(\langle A_1, B_1\rangle, o) \leq CbODown(\langle A_2, B_2\rangle, o).$$

По определению 3.5 из

$$CbODown(\langle A_2, B_2\rangle, o) \in J$$

следует

$$CbODown(\langle A_1, B_1\rangle, o) \in J.$$

Если выбран $f \in F$, то опять по лемме 1.3 имеем

$$CbOUp(\langle A_1, B_1\rangle, f) \leq CbOUp(\langle A_2, B_2\rangle, f).$$

Снова определение 3.5 позволяет из

$$CbOUp(\langle A_2, B_2\rangle, f) \in J$$

вывести

$$CbOUp(\langle A_1, B_1\rangle, f) \in J.$$

Поэтому имеем включение событий

$$[X_{t+1} \in J \mid X_t = j] \subseteq [X_{t+1} \in J \mid X_t = i],$$

то есть требуемое соотношение

$$\mathbf{P}[X_{t+1} \in J \mid X_t = i] \geq \mathbf{P}[X_{t+1} \in J \mid X_t = j].$$

□

Это дает нам основание назвать цепь Маркова из алгоритма 3 *монотонной*.

Теперь перейдем к исследованию спаривающейся цепи Маркова из алгоритма 4.

Сначала перечислим все упорядоченные пары фрагментов кандидатов для формального контекста, заданного в начале настоящего параграфа:

1	$\emptyset \subseteq \emptyset$
2	$\{f_1\} \subseteq \{f_1\}$
3	$\{f_4\} \subseteq \{f_4\}$
4	$\{f_1, f_2\} \subseteq \{f_1, f_2\}$
5	$\{f_1, f_3\} \subseteq \{f_1, f_3\}$
6	$\{f_1, f_2, f_3, f_4\} \subseteq \{f_1, f_2, f_3, f_4\}$
7	$\emptyset \subseteq \{f_4\}$
8	$\emptyset \subseteq \{f_1\}$
9	$\{f_1\} \subseteq \{f_1, f_2\}$
10	$\{f_1\} \subseteq \{f_1, f_3\}$
11	$\emptyset \subseteq \{f_1, f_2\}$
12	$\emptyset \subseteq \{f_1, f_3\}$
13	$\{f_4\} \subseteq \{f_1, f_2, f_3, f_4\}$
14	$\{f_1, f_2\} \subseteq \{f_1, f_2, f_3, f_4\}$
15	$\{f_1, f_3\} \subseteq \{f_1, f_2, f_3, f_4\}$
16	$\{f_1\} \subseteq \{f_1, f_2, f_3, f_4\}$
17	$\emptyset \subseteq \{f_1, f_2, f_3, f_4\}$

Тогда матрица перехода для цепи Маркова из алгоритма 4 имеет блочный вид

$$\left\| \begin{array}{c|c} S & \mathbf{O} \\ \hline R & Q \end{array} \right\|$$

Блок S в точности совпадает с 6×6-матрицей переходов цепи Маркова для алгоритма 3. 6×11-матрица $\mathbf{O}$ состоит из одних нулей. 11×6-матрица R имеет вид:

$$\left\|\begin{array}{cccccc} 2/7 & 0 & 1/7 & 0 & 0 & 0 \\ 1/7 & 1/7 & 0 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 0 & 1/7 & 0 & 1/7 \\ 1/7 & 1/7 & 0 & 0 & 1/7 & 1/7 \\ 1/7 & 0 & 0 & 1/7 & 0 & 0 \\ 1/7 & 0 & 0 & 0 & 1/7 & 0 \\ 0 & 0 & 1/7 & 0 & 0 & 3/7 \\ 0 & 0 & 0 & 1/7 & 0 & 2/7 \\ 0 & 0 & 0 & 0 & 1/7 & 2/7 \\ 0 & 0 & 0 & 0 & 0 & 1/7 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right\|$$

Наконец, 11×11-матрица Q имеет вид:

$$\left\|\begin{array}{ccccccccccc} 1/7 & 0 & 0 & 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 0 \\ 0 & 2/7 & 0 & 0 & 0 & 0 & 1/7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2/7 & 0 & 0 & 0 & 0 & 0 & 1/7 & 0 & 0 \\ 0 & 0 & 0 & 2/7 & 0 & 0 & 0 & 1/7 & 0 & 0 & 0 \\ 0 & 1/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 & 0 \\ 0 & 1/7 & 0 & 1/7 & 0 & 1/7 & 1/7 & 1/7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 0 & 0 & 0 & 0 \\ 1/7 & 0 & 0 & 1/7 & 0 & 0 & 0 & 2/7 & 0 & 0 & 0 \\ 1/7 & 0 & 1/7 & 0 & 0 & 0 & 0 & 0 & 2/7 & 0 & 0 \\ 1/7 & 0 & 1/7 & 1/7 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 0 & 0 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \end{array}\right\|$$

Матрица S в качестве левого верхнего блока возникает из-за того, что при ограничении *эргодическими* состояниями $\langle A, B\rangle \leq \langle A, B\rangle$ алгоритм 4 сводится к алгоритму 3.

Из предыдущих рассуждений следует, что спаривающая цепь Маркова, в общем случае, не является обратимой и стационарное распределение не является равномерным.

То обстоятельство, что вероятность для спаривающейся цепи оказаться на шаге $t \to \infty$ в каком-то *невозвратном* состоянии $\langle A_1, B_1\rangle < \langle A_2, B_2\rangle$ (теорема 3.2) находит свое проявление в предельном соотношении $\lim\limits_{t\to\infty} Q^t = \mathbf{O}$, где $\mathbf{O}$ - квадратная матрица, состоящая из одних нулей.

Заметим, что в некоторых других случаях (например, для формального контекста из уравнения (1.21), определяющего Булеан) алгоритмы 2 и 3, могут задавать обратимые цепи Маркова с равномерным стационарным распределением. Подробное доказательство этого факта, можно найти в параграфе 3.5.

При этом цепь из алгоритма 2 будет монотонной. Фактически, для Булеана эти цепи Маркова совпадают со случайным и ленивым случайным блужданиями по вершинам Булеана, соответственно.

3.3 Скорость склеивания спаривающей цепи: случай Булеана

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.

Мы воспроизведем здесь эти результаты из статьи автора [15], так как они могут служить дополнительным доводом о том, что предложенный подход эффективен с вычислительной точки зрения.

До конца этого параграфа мы ограничимся случаем Булеана (уравнение (1.21) параграфа 1.2 задает формальный контекст, определяющий Булеан).

Определение 3.7. ***Расстояние*** $\rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle)$ *между кандидатами* $\langle A_1, B_1\rangle$ *и* $\langle A_2, B_2\rangle$ *определяется как число позиций, в которых отличаются битовые строки* B_1 *и* B_2*. Другими словами, расстояние равно минимальному количеству ребер между соответствующими вершинами гиперкуба.*

Это расстояние на Булеане называется *метрикой Хэмминга.*

Оказывается, что после применения операций «замыкай-по-одному» в случае Булеана, это расстояние не увеличивается. Точнее,

Лемма 3.5. *Пусть* $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$. *Тогда*

$$\rho(CbOUp(\langle A_1, B_1\rangle, f), CbOUp(\langle A_2, B_2\rangle, f)) = \\ = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle), \text{ если } f \in B_1 \wedge f \in B_2.$$

$$\rho(CbOUp(\langle A_1, B_1\rangle, f), CbOUp(\langle A_2, B_2\rangle, f)) = \\ = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle) - 1, \text{ если } f \notin B_1 \wedge f \in B_2.$$

$$\rho(CbOUp(\langle A_1, B_1\rangle, f), CbOUp(\langle A_2, B_2\rangle, f)) = \\ = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle), \text{ если } f \notin B_1 \wedge f \notin B_2.$$

$$\rho(CbODown(\langle A_1, B_1\rangle, o), CbODown(\langle A_2, B_2\rangle, o)) = \\ = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle), \text{ если } o \in A_1 \wedge o \in A_2.$$

$$\rho(CbODown(\langle A_1, B_1\rangle, o), CbODown(\langle A_2, B_2\rangle, o)) = \\ = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle) - 1, \text{ если } o \in A_1 \wedge o \notin A_2.$$

$$\rho(CbODown(\langle A_1, B_1\rangle, o), CbODown(\langle A_2, B_2\rangle, o)) = \\ = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle), \text{ если } o \notin A_1 \wedge o \notin A_2.$$

□

Заметим, что в общем случае это расстояние может увеличиться. Рассмотрим опять формальный контекст из начала параграфа 3.2. Здесь $\rho(\langle\{o_1, o_2, o_3\}, \emptyset\rangle, \langle\{o_3\}, \{f_4\}\rangle) = 1$, но

$$\rho(CbOUp(\langle\{o_1, o_2, o_3\}, \emptyset\rangle, f_1), CbOUp(\langle\{o_3\}, \{f_4\}\rangle, f_1)) = \\ = \rho(\langle\{o_1, o_2\}, \{f_1\}\rangle, \langle\emptyset, \{f_1, f_2, f_3, f_4\}\rangle) = 3. \quad (3.3)$$

Нам теперь понадобится один известный класс распределений вероятностей [55]:

Определение 3.8. ***Геометрическим*** *распределением вероятностей называется целочисленная случайная величина* T *с* $\boldsymbol{P}[T = k] = (1-p)^{k-1} \cdot p$, *где* $0 < p \leq 1$ *и* $k \geq 1$.

Лемма 3.6. *Время* T_n *ожидания уменьшения расстояния с*

$$\rho(\langle O, \emptyset\rangle, \langle\emptyset, F\rangle) = n$$

до $n-1$ *имеет геометрическое распределение вероятностей с* $p = 1$. *Время* T_j *ожидания уменьшения расстояния с* j *до* $j-1$ *имеет геометрическое распределение вероятностей с* $p = \frac{j}{n}$.

Доказательство следует из предыдущей леммы и разбора случаев выбора для следующего замыкаемого объекта или признака. □

Сформулируем и докажем теперь классическую лемму:

Лемма 3.7. *Для целочисленной случайной величины T с геометрическим распределением вероятностей выполнено $\boldsymbol{E}[T] = 1/p$ и $\boldsymbol{D}[X] = (1-p)/p^2$.*

Доказательство. Воспользуемся производящей функцией

$$\psi_T(z) = \boldsymbol{E}[z^T] = \frac{p \cdot z}{1 - (1-p) \cdot z}.$$

Имеем

$$\boldsymbol{E}[T] = \psi'_T(1) = 1/p$$

и

$$\boldsymbol{D}[T] = \psi''_T(1) + \psi'_T(1) - (\psi'_T(1))^2 = (1-p)/p^2.$$

□

Теорема 3.4. *Среднее время склеивания для n-мерного гиперкуба равно*

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

Доказательство. Из-за линейности среднего имеем

$$\boldsymbol{E}[\sum_{j=1}^{n} T_j] = \sum_{j=1}^{n} \boldsymbol{E}[T_j].$$

По лемме 3.7 имеем $\boldsymbol{E}[T_j] = \frac{n}{j}$. □

Вспомним следующую классическую лемму П.Л.Чебышева:

Лемма 3.8. $\boldsymbol{P}[\mid T - \boldsymbol{E}[T] \mid \geq \varepsilon] \leq \boldsymbol{D}[T]/\varepsilon^2$. □

Теорема 3.5. $\boldsymbol{P}\left[\sum_{j=1}^{n} T_j \geq (1+\varepsilon) \cdot n \cdot \ln(n)\right] \to 0$ *при* $n \to \infty$ *для любого* $\varepsilon > 0$.

Доказательство. Из-за независимости T_j имеем

$$\boldsymbol{D}[\sum_{j=1}^{n} T_j] = \sum_{j=1}^{n} \boldsymbol{D}[T_j].$$

По лемме 3.7 имеем $\boldsymbol{D}[T_j] \approx n^2/j^2$. Поэтому $\sum_{j=1}^{n} \boldsymbol{D}[T_j] = O(n^2)$. Для $T = \sum_{j=1}^{n} T_j$ по лемме 3.8 при достаточно больших n имеем

$$\boldsymbol{P}[T \geq (1+\varepsilon) \cdot n \cdot \ln(n)] \leq$$
$$\leq \boldsymbol{P}[\mid T - \boldsymbol{E}[T] \mid \geq \varepsilon \cdot n \cdot \ln(n)] \leq \frac{\boldsymbol{D}[T]}{\varepsilon^2 \cdot n^2 \cdot \ln^2(n)}.$$

При $n \to \infty$, имеем требуемый результат. □

Хотелось бы обратить внимание читателя на следующие замечательные обстоятельства:

1. Для 32-мерного гиперкуба среднее время склеивания

$$\boldsymbol{E}[T] = 32 \cdot \sum_{j=1}^{32} \frac{1}{j} \leq 130.$$

   Чтобы выбрать случайное подмножество из 32 признаков, нужно использовать 32 раза датчик случайных чисел, так что наша оценка не сильно (только логарифмически) хуже.

2. Но в 32-мерном гиперкубе 4294967296 сходств, подавляющее большинство которых не будет вычисляться в процессе вероятностного поиска сходств.

3. Эксперименты показывают, что и в общем случае время склеивания мало по сравнению с числом различных кандидатов.

3.4 Скорость перемешивания: постановка задачи

Как было указано в параграфе 3.2 в ВКФ-методе мы используем спаривающую цепь Маркова, порождаемую алгоритмом 4, которая имеет очевидное условие остановки, тогда как алгоритмы 2 и 3 имеют проблему определения числа T, сколько раз должен выполняться шаг соответствующей цепи Маркова.

В западной литературе [63, 68–70, 72, 74] вопрос об остановке цепи Маркова активно исследовался, начиная с начала 80-х годов прошлого века. Однако были получены лишь частичные результаты. Особенно хорошо поддаются исследованию случаи обратимых цепей Маркова, порождаемых сверткой распределений вероятностей на группах, то есть *случайные блуждания на группах*.

Нам эти результаты могут быть полезны, когда для случая Булеана всех подмножеств n-элементного множества, мы получаем случайное блуждание на группе $\mathbb{Z}_2^n$. Как было указано в параграфе 3.1 алгоритмы 2 и 3 задают случайное и ленивое случайное блуждание на Булеане, соответственно.

Вычислительные эксперименты, проведенные Е.Ю.Сидоровой [54] в рамках дипломной работы, показали, что из-за большей сложности внутреннего цикла в алгоритме 2, он уступает алгоритму 3 более, чем в 2 раза, что нивелирует его чуть более высокую скорость перемешивания (в случае Булеана ровно в 2 раза).

В параграфе 3.5 мы исследуем время перемешивания монотонной цепи Маркова из алгоритма 3 для Булеана в надежде на продвижение исследования алгоритма 3 для самых общих формальных контекстов.

Следует отметить, что для случая спаривающей цепи Маркова, порождаемой алгоритмом 4, подобные результаты о времени перемешивания будут полезны, так как после склеивания алгоритм 4 совпадает с алгоритмом 3. Таким образом, продолжая вычисления, можно добиться близости в метрике тотального изменения к стационарному распределению.

Впрочем, для случая Булеана это оказывается излишним: распределение состояний склеивания алгоритма 4 уже является равномерным (=стационарным).

Основным результатом теории конечных цепей Маркова является эргодическая теорема 3.6.

Цепь Маркова, определяемая стохастической матрицей P переходов, называется *неприводимой*, если любые два ее состояния достижимы друг из друга с положительной вероятностью:

$$\forall i, j \in S \exists t \left[(P^t)_{i,j} > 0 \right]. \tag{3.4}$$

Цепь Маркова называется *апериодичной*, если наибольший общий делитель шагов возвращения в исходное состояние равен 1:

$$\forall i \in S \left[\gcd\{t > 0 | (P^t)_{i,i} > 0\} = 1 \right]. \tag{3.5}$$

Стационарным называется такое распределение вероятностей π на пространстве состояний S цепи Маркова с матрицей переходов P, что выполняется соотношение:

$$\forall i \in S \left[\pi_i = \sum_{j \in S} \pi_j \cdot (P)_{j,i} \right]. \tag{3.6}$$

Теорема 3.6. *Для неприводимой апериодичной цепи Маркова с матрицей переходов P существует единственное стационарное распределение π. Более того*

$$\forall i, j \in S \left[\lim_{t \to \infty} (P^t)_{i,j} = \pi_j \right].$$

Другими словами, при неограниченном возрастании $t \to \infty$ все строки матрицы P^t переходов за t шагов будут сходиться к стационарному распределению π, рассматриваемого как вектор-строка.

Напомним, что метрика *тотального изменения*, играющая в дальнейшем изложении ключевую роль, имеет вид:

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \cdot \sum_{i \in S} |\mu_i - \nu_i|. \tag{3.7}$$

Полезна следующая лемма, легко выводимая из тождества (3.7):

Лемма 3.9. $\|\mu - \nu\|_{TV} = \frac{1}{2} \cdot \boldsymbol{max}_{h:S\to[-1,1]} \sum_{i\in S} h(i) \cdot (\mu_i - \nu_i)$.

С использованием леммы 3.9 из теоремы 3.6 легко доказать:

Теорема 3.7. *Для любого начального распределения $\mu^{(0)}$ на состояниях неприводимой апериодичной цепи Маркова с матрицей переходов P имеем*

$$\lim_{t\to\infty} \|\mu^{(0)} \cdot P^t - \pi\|_{TV} = 0.$$

Заметим, что $\mu^{(0)} \cdot P^t$ будет распределением вероятностей состояний цепи Маркова после t шагов с начальным распределением $\mu^{(0)}$.

Рассмотрим начальные распределения δ_j, *сконцентрированные* в состоянии $j \in S$ с $(\delta_j)_i = \delta_{j,i}$ (где $\delta_{j,i}$ - символ Дирака).

Определение 3.9. *Для заданного порога $\varepsilon > 0$* ***временем перемешивания*** *называется такое минимальное целое число $T = \tau(\varepsilon)$, что для любого стартового состояния $j \in S$ выполняется*

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

3.5 Скорость перемешивания: частичные результаты

В этом параграфе мы докажем вариант результата Перси Дьякониса [63] о сильной концентрации времени перемешивания для монотонной цепи Маркова, порождаемой алгоритмом 3, примененного к формальному контексту Булеана (уравнение (1.21)). Замечательным фактом является то обстоятельство, что немонотонная цепь Маркова для Булеана совпадает с классической моделью Пауля и Татьяны Эренфестов [64] из статистической механики, предложенной ими в начале 20 века.

Модель Эренфестов [55] состоит из двух урн с номерами 0 и 1 и n шаров, пронумерованных числами $1, 2, \ldots, n$.

Конфигурацией называется размещение шаров по урнам. Ясно, что имеется 2^n конфигураций, однозначно соответствующих подмножествам B шаров, находящихся в урне с номером 1.

Начальная конфигурация - та, в которой все шары находятся в урне с номером 0, что соответствует подмножеству $B_0 = \emptyset$.

На каждом шаге t равновероятно выбирается один из шаров $f \in \{1, 2, \ldots, n\}$ любой из урн, который перекладывается в другую урну. Пусть в момент времени t цепь Маркова находится в конфигурации, соответствующей подмножеству B_t. Если $f \notin B_t$, то цепь переходит в состояние, соответствующее подмножеству $B_{t+1} = B_t \cup \{f\}$. Если $f \in B_t$, то цепь переходит в состояние $B_{t+1} = B_t \setminus \{f\}$.

Эта процедура (*диффузия Эренфестов*) полностью совпадает с алгоритмом 2 немонотонной цепи Маркова из параграфа 3.1, если там следить только за второй компонентой (*фрагментом*) ВКФ-кандидата.

К сожалению, это правило задает периодическую цепь Маркова (с периодом 2). Это легко увидеть, если рассмотреть двудольный граф Булеана всех подмножеств и заметить, что каждый шаг переводит состояние из одной доли в состояния другой доли.

Формально, мы имеем $\gcd\{t > 0 | (P^t)_{j,j} > 0\} = 2$ для любого состояния $j \in S$, что противоречит условию (3.5).

Для получения апериодической цепи достаточно с вероятностью $\frac{1}{2}$ не изменять текущую конфигурацию и с вероятностью $\frac{1}{2n}$ (вместо $\frac{1}{n}$) выбирать перекладываемый шар.

Легко проверить, что так модифицированная диффузия Эренфестов полностью совпадает с алгоритмом 3 монотонной цепи Маркова из параграфа 3.1, если там следить только за второй компонентой (*фрагментом*) кандидата.

Для модифицированной диффузии Эренфестов и будет оцениваться время перемешивания из определения 3.9. В дальнейшем изложении прилагательное «модифицированная» будет опускаться.

Так как $(P)_{i,i} = \frac{1}{2} > 0$, то модель Эренфестов задает апериодичную цепь Маркова.

Легко убедиться в том, что (модифицированная) модель Эренфестов задает неприводимую цепь Маркова. Для проверки условия (3.4) достаточно взять $t = n$: если конфигурация $i \in S$ соответствует

подмножеству B_1, а состояние $j \in S$ - подмножеству B_2, то

$$(P^n)_{i,j} \geq (P)_{i,i(1)j} \cdot (P)_{i(1)j,i(2)j} \cdot \ldots \cdot (P)_{i(n-1)j,j} \geq \left(\frac{1}{2n}\right)^n > 0,$$

где $i(k)j$ соответствует $\{f \in B_2 \mid f \leq k\} \cup \{f \in B_1 \mid f > k\}$.

Другими словами, мы можем перекладывать шары по порядку так, чтобы конфигурация i переходила в конфигурацию j. Если на k-ом шаре конфигурации не различаются, то мы делаем тождественный переход.

Тот факт, что стационарным распределением для модели Эренфестов является равномерное, легче всего установить через обратимость.

Напомним, что цепь Маркова называется *обратимой*, если выполняются *тождества баланса*:

$$\pi_i \cdot (P)_{i,j} = \pi_j \cdot (P)_{j,i}, \tag{3.8}$$

для некоторого распределения вероятностей π.

Легко проверить, что модель Эренфестов обратима с балансом относительно равномерного распределения: $(P)_{i,j} = 0 = (P)_{j,i}$, если i и j отличаются более чем по одному шару; $(P)_{i,j} = \frac{1}{2n} = (P)_{j,i}$, если i и j отличаются ровно по одному шару; случай $i = j$ тривиален.

Теперь осталось применить следующую лемму:

Лемма 3.10. *Если неприводимая апериодичная цепь Маркова удовлетворяет тождеству баланса относительно распределения π, то π является ее стационарным распределением.*

Доказательство.

$$\sum_{j \in S} \pi_j \cdot (P)_{j,i} = \sum_{j \in S} \pi_i \cdot (P)_{i,j} = \pi_i \cdot \sum_{j \in S} (P)_{i,j} = \pi_i.$$

□

Итак, мы установили, что стационарное распределение модели Эренфестов является равномерным.

В случае Булеана всех подмножеств n-элементного универсума F мы имеем пространство состояний $\{0,1\}^n$, которое мы отождествим с группой $\mathbb{Z}_2^n$ - прямым произведением циклических групп $\mathbb{Z}_2$.

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

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

Рассмотрим теперь распределение вероятностей μ, задаваемое на $S = \mathbb{Z}_2^n$ равенствами

$$\begin{aligned} \mu(0) &= \frac{1}{2}; \\ \mu(e_1) = \ldots &= \mu(e_n) = \tfrac{1}{2n}. \end{aligned} \tag{3.9}$$

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

Ясно, что для модели Эренфестов выполнено $(P)_{i,j} = \mu((-i)+j)$, где конфигурации i и j понимаются как элементы абелевой группы $\mathbb{Z}_2^n$.

Как обычно, *свертка* распределения вероятностей μ с распределением вероятностей ν на абелевой группе $G = \langle S; +, 0, - \rangle$ задается равенством

$$\mu * \nu(j) = \sum_{i \in S} \mu(i) \cdot \nu((-i) + j). \tag{3.10}$$

t-кратная свертка распределения вероятностей μ с собой обозначается через

$$\mu^{*t} = \mu * \mu^{*(t-1)}.$$

Теперь с помощью формулы полной вероятности легко доказывается следующая

Лемма 3.11. $(P^t)_{i,j} = \mu^{*t}((-i) + j)$

Для замены свертки произведением необходимо сделать *преобразование Фурье* из теории представления конечных групп [53]. Для группы $\mathbb{Z}_2^n$ оно называется *преобразованием Уолша-Адамара* [27].

Определение 3.10. ***Характером*** *группы* $\mathbb{Z}_2^n$ *называется такое отображение* $\chi : \mathbb{Z}_2^n \to \mathbb{T} = \{z \in \mathbb{C} \mid |z| = 1\}$, *что выполняется*

$$\forall x, y \in \mathbb{Z}_2^n \left[\chi(x + y) = \chi(x) \cdot \chi(y)\right].$$

Вектор весов $w = (w_1, \ldots, w_n) \in \mathbb{Z}_2^n$ задает характер χ_w группы $\mathbb{Z}_2^n$ по правилу:

$$\chi_w(x) = (-1)^{w \cdot x} = (-1)^{\sum_{k=1}^n w_k \cdot x_k}. \tag{3.11}$$

Индукцией по размерности n группы легко устанавливается

Лемма 3.12. *Для любого характера* χ *группы* $\mathbb{Z}_2^n$ *найдется такой вектор* $w = (w_1, \ldots, w_n) \in \mathbb{Z}_2^n$, *что* $\chi = \chi_w$.

Рассмотрим линейное пространство функций $L(\mathbb{Z}_2^n) = \{g : \mathbb{Z}_2^n \to \mathbb{C}\}$ на группе $\mathbb{Z}_2^n$.

Следующие две леммы устанавливают ортогональность характеров (как векторов $L(\mathbb{Z}_2^n)$):

Лемма 3.13.

$$\sum_{x \in \mathbb{Z}_2^n} \chi_w(x) = \begin{cases} 2^n, & \text{если } w = 0 \\ 0, & \text{если } w \neq 0. \end{cases}$$

Доказательство. Если $w = 0$, то $\sum_{x \in \mathbb{Z}_2^n} \chi_0(x) = |\mathbb{Z}_2^n| = 2^n$, так как $\forall x \left[\chi_0(x) = 1\right]$. В противном случае найдется такое $y \in \mathbb{Z}_2^n$, что $\chi_w(y) = -1$. Тогда

$$-\sum_{x \in \mathbb{Z}_2^n} \chi_w(x) = \chi_w(y) \cdot \sum_{x \in \mathbb{Z}_2^n} \chi_w(x) = \sum_{x \in \mathbb{Z}_2^n} \chi_w(y + x) = \sum_{x \in \mathbb{Z}_2^n} \chi_w(x),$$

что доказывает второе равенство. □

Лемма 3.14.

$$\sum_{x \in \mathbb{Z}_2^n} \chi_w(x) \cdot \chi_v(x) = \begin{cases} 2^n, & \text{если } w = v \\ 0, & \text{если } w \neq v. \end{cases}$$

Доказательство. Из равенства (3.11) сразу выводится, что $\chi_w \cdot \chi_v = \chi_{w+v}$. Но равенство $w = v$ в $\mathbb{Z}_2^n$ эквивалентно равенству $w + v = 0$. Теперь результат следует из леммы 3.13. □

Определение 3.11. ***Преобразованием Фурье*** *элемента* $g \in L(\mathbb{Z}_2^n)$ *называется*

$$\hat{g}(w) = \langle g, \chi_w \rangle = \sum_{x \in \mathbb{Z}_2^n} g(x) \cdot \chi_w(x).$$

Преобразование Фурье свертку распределений вероятностей переводит в произведение преобразований Фурье:

Лемма 3.15. *Для любых распределений вероятностей* μ *и* ν *на группе* $\mathbb{Z}_2^n$ *выполняется* $\widehat{\mu * \nu} = \hat{\mu} \cdot \hat{\nu}$.

Доказательство.

$$\begin{aligned}
\widehat{\mu * \nu}(w) &= \sum_{x \in \mathbb{Z}_2^n} (\mu * \nu)(x) \cdot \chi_w(x) = \\
&= \sum_{x \in \mathbb{Z}_2^n} \Big(\sum_{y \in \mathbb{Z}_2^n} \mu(y) \cdot \nu((-y) + x) \Big) \cdot \chi_w(y + ((-y) + x)) = \\
&= \sum_{y \in \mathbb{Z}_2^n} \Big(\sum_{x \in \mathbb{Z}_2^n} \mu(y) \cdot \nu((-y) + x) \cdot \chi_w((-y) + x) \Big) \cdot \chi_w(y) = \\
&= \sum_{y \in \mathbb{Z}_2^n} \mu(y) \cdot \hat{\nu}(w) \cdot \chi_w(y) = \hat{\mu}(w) \cdot \hat{\nu}(w).
\end{aligned}$$

□

Отсюда индукцией по t получаем важное соотношение:

$$\widehat{\mu^{*t}} = \hat{\mu}^t. \tag{3.12}$$

Вычислим коэффициент Фурье $\hat{\nu}(0)$ для распределения вероятностей ν, где $0 \in \mathbb{Z}_2^n$:

Лемма 3.16. *Для любого распределения вероятностей* ν *на группе* $\mathbb{Z}_2^n$ *выполняется* $\hat{\nu}(0) = 1$.

Ясно, что Фурье-образ $\hat{g}$ принадлежит $L(\mathbb{Z}_2^n)$. Имеет место *формула обращения Фурье.*

Лемма 3.17. *Для $g \in L(\mathbb{Z}_2^n)$ имеем*

$$g = \frac{1}{2^n} \cdot \sum_{w \in \mathbb{Z}_2^n} \hat{g}(w) \cdot \chi_w = \frac{1}{2^n} \cdot \sum_{w \in \mathbb{Z}_2^n} \langle g, \chi_w \rangle \cdot \chi_w.$$

Доказательство.

$$\frac{1}{2^n} \cdot \sum_{w \in \mathbb{Z}_2^n} \hat{g}(w) \cdot \chi_w(y) = \frac{1}{2^n} \cdot \sum_{w \in \mathbb{Z}_2^n} \sum_{x \in \mathbb{Z}_2^n} g(x) \cdot \chi_w(x) \cdot \chi_w(y) =$$

$$= \frac{1}{2^n} \cdot \sum_{w \in \mathbb{Z}_2^n} \sum_{x \in \mathbb{Z}_2^n} g(x) \cdot \chi_w(x+y) = \frac{1}{2^n} \cdot \sum_{x \in \mathbb{Z}_2^n} \left(g(x) \cdot \sum_{w \in \mathbb{Z}_2^n} \chi_w(x+y) \right) =$$

$$= \frac{1}{2^n} \cdot \sum_{x \in \mathbb{Z}_2^n} \left(g(x) \cdot \sum_{w \in \mathbb{Z}_2^n} \chi_{x+y}(w) \right) = \frac{1}{2^n} \cdot \sum_{x \in \mathbb{Z}_2^n} (g(x) \cdot 2^n \cdot \delta_{x,y}) = g(y).$$

□

Другими словами, лемма 3.17 доказывает полноту ортогонального базиса характеров в пространстве $L(\mathbb{Z}_2^n)$.

Разложим равномерное распределение $\pi(x) = 2^{-n}$ по этому базису, используя результат леммы 3.13:

Лемма 3.18. $\hat{\pi}(w) = \delta_0(w) = \delta_{0,w}$.

Вводя обозначение $\Sigma(w) = w_1 + \ldots + w_n$, где $w = (w_1, \ldots, w_n) \in \mathbb{Z}_2^n$, получим разложение по базису характеров распределения вероятностей μ, задаваемое равенствами (3.9):

Лемма 3.19.

$$\hat{\mu}(w) = 1 - \frac{\Sigma(w)}{n}.$$

Доказательство. Заметим, что $\chi_w(e_j) = 1-2\cdot w_j$ и $\chi_w(0) = 1$. Тогда

$$\hat{\mu}(w) = \sum_{x\in\mathbb{Z}_2^n} \mu(x)\cdot\chi_w(x) = \frac{1}{2}\cdot\chi_w(0) + \frac{1}{2n}\cdot[\chi_w(e_1)+\ldots+\chi_w(e_n)] =$$
$$= \frac{1}{2} + \frac{1}{2n}\cdot[(1-2\cdot w_1)+\ldots+(1-2\cdot w_n)] =$$
$$= \frac{1}{2} + \frac{n}{2n} - \frac{1}{2n}\cdot[2\cdot w_1+\ldots+2\cdot w_n] =$$
$$= \left(\frac{1}{2}+\frac{1}{2}\right) - \frac{w_1+\ldots+w_n}{n} = 1 - \frac{\Sigma(w)}{n}.$$

□

Используя равенство (3.12), из предыдущей леммы получаем:

$$\widehat{\mu^{*t}}(w) = \left(1-\frac{\Sigma(w)}{n}\right)^t. \tag{3.13}$$

Еще один нужный нам результат носит название *формулы Планшереля*:

Лемма 3.20. *Для $g \in L(\mathbb{Z}_2^n)$ имеем*

$$\|\hat{g}\|^2 = 2^n\cdot\|g\|^2.$$

Доказательство.

$$\|\hat{g}\|^2 = \langle\hat{g},\hat{g}\rangle = \sum_{w\in\mathbb{Z}_2^n}\hat{g}(w)\cdot\overline{\hat{g}(w)} =$$
$$= \sum_{x,y\in\mathbb{Z}_2^n}\left[g(x)\cdot\overline{g(y)}\cdot\sum_{w\in\mathbb{Z}_2^n}\chi_w(x)\cdot\chi_w(y)\right] =$$
$$= 2^n\cdot\sum_{x\in\mathbb{Z}_2^n} g(x)\cdot\overline{g(x)} = 2^n\cdot\|g\|^2.$$

□

Воспользуемся *неравенством Коши-Буняковского*:

$$(\|\nu - \pi\|_{TV})^2 = \frac{1}{4} \cdot \left(\sum_{i \in \mathbb{Z}_2^n} 1 \cdot |\nu_i - \pi_i| \right)^2 \leq \frac{1}{4} \cdot 2^n \cdot \|\nu - \pi\|^2. \quad (3.14)$$

Мы будем применять этот результат к $(\nu - \pi) \in L(\mathbb{Z}_2^n)$, где $\nu = \mu^{*t}$, μ определяется равенствами (3.9), а π - равномерное распределение на $\mathbb{Z}_2^n$.

Следующий результат является вариантом неравенства Дьякониса-Шахшахани:

Лемма 3.21. *Для μ, определяемого равенствами (3.9), и равномерного распределения π имеем*

$$\left(\|\mu^{*t} - \pi\|_{TV}\right)^2 \leq \frac{1}{4} \cdot \sum_{w \neq 0} \left(1 - \frac{\Sigma(w)}{n}\right)^{2t}.$$

Доказательство.

$$\left(\|\mu^{*t} - \pi\|_{TV}\right)^2 \leq \frac{1}{4} \cdot 2^n \cdot \|\mu^{*t} - \pi\|^2 = \frac{1}{4} \cdot \|\widehat{\mu^{*t}} - \hat{\pi}\|^2 =$$
$$= \frac{1}{4} \cdot \sum_{w \neq 0} \left(1 - \frac{\Sigma(w)}{n}\right)^{2t}.$$

□

Первый основной результат этого параграфа таков:

Теорема 3.8. *Пусть μ определяется равенствами (3.9), а π - равномерное распределение. Для $t \geq \frac{1}{2} \cdot n \cdot (\log n + c)$ имеем*

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

Доказательство. Имеем

$$\left(\|\mu^{*t}-\pi\|_{TV}\right)^2 \le \frac{1}{4}\cdot\sum_{w\neq 0}\left(1-\frac{\Sigma(w)}{n}\right)^{2t} =$$

$$= \frac{1}{4}\cdot\sum_{k=1}^{n}\binom{n}{k}\cdot\left(1-\frac{k}{n}\right)^{2t} \le \frac{1}{4}\cdot\sum_{k=1}^{n}\frac{n^k}{k!}\cdot e^{-\frac{2tk}{n}}.$$

Первое неравенство - результат леммы 3.21, равенство - группировка $w \in \mathbb{Z}_2^n$ по $k = \Sigma(w)$, а последнее неравенство следует из $\binom{n}{k} = \frac{n!}{k!\cdot(n-k)!} \le \frac{n^k}{k!}$ и $\left(1-\frac{k}{n}\right)^{2t} \le e^{-\frac{2tk}{n}}$, которое, в свою очередь, выводится из неравенства $(1-x) \le e^{-x}$.

Продолжим оценку, подставляя $t \ge \frac{1}{2}\cdot n\cdot(\log n + c)$:

$$\left(\|\mu^{*t}-\pi\|_{TV}\right)^2 \le \frac{1}{4}\cdot\sum_{k=1}^{n}\frac{n^k}{k!}\cdot e^{-\frac{2tk}{n}} \le \frac{1}{4}\cdot\sum_{k=1}^{n}\frac{n^k}{k!}\cdot e^{-\frac{k\cdot n\cdot(\log n+c)}{n}} =$$

$$= \frac{1}{4}\cdot\sum_{k=1}^{n}\frac{1}{k!}\cdot e^{-k\cdot c} \le \frac{1}{4}\cdot\sum_{k=1}^{\infty}\frac{1}{k!}\cdot e^{-k\cdot c} = \frac{1}{4}\cdot\left(e^{e^{-c}}-1\right).$$

□

Теперь получаем оценку сверху для времени перемешивания:

$$\tau(\varepsilon) \le \frac{1}{2}\cdot n\cdot(\log n - \log\log(4\varepsilon^2+1)) \tag{3.15}$$

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

Для вывода оценки снизу на скорость перемешивания будем использовать лемму 3.1.

Другими словами, для любого подмножества $R \subseteq \mathbb{Z}_2^n$ имеем

$$\|\mu^{*t}-\pi\|_{TV} \ge \pi(R) - \mu^{*t}(R).$$

Мы выберем в качестве такого подмножества

$$R_d = \{x \in \mathbb{Z}_2^n \mid |n-2\Sigma(x)| < d\cdot\sqrt{n}\}, \tag{3.16}$$

то есть элементы с числом единиц, близким к половине.

Для дальнейшего обозначим подмодульное выражение через

$$\phi(x) = n - 2\Sigma(x) = \sum_{k=1}^{n} \chi_{e_k}(x), \tag{3.17}$$

где $e_1 = (1, 0, \ldots, 0); \ldots; e_n = (0, 0, \ldots, 1)$ - единичные орты.

Для любой функции $g \in L(\mathbb{Z}_2^n)$ и любого распределения вероятностей ν на $\mathbb{Z}_2^n$ определим *среднее* (значение):

$$\boldsymbol{E}_\nu(g) = \sum_{i \in \mathbb{Z}_2^n} g(x) \cdot \nu(x) \tag{3.18}$$

и *дисперсию*:

$$\boldsymbol{D}_\nu(g) = \boldsymbol{E}_\nu((g - \boldsymbol{E}_\nu(g))^2). \tag{3.19}$$

Вычислим среднее значение для ϕ относительно равномерного распределения π.

Лемма 3.22.

$$\boldsymbol{E}_\pi(\phi) = 0.$$

Доказательство.

$$\boldsymbol{E}_\pi(\phi) = \frac{1}{2^n} \cdot \sum_{y \in \mathbb{Z}_2^n} \phi(y) = \frac{1}{2^n} \cdot \sum_{k=1}^{n} \sum_{y \in \mathbb{Z}_2^n} \chi_{e_k}(y) = 0$$

по лемме 3.13. □

Вычислим дисперсию для ϕ относительно равномерного распределения π.

Лемма 3.23.

$$\boldsymbol{D}_\pi(\phi) = n.$$

Доказательство. Воспользуемся легко проверяемым из равенства (3.19) тождеством $\boldsymbol{D}_\pi(\phi) = \boldsymbol{E}_\pi(\phi^2) - (\boldsymbol{E}_\pi(\phi))^2$. Тогда по лемме 3.22

$$\boldsymbol{D}_\pi(\phi) = \boldsymbol{E}_\pi(\phi^2) = \frac{1}{2^n} \sum_{y \in \mathbb{Z}_2^n} \sum_{k,l=1}^{n} \chi_{e_k}(y) \cdot \chi_{e_l}(y) =$$

$$= \frac{1}{2^n} \cdot \sum_{k,l=1}^{n} \sum_{y \in \mathbb{Z}_2^n} \chi_{e_k}(y) \cdot \chi_{e_l}(y) = \frac{1}{2^n} \cdot \sum_{k=1}^{n} 2^n = n$$

по лемме 3.14. □

Вычислим среднее значение для ϕ относительно распределения μ^{*t}, где μ задается равенствами (3.9):

Лемма 3.24.

$$\boldsymbol{E}_{\mu^{*t}}(\phi) = n \cdot \left(1 - \frac{1}{n}\right)^t.$$

Доказательство.

$$\boldsymbol{E}_{\mu^{*t}}(\phi) = \sum_{k=1}^{n} \sum_{y \in \mathbb{Z}_2^n} \mu^{*t}(y) \cdot \chi_{e_k}(y) = \sum_{k=1}^{n} [\hat{\mu}(e_k)]^t = n \cdot \left(1 - \frac{1}{n}\right)^t,$$

где используется определение 3.11 и равенство (3.13), в котором для всех ортов e_k имеем $\Sigma(e_k) = 1$. □

Оценим дисперсию для ϕ относительно распределения μ^{*t}, где μ задается равенствами (3.9):

Лемма 3.25.

$$\boldsymbol{D}_{\mu^{*t}}(\phi) \leq n.$$

Доказательство. Опять используем тождество $\boldsymbol{D}_{\mu^{*t}}(\phi) = \boldsymbol{E}_{\mu^{*t}}(\phi^2) - (\boldsymbol{E}_{\mu^{*t}}(\phi))^2$. По лемме 3.24 $(\boldsymbol{E}_{\mu^{*t}}(\phi))^2 = n^2 \cdot \left(1 - \frac{1}{n}\right)^{2t}$.

Получим выражение для

$$\boldsymbol{E}_{\mu^{*t}}(\phi^2) = \sum_{y\in\mathbb{Z}_2^n} \mu^{*t}(y)\cdot\sum_{k,l=1}^{n}\chi_{e_k}(y)\cdot\chi_{e_l}(y) =$$

$$= \sum_{y\in\mathbb{Z}_2^n} \mu^{*t}(y)\cdot\sum_{k,l=1}^{n}\chi_{e_k+e_l}(y) = n\cdot[\hat{\mu}(0)]^t + \sum_{k,l=1:k\neq l}^{n}[\hat{\mu}(e_k+e_l)]^t,$$

так как $\chi_{e_k}\cdot\chi_{e_l} = \chi_{e_k+e_l}$, причем при $k = l$ будет $e_k + e_l = 0$, далее используется определение 3.11 совместно с формулой (3.13).

По формуле (3.13) $[\hat{\mu}(0)]^t = 1$ и $[\hat{\mu}(e_k+e_l)]^t = \left(1-\frac{2}{n}\right)^t$. Собираем все вместе

$$\boldsymbol{E}_{\mu^{*t}}(\phi^2) = n\cdot 1 + n\cdot(n-1)\cdot\left(1-\frac{2}{n}\right)^t.$$

Наконец, получаем

$$\boldsymbol{D}_{\mu^{*t}}(\phi^2) = n\cdot 1 + n\cdot(n-1)\cdot\left(1-\frac{2}{n}\right)^t - n^2\cdot\left(1-\frac{1}{n}\right)^{2t} \leq n,$$

так как $n^{t+1}\cdot(n-1)\cdot(n-2)^t \leq n^2\cdot(n-1)^{2t}$. □

Также нам понадобится *неравенство Чебышева*:

$$\nu\{x\in\mathbb{Z}_2^n \mid |g(x) - \boldsymbol{E}_\nu(g)| \geq \alpha\} \leq \frac{\boldsymbol{D}_\nu(g)}{\alpha^2}. \tag{3.20}$$

Сначала применяем неравенство Чебышева к $g = \phi$ и равномерному распределению π, чтобы оценить вероятность попадания в множество R_d:

Лемма 3.26.

$$\pi(R_d) \geq 1 - \frac{1}{d^2}.$$

Доказательство.

$$\pi(R_d) = 1 - \pi\{x \in \mathbb{Z}_2^n \mid |\phi(x)| \geq d \cdot \sqrt{n}\} = \\ = 1 - \pi\{x \in \mathbb{Z}_2^n \mid |\phi(x) - \boldsymbol{E}_\pi(\phi)| \geq d \cdot \sqrt{n}\} \geq \\ \geq 1 - \frac{\boldsymbol{D}_\pi(\phi)}{d^2 \cdot n} = 1 - \frac{1}{d^2}.$$

□

Теперь можно доказать второй основной результат этого параграфа:

Теорема 3.9. *Пусть μ определяется равенствами (3.9), а π - равномерное распределение. Для $t = \frac{1}{2} \cdot n \cdot (\log n - 2 \cdot \log(2d))$ и достаточно больших n имеем*

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

Доказательство. По лемме 3.1 $\|\mu^{*t} - \pi\|_{TV} \geq \pi(R_d) - \mu^{*t}(R_d)$, где R_d задается равенством (3.16).

Лемма 3.26 дает нам оценку $\pi(R_d) \geq 1 - \frac{1}{d^2}$. Осталось оценить сверху $\mu^{*t}(R_d)$.

Так как $|\phi(x)| < d \cdot \sqrt{n}$ влечет $\boldsymbol{E}_{\mu^{*t}}(\phi) - d \cdot \sqrt{n} \leq |\phi(x) - \boldsymbol{E}_{\mu^{*t}}(\phi)|$, то

$$R_d = \{x \in \mathbb{Z}_2^n \mid |\phi(x)| < d \cdot \sqrt{n}\} \subseteq \\ \subseteq \{x \in \mathbb{Z}_2^n \mid |\phi(x) - \boldsymbol{E}_{\mu^{*t}}(\phi)| \geq \boldsymbol{E}_{\mu^{*t}}(\phi) - d \cdot \sqrt{n}\}.$$

Применение неравенства Чебышева (формула (3.20)) дает

$$\mu^{*t}(R_d) \leq \frac{\boldsymbol{D}_{\mu^{*t}}(\phi)}{\left(\boldsymbol{E}_{\mu^{*t}}(\phi) - d \cdot \sqrt{n}\right)^2}.$$

Лемма 3.25 гарантирует, что $\boldsymbol{D}_{\mu^{*t}}(\phi) \leq n$. Осталось оценить знаменатель.

По лемме 3.24 для $t = \frac{1}{2} \cdot n \cdot (\log n - 2 \cdot \log(2d))$ имеем

$$\boldsymbol{E}_{\mu^{*t}}(\phi) = n \cdot \exp\left[\log\left(1 - \frac{1}{n}\right) \cdot \frac{1}{2} \cdot n \cdot (\log n - 2 \cdot \log(2d))\right] =$$
$$= n \cdot \exp\left[\left(-\frac{1}{n} - \frac{1}{2n^2} \cdot (1 + o(1))\right) \cdot \frac{1}{2} \cdot n \cdot (\log n - 2 \cdot \log(2d))\right] =$$
$$= n \cdot \exp\left[\left(-\frac{\log n - 2 \cdot \log(2d)}{2} - \frac{\log n - 2 \cdot \log(2d)}{4n} \cdot (1 + o(1))\right)\right] =$$
$$= n \cdot \frac{2d}{\sqrt{n}} \exp\left[-\frac{\log n - 2 \cdot \log(2d)}{4n} \cdot (1 + o(1))\right] \geq \frac{3}{4} \cdot 2d \cdot \sqrt{n}$$

для достаточно больших n, так как $\frac{2 \cdot \log(2d) - \log n}{4n} \to 0$ при $n \to \infty$.

Поэтому

$$\left(\boldsymbol{E}_{\mu^{*t}}(\phi) - d \cdot \sqrt{n}\right)^2 \geq \left(\frac{3}{2} \cdot d \cdot \sqrt{n} - d \cdot \sqrt{n}\right)^2 = \frac{d^2 \cdot n}{4},$$

что дает в результате

$$\mu^{*t}(R_d) \leq \frac{n}{\frac{d^2 \cdot n}{4}} \leq \frac{4}{d^2}.$$

Собирая все вместе, имеем

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

□

Оба ключевых результата этого параграфа вместе доказывают эффект быстрого перемешивания для монотонной цепи Маркова, определяемой алгоритмом 3 на формальном контексте Булеана.

К сожалению, никаких общих результатов о времени перемешивания алгоритма 3 для произвольного формального контекста автору неизвестно. Хочется надеяться, что в этом открытом вопросе появится прогресс.

По поводу значения результатов настоящего параграфа можно сделать несколько замечаний:

Во-первых, техника преобразований Фурье из теории представлений конечных групп не может быть расширена на цепи Маркова, порождаемые алгоритмом 3 из произвольных формальных контекстов. Препятствием, например, является неравномерность стационарного распределения, что было установлено в параграфе 3.2.

Во-вторых, стационарное распределение часто не имеет легко вычисляемого вида. Пример такой ситуации имеется в параграфе 3.2. Да и смысла порождать кандидаты с распределением, близким к стационарному, особого нет: результат ключевой теоремы 4.1 не зависит от вида распределения ВКФ-гипотез.

В-третьих, логические условия, налагаемые на кандидаты, чтобы стать ВКФ-гипотезой (порог числа родителей и запрет на контрпримеры), изменяют распределение ВКФ-гипотез, выдаваемых алгоритмом 6.

Наконец, «ленивые» вычисления операций «замыкай-по-одному» из параграфа 1.3 тоже изменяют распределение порождаемых кандидатов.

Тем не менее, исследование скорости сходимости монотонной цепи Маркова может дать возможность заменить алгоритм 4 спаривающей цепи Маркова алгоритмом 3 монотонной цепи Маркова.

Однако против такой замены спаривающей цепи Маркова монотонной цепью из алгоритма 3 говорят многие теоретические результаты, установленные в настоящей работе.

Автор полагает, что методом, который может привести к продвижению в задаче оценки скорости перемешивания монотонной цепи Маркова, примененной к произвольному формальному контексту, является «спаривание» [69]. Как это работает в других ситуациях, можно посмотреть в работах [68, 74].

Другим перспективным методом является «эволюционирующие множества» [70, 72].

Основные выводы

1. Алгоритмы вероятностного нахождения сходств (немонотонный, монотонный и спаривающий) соответствуют цепям Маркова (теорема 3.1 и замечание после нее).

2. Алгоритм 4 спаривающей цепи Маркова останавливается с вероятностью единица (следствие из теоремы 3.2).

3. В случае малого числа длинных траекторий ими можно пренебречь с оценкой изменений результатов в теореме 3.3.

4. Для случая Булеана (множества всех подмножеств признаков) среднее время работы алгоритма 4 имеет порядок $O(n \cdot \ln n)$.

5. Для случая Булеана можно заменить алгоритм 4 на алгоритм 3, используя оценку (3.15) времени перемешивания.

Глава 4

Машинное обучение, основанное на теории решеток

Мы предлагаем использовать вероятностный подход к машинному обучению, с использованием техники теории решеток, для порождения причин из обучающей выборки сложно-структурированных прецедентов.

В первом параграфе мы опишем генезис развиваемых процедур в рамках логико-комбинаторного подхода проф. В.К. Финна и его учеников, перечислим еще раз проблемы, с которыми он сталкивается, а затем сформулируем новый подход, который мы называем ВКФ-метод, как дань уважения нашему учителю Виктору Константиновичу Финну.

Параграф 4.2 посвящен описанию процедур машинного обучения, основанных на сходстве, для порождения причин, достаточных для проведения правдоподобных рассуждений. Для доопределения по аналогии будет установлен ключевой результат (теорема 4.1) о надежности (оценка правдоподобия получаемых результатов).

Описание программной реализации ВКФ-метода содержится в параграфе 4.3.

Параграф 4.4 описывает апробацию разработанного подхода на

массивах из репозитория данных для тестирования алгоритмов машинного обучения.

4.1 Истоки: ДСМ-метод

В начале 1980-х проф. В.К.Финн [56] придумал ДСМ-метод, который объединяет несколько когнитивных процедур:

1. индуктивное обобщение данных (развивая идеи Д.С.Милля [44]);
2. предсказание целевого свойства у новых объектов по аналогии с обучающими примерами;
3. абдуктивное принятие гипотез (развивая идеи Ч.С.Пирса [50]).

ДСМ-гипотезами будут называться сходства обучающих примеров, удовлетворяющие дополнительным условиям. Минимальные требования - число родителей не менее двух и «запрет контр-примеров» (чтобы ни один контр-пример не выжил в смысле главы 2).

Развитие ДСМ-метода привело к созданию программных ДСМ-систем [31], которые применяются к самым разнообразным предметным областям:

- социология и социальная психология [46];
- фармакология [4];
- медицина [25, 49];
- датировки исторических источников [29];
- биология [66];
- почерковедческая экспертиза;
- техническая диагностика.

Однако 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 (пример Булеана, демонстрирующий этот феномен, приведен в параграфе 1.2).

Во-вторых, С.О. Кузнецовым [39], М.И. Забежайло и другими авторами были доказаны NP- и $\#P$-полнота для многих ДСМ-процедур.

В-третьих, появление фантомных сходств, которые были исследованы в главе 2, снижает качество предсказания по аналогии, аналогично феномену «переобучения» в других процедурах машинного обучения.

Чтобы справиться с возникающими эффектами автором предлагается новый вероятностно-комбинаторный подход.

ВКФ-метод использует вероятностные модификации двух процедур из ДСМ-метода:

1. индуктивное обобщение обучающих примеров в структурных фрагментах - гипотезах о причинах проявления исследуемого свойства;

2. предсказание целевого свойства у тестовых примеров с помощью порожденных гипотез (по аналогии с обучающими примерами).

Абдукция - условие принятия порожденных гипотез - первоначально была дополнена процедурой абдуктивного уточнения множества гипотез.

Абдуктивное уточнение заключается в применении операции

$$CbODown(\langle A, B\rangle, o)$$

к каждому исходному обучающему примеру o и каждой порожденной на шаге индукции ВКФ-гипотезе $\langle A, B\rangle$.

Это казалось необходимым для увеличения шансов найти ВКФ-гипотезу, которая правильно объясняет исходный обучающий пример. Из-за вероятностного характера порождения гипотез все ВКФ-гипотезы, включающиеся в выбранный пример, могли быть пропущены.

Однако, как показали эксперименты на реальных данных, отказ от абдуктивного уточнения за счет увеличения объема порождаемых ВКФ-гипотез на этапе индукции, контролируемого с помощью результата теоремы 4.1, приводит к уменьшению общего количества ВКФ-гипотез на несколько порядков (и соответствующему сокращению времени работы) без снижения качества предсказания тестовых примеров.

Так как меньшее количество гипотез легче анализировать экспертам, мы предпочли отказаться от процедуры абдуктивного уточнения. Следует, однако, отметить, что нахождение уникальных обучающих примеров (не объясняемых порожденными гипотезами) представляет интерес в качестве процедуры, позволяющей пополнять обучающую выборку их аналогами для выявления дополнительных механизмов, вызывающих проявление целевого свойства.

4.2 Процедуры ВКФ-метода

Мы изменим порядок изложения процедур машинного обучения, основанного на сходстве, с той целью, чтобы легче было устанавливать их свойства. Они будут рассматриваться в следующем порядке:

1. предсказание целевого свойства у тестовых объектов по аналогии с обучающими примерами;

2. индуктивное обобщение обучающих примеров (используя цепь Маркова для порождения вероятностным образом заранее предписанного числа гипотез);

3. абдуктивное уточнение и принятие гипотез (порождая дополнительные гипотезы для объяснения исходных примеров).

Предсказание целевого свойства по аналогии с обучающими примерами осуществляется с помощью следующего алгоритма:

Data: расширенная выборка S^+ ВКФ-гипотез, файл (τ)-примеров
Result: предсказанные свойства (τ)-примеров
$X := (\tau)$-примеры;
for $(o \in X)$ **do**
  $PredictPositively(o) :=$ **false**;
  for $(\langle A, B\rangle \in S^+)$ **do**
    if $(B \subseteq o')$ **then**
      $PredictPositively(o) :=$ **true**;
    end
  end
end

Algorithm 5: Процедура предсказания по аналогии

Процедура предсказания по аналогии (алгоритм 5) пытается найти вложение хотя бы одного фрагмента $B \subseteq o'$, соответствующего хотя одной из порожденных (индукцией и абдукцией) ВКФ-гипотез $\langle A, B\rangle$, в каждый (τ)-пример o. Если такое вложение случается, то для этого (τ)-примера предсказывается наличие целевого свойства по аналогии с родителями $A \subseteq O$ ВКФ-гипотезы $\langle A, B\rangle$, чей фрагмент B вложился. Иначе предсказывается отсутствие целевого свойства у этого (τ)-примера o.

Установим полезную характеризацию предсказания по аналогии. Начнем с формального определения:

Определение 4.1. *Объект o, описываемый фрагментом $o' \subseteq F$ (множеством признаков),* ***предсказывается положительным*** *с помощью ВКФ-гипотезы $\langle A, B\rangle$, если $B \subseteq o'$.*

Если число признаков равно $n = |F|$, то можно рассматривать вершины n-мерного гиперкуба $\{0, 1\}^n$.

Каждый объект o, предъявляемый для предсказания, задает семейство нижних полупространств в $\boldsymbol{R}^n$:

Определение 4.2. ***Нижнее полупространство*** $H^{\downarrow}_{\varkappa}(o)$, *определяемое объектом* o *с фрагментом* $o' \subseteq F$, *задается линейным неравенством*

$$x_{j_1} + \ldots + x_{j_k} < \varkappa,$$

где $F \setminus o' = \{f_{j_1}, \ldots, f_{j_k}\}$ *и* $0 < \varkappa < 1$. *Допускается также вырожденное нижнее полупространство* $0 < \varkappa$, *соответствующее* $o' = F$, *и совпадающее со всем* $\boldsymbol{R}^n$. *Класс нижних полупространств, определяемых объектами, обозначим через* $(Sub \downarrow)$.

Лемма 4.1. *Пример* o *предсказывается положительным тогда и только тогда, когда в любом его нижнем полупространстве содержится хотя одна ВКФ-гипотеза.*

Доказательство. По определению 4.1 условие, что ВКФ-гипотеза $\langle A, B\rangle$ предсказывает объект o положительным, эквивалентно $B \subseteq o'$, то есть условию $B \cap (F \setminus o') = \emptyset$. Но в обозначения определения 4.2 это означает, что $\forall i\,[f_{j_i} \notin B]$. Последнее эквивалентно условию $0 = x_{j_1} + \ldots + x_{j_k} < \varkappa$ для любого $0 < \varkappa < 1$. □

Индуктивное обобщение обучающих примеров осуществляется сле-

дующей 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;
  $hasObstacle :=$ **false**;
  for $(c \in C)$ **do**
    if $(B \subseteq c')$ **then**
      $hasObstacle :=$ **true**;
    end
  end
  if $(hasObstacle =$ ***false***$)$ **then**
    $S := S \cup \{\langle A, B\rangle\}$;
    $i := i + 1$;
  end
end

Algorithm 6: Процедура индуктивного обобщения

Проверка условия $(B \subseteq c')$ в алгоритме 6 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 4) рассмотрим задачу вероятно приближенно корректного (*probably approximately correct* (PAC-)) обучения [6], [75].

Зафиксируем $\varepsilon > 0$ - точность предсказания.

Определение 4.3. *Объект o назовем ε-**важным**, если суммарная вероятность появления таких ВКФ-гипотез $\langle A, B\rangle$, что $B \in H^{\downarrow}_{\varkappa}(o)$ будет больше ε.*

*Семейство ВКФ-гипотез назовем ε-**сетью**, если для каждого ε-важного объекта найдется хотя бы одна ВКФ-гипотеза из этого семейства, которая предскажет этот объект положительно.*

Теперь мы будем проводить рассуждения, аналогичные рассуждениям В.Н. Вапника и А.Я. Червоненкиса [8], хотя нас будет интересовать только вероятность ошибки «первого рода» (отказ от положительного предсказания).

Другими словами, требуется найти такое число N, зависящее от ε и δ, что с вероятностью, большей $1-\delta$, случайная выборка объема N будет образовывать ε-сеть.

Для того, чтобы прямо сравнить с результатами В.Н. Вапника и А.Я. Червоненкиса, несколько расширим класс подмножеств.

Определение 4.4. ***Нижнее полупространство*** *задается линейным неравенством*

$$\sum_{j=1}^{n} w_j \cdot x_j < \varkappa,$$

где $\varkappa > 0$ и вектор нормали $\langle w_1, \ldots, w_n \rangle$ направлен в неотрицательный ортант $\boldsymbol{R}^n_+$, то есть $\forall j \, [w_j \geq 0]$. Семейство нижних полупространств обозначим через $(Lin \downarrow)$

Определение 4.5. ***Расколотым*** *называется такое множество в неотрицательном ортанте (без начала координат) $\boldsymbol{R}^n_+ \backslash \{\langle 0, \ldots, 0 \rangle\}$, что любое его подмножество может быть отколото нижним полупространством $H \in (Lin \downarrow)$.*

*Максимальная мощность $dim_{VC\downarrow}$ расколотого подмножества называется **размерностью Вапника-Червоненкиса**.*

*Максимальное число подмножеств, откалываемых нижними полупространствами $H \in (Lin \downarrow)$ у множества мощности l, задает **функцию роста** $m^{Lin\downarrow}(l)$.*

Сами В.Н. Вапник и А.Я. Червоненкис [8] называли свою характеристику *емкостью* класса подмножеств.

Нам потребуется лемма Радона из теории выпуклых тел:

Лемма 4.2. *Для любых $n+2$ точек $\{v_0, v_1, \ldots, v_{n+1}\}$ в $\boldsymbol{R}^n$ надется такое разбиение на два непустых подмножества, что их линейные оболочки пересекаются.*

Доказательство. Рассмотрим $n+1$ вектор $\{v_1 - v_0, \ldots, v_{n+1} - v_0\}$ в $\mathbf{R}^n$. Они линейно зависимы, то есть найдутся такие не все одновременно равные нулю числа $\{\lambda_1, \ldots, \lambda_{n+1}\}$, что $\sum_{j=1}^{n+1} \lambda_j \cdot (v_j - v_0) = 0$. Обозначим через λ_0 число $-\sum_{j=1}^{n+1} \lambda_j$. Пусть $\Lambda = \{\lambda_{j_1}, \ldots, \lambda_{j_l}\}$ - множество всех неотрицательных чисел среди $\{\lambda_0, \lambda_1, \ldots, \lambda_{n+1}\}$. Тогда

$$\sum_{i=1}^{l} \lambda_{j_i} \cdot v_{j_i} = \sum_{j=0:\lambda_j \notin \Lambda}^{n+1} \lambda_j \cdot v_j,$$

причем и слева и справа стоят положительные коэффициенты. Положим теперь $\mu_j = \frac{\lambda_j}{\sum_{i=1}^{l} \lambda_{j_i}}$. Тогда

$$\sum_{j=0:\lambda_j \in \Lambda}^{n+1} \mu_j \cdot v_j = \sum_{j=0:\lambda_j \notin \Lambda}^{n+1} \mu_j \cdot v_j$$

задает представление общей точки выпуклых оболочек $\{v_j \mid \lambda_j \in \Lambda\}$ и его дополнения $\{v_j \mid \lambda_j \notin \Lambda\}$. □

Лемма 4.3. *$dim_{VC\downarrow}(\boldsymbol{R}^n_+) = n$.*

Доказательство. Ясно, что вершины единичного симплекса образуют n-элементное расколотое подмножество. Докажем, что никакое $(n+1)$-элементное подмножество точек в $\mathbf{R}^n_+$ не является расколотым. Добавим к нашему множеству начало координат. По лемме 4.2 Радона найдется разбиение этих $n+2$ точек на два непересекающихся подмножества, чьи выпуклые оболочки пересекаются. Отберем подмножество исходных точек, попавших в одну группу с началом координат. Ясно, что это множество не может быть отколото никаким нижним полупространством, так как комбинации точек, лежащих в нижнем полупространстве, сами лежат в нижнем полупространстве, а комбинации точек из верхнего полупространства лежат в верхнем. □

Теперь, дословно повторяя рассуждения Вапника-Червоненкиса [7], получаем следующий результат:

Лемма 4.4. $m^{Lin\downarrow}(l) \leq \sum_{i=0}^{n} \binom{l}{i}$. □

Лемма 4.5. *Для $l \geq n$ верно $m^{Lin\downarrow}(l) \leq \left(\frac{e \cdot l}{n}\right)^n$.*

Доказательство. По лемме 4.4 нужно доказать $\sum_{i=0}^{n} \binom{l}{i} \leq \left(\frac{e \cdot l}{n}\right)^n$ при $l \geq n$. Но

$$\left(\frac{n}{l}\right)^n \cdot \sum_{i=0}^{n} \binom{l}{i} \leq \sum_{i=0}^{n} \binom{l}{i} \cdot \left(\frac{n}{l}\right)^i \leq$$

$$\leq \sum_{i=0}^{l} \binom{l}{i} \cdot \left(\frac{n}{l}\right)^i = \left(1 + \frac{n}{l}\right)^l \leq e^n.$$

□

Лемма 4.5 слегка усиливает оригинальную оценку В.Н. Вапника и А.Я. Червоненкиса [7], но изложена, например, в учебнике В.В. Вьюгина [26].

Но в нашем случае функция роста оказывается независимой от l:

Определение 4.6. *Максимальное число подмножеств, откалываемых нижними полупространствами $H \in (Sub\downarrow)$ у множества мощности l, задает* ***функцию роста*** $m^{Sub\downarrow}(l)$.

Лемма 4.6. *Для $l \geq n$ верно $m^{Sub\downarrow}(l) = 2^n$.*

Доказательство. Легко проверить, что вершины единичного симплекса образуют расколотое множество относительно $(Sub\downarrow)$.

С другой стороны, $(Sub\downarrow)$ содержит 2^n элементов. □

Применим метод повторной выборки Вапника-Червоненкиса [7], [6]. В следующей лемме $|S_2 \cap H|$ понимается с учетом кратности, так как элементы повторной выборки S_2 могут повторяться.

Лемма 4.7. *Для любого ε при $l > \frac{2}{\varepsilon}$ для независимых случайных выборок S_1 и S_2 ВКФ-гипотез объемов l имеем оценку:*

$$\boldsymbol{P}^l\{S_1 : \exists H \in (Sub\downarrow)\,[S_1 \cap H = \emptyset, \boldsymbol{P}H > \varepsilon]\} \leq$$
$$\leq 2 \cdot \boldsymbol{P}^{2l}\{S_1 S_2 : \exists H \in (Sub\downarrow)\,[S_1 \cap H = \emptyset, |S_2 \cap H| > \varepsilon \cdot l/2]\}.$$

Доказательство. Для выборки S_1 рассмотрим нижнее полупространство $H \in (Sub\downarrow)$, удовлетворяющее условиям $S_1 \cap H = \emptyset$ и $\boldsymbol{P}H > \varepsilon$. Из неравенства треугольника следует, что

$$\boldsymbol{P}^l\{S_2 : l \cdot \boldsymbol{P}H - |S_2 \cap H| \leq \varepsilon \cdot l/2\} \leq \boldsymbol{P}^l\{S_2 : |S_2 \cap H| > \varepsilon \cdot l/2\}.$$

Покажем, что при $l > \frac{2}{\varepsilon}$ выполняется

$$\boldsymbol{P}^l\{S_2 : l \cdot \boldsymbol{P}H - |S_2 \cap H| \leq \frac{\varepsilon \cdot l}{2}\} =$$
$$= \boldsymbol{P}^l\{S_2 : l \cdot \boldsymbol{P}H - \frac{\varepsilon \cdot l}{2} \leq |S_2 \cap H|\} \geq \frac{1}{2}.$$

Это - вероятность для биномиальной случайной величины $|S_2 \cap H|$ быть не меньше своего математического ожидания $l \cdot \boldsymbol{P}H$ за вычетом $\frac{\varepsilon \cdot l}{2} > 1$ (при $l > \frac{2}{\varepsilon}$). Известно, что медиана биномиального распределения отличается от его среднего меньше, чем на единицу. Это и доказывает нужное нам неравенство.

Из-за независимости выборок S_1 и S_2 имеем

$$\frac{1}{2} \cdot \boldsymbol{P}^l\{S_1 : S_1 \cap H = \emptyset, \boldsymbol{P}H > \varepsilon\} \leq$$
$$\leq \boldsymbol{P}^l\{S_2 : l \cdot \boldsymbol{P}H - |S_2 \cap H| \leq \frac{\varepsilon \cdot l}{2}\} \cdot \boldsymbol{P}^l\{S_1 : S_1 \cap H = \emptyset, \boldsymbol{P}H > \varepsilon\} \leq$$
$$\leq \boldsymbol{P}^l\{S_2 : |S_2 \cap H| > \varepsilon \cdot l/2\} \cdot \boldsymbol{P}^l\{S_1 : S_1 \cap H = \emptyset, \boldsymbol{P}H > \varepsilon\} =$$
$$= \boldsymbol{P}^{2l}\{S_1 S_2 : S_1 \cap H = \emptyset, |S_2 \cap H| > \varepsilon \cdot l/2\}.$$

Объединяя события в левой и в правой частях по $H \in (Sub\downarrow)$, удовлетворяющим условиям $S_1 \cap H = \emptyset$ и $\boldsymbol{P}H > \varepsilon$, получаем утверждение леммы. $\square$

Лемма 4.8. *Для любого ε для двух независимых случайных выборок S_1 и S_2 ВКФ-гипотез объемов l имеем оценку:*

$$\boldsymbol{P}^{2l}\{S_1S_2 : \exists H \in (Sub\downarrow)\,[S_1 \cap H = \emptyset, |S_2 \cap H| > \varepsilon \cdot l]\} \leq \\ \leq m^{Sub\downarrow}(2l) \cdot 2^{-\varepsilon l}.$$

Доказательство. Зададим отображение g выборки S_1S_2 в мультимножество над $\{0,1\}^n$, фиксирующее состав выборки. Вероятность $\boldsymbol{P}^{2l}$ на $(\{0,1\}^n)^{2l}$ индуцирует вероятностную меру $g(\boldsymbol{P})$ на составах выборки объема $2l$. Из-за независимости и одинаковой распределенности ВКФ-гипотез в разных прогонах спаривающей цепи Маркова вероятности на выборках одного состава одинаковы. Фиксируем состав выборки ν. Для конкретного нижнего полупространства $H \in (Sub\downarrow)$ условная вероятность оценивается через гипергеометрическое распределение:

$$\boldsymbol{P}\{S_1S_2 : [S_1 \cap H = \emptyset, |S_2 \cap H| > \varepsilon \cdot l] \mid g(S_1S_2) = \nu\} \leq \\ \leq \frac{\binom{l}{\varepsilon l}}{\binom{2l}{\varepsilon l}} = \frac{l! \cdot (\varepsilon l)! \cdot (2l - \varepsilon l)!}{(\varepsilon l)! \cdot (l - \varepsilon l)! \cdot (2l)!} = \\ = \frac{l \cdot (l-1) \cdot \ldots \cdot (l - \varepsilon l + 1)}{(2l) \cdot (2l-1) \cdot \ldots \cdot (2l - \varepsilon l + 1)} \leq 2^{-\varepsilon l}.$$

Тогда условная вероятность на выборках заданного состава

$$\boldsymbol{P}\,\{S_1S_2 : \exists H \in (Sub\downarrow)[S_1 \cap H = \emptyset, |S_2 \cap H| > \varepsilon \cdot l] \mid \\ \mid g(S_1S_2) = \nu\} \leq m^{Sub\downarrow}(2l) \cdot 2^{-\varepsilon l}.$$

Правая часть не зависит от состава ν, поэтому, интегрируя по мере $g(\boldsymbol{P})$, получаем утверждение леммы. □

Собираем вместе результаты трех предыдущих лемм и получаем основной результат этого параграфа.

Теорема 4.1. *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.*

Доказательство. По леммам 4.7 и 4.8 для $N > \frac{2}{\varepsilon}$ имеем оценку

$$\boldsymbol{P}^N\{S \subseteq \{0,1\}^n \mid \exists H \in (Sub\downarrow)\,[S \cap H = \emptyset, \boldsymbol{P}H > \varepsilon]\} \leq \\ \leq 2 \cdot m^{Sub\downarrow}(2N) \cdot 2^{-\varepsilon N/2}.$$

Теперь по лемме 4.6 остается решить неравенство $2 \cdot 2^n \cdot 2^{-\varepsilon N/2} \leq \delta$ относительно N, чтобы получить утверждение теоремы. □

Отметим некоторую специфику нашего подхода относительно классической парадигмы Вапника-Червоненкиса:

1. Мы ограничиваемся рассмотрением точек (фрагментов ВКФ-гипотез) в вершинах единичного гиперкуба. Поэтому удается получить лучшую оценку (не зависящую от длины обучающей выборки, лишь бы эта длина была больше размерности n пространства. Поэтому метод повторной выборки не дает дополнительного завышающего множителя. Хотя оценка все равно сильно завышена по другим причинам.

2. Тестовые примеры задают множества точек (отсекаемые гиперплоскостями) на гиперкубе, где должны оказаться фрагменты ВКФ-гипотез, чтобы предсказать тестовые примеры положительно. Это двойственно парадигме Вапника-Червоненкиса (там гипотезы определяют области пространства, куда должны попасть тестовые точки).

3. Спаривающая цепь Маркова используется для порождения ВКФ-гипотез из обучающих примеров, при этом независимые траектории порождают независимые элементы решетки кандидатов (с распределением первого попадания в ВКФ-гипотезы).

4. Мы рассматриваем только ошибки первого рода, когда положительный тестовый пример не предсказывается положительным. Про неправильное предсказание отрицательных примеров речь не идет.

Еще одной исследовавшейся процедурой ВКФ-метода является алгоритм абдуктивного уточнения множества гипотез. Эта процедура является расширением условия принятия ДСМ-гипотез, сформулированная В.К. Финном по результатам анализа идей Ч.С. Пирса об абдукции.

В своем первоначальном виде (у Ч.С. Пирса) абдукция задавала схему принятия гипотез посредством проверки того, что они объясняют эмпирические факты. Уточнение В.К. Финна делает процедуру абдукции конструктивной (гипотезы порождаются с помощью индуктивного обобщения обучающих примеров). При этом свойства самих примеров служат эмпирическими фактами, требующими объяснения. Объяснение же состоит в предъявлении (гипотетической) причины для наблюдаемого эффекта у каждого обучающего примера.

Как уже было указано ранее, вероятностный характер алгоритма 6 не позволяет быть уверенным в том, что мы не пропустили все ВКФ-гипотезы, фрагменты которых содержатся в объясняемом примере. Логико-комбинаторный ДСМ-метод лишен указанного недостатка. Чтобы частично устранить указанный дефект ВКФ-метод может досчитывать дополнительные ВКФ-гипотезы с помощью аб-

дуктивного уточнения множества ВКФ-гипотез:

Data: выборка S ВКФ-гипотез, внешняя функция
$CbODown(\ ,\ )$ операции «замыкай-по-одному-вниз»
Result: расширенная выборка S^+ ВКФ-гипотез
$S^+ := \emptyset$;
$O :=$ (+)-примеры, $C :=$ (-)-примеры; **for** ($o \in O$ *and* $\langle A, B\rangle \in S$) **do**
  вычислить $\langle X, Y\rangle := CbODown(\langle A, B\rangle, o)$;
  $Explained(o) :=$ **false**; $hasObstacle :=$ **false**;
  for ($c \in C$) **do**
    if ($Y \subseteq c'$) **then**
      $hasObstacle :=$ **true**;
    end
  end
  if ($hasObstacle =$ ***false***) **then**
    $S^+ := S^+ \cup \{\langle X, Y\rangle\}$;
    $Explained(o) :=$ **true**;
  end
end

Algorithm 7: Процедура абдуктивного уточнения

Как видно из алгоритма 7 абдуктивное уточнение заключается в применении оператора $CbODown$ к каждому исходному обучающему примеру и каждой порожденной на шаге индукции ВКФ-гипотезе.

Проверка условия ($Y \subseteq c'$) в алгоритме 7 означает, что фрагмент Y кандидата $\langle X, Y\rangle$ вкладывается в фрагмент (множество признаков) контр-примера c, так проверяется условие «запрета контр-примеров».

Абдуктивное уточнение казалось необходимым для увеличения шансов найти ВКФ-гипотезу, которая правильно объясняет исходный обучающий пример. Из-за вероятностного характера порождения гипотез все ВКФ-гипотезы, включающиеся в выбранный пример, могли быть пропущены.

Однако, как показали эксперименты на реальных данных (см.

параграф 4.4), отказ от абдуктивного уточнения за счет увеличения объема порождаемых ВКФ-гипотез на этапе индукции, контролируемого с помощью результата теоремы 4.1, приводит к уменьшению общего количества ВКФ-гипотез на несколько порядков (и соответствующему сокращению времени работы) без снижения качества предсказания тестовых примеров.

Так как меньшее количество ВКФ-гипотез легче анализировать экспертам, мы предпочли отказаться от процедуры абдуктивного уточнения. Еще раз подчеркнем, что нахождение уникальных обучающих примеров (не объясняемых порожденными гипотезами) представляет интерес в качестве процедуры, позволяющей пополнять обучающую выборку их аналогами для выявления дополнительных механизмов, вызывающих проявление целевого свойства.

4.3 Программная реализация

Описанные выше алгоритмы были запрограммированы автором в программной системе, получившей название ВКФ-система:

- Программа реализована как консольное приложение *client.exe* с использованием библиотеки разделяемого доступа (libvkf.so под Linux, vkf.dll под Windows).
- Программа платформенно независима: она собиралась и запускалась под Windows и под Linux.
- Компилятор C++: под Linux - GNU C++ toolset (version 4.9.1 или более поздние), под Windows - Microsoft Visual BuildTools 2017.

При программной реализации системы мы отказались от классического варианта спаривающей цепи Маркова (алгоритм 4) в пользу

описываемой ниже «ленивой» версии:

Data: множество обучающих (+)-примеров
Result: случайный кандидат $\langle A_1, B_1\rangle$
$O :=$ (+)-примеры, $F :=$ признаки; $I \subseteq O \times F$ - формальный контекст;
$R := O \cup F$; $\langle A_1, B_1\rangle := \langle O, O'\rangle$; $\langle A_2, B_2\rangle := \langle F', F\rangle$;
$moveUp := true$;
while $(\langle A_1, B_1\rangle \neq \langle A_2, B_2\rangle)$ **do**
  Выбираем случайный элемент $r \in R$;
  if $(r \in O \&\& moveUp)$ **then**
    $B_1 := A_1'$; $B_2 := A_2'$;
  end
  if $(r \in O)$ **then**
    $B_1 := B_1 \cap (\{r\}')$; $B_2 := B_2 \cap (\{r\}')$;
  end
  if $(r \in F \&\& !moveUp)$ **then**
    $A_1 := B_1'$; $A_2 := B_2'$;
  end
  if $(r \in F)$ **then**
    $A_1 := A_1 \cap (\{r\}')$; $A_2 := A_2 \cap (\{r\}')$;
  end
end
if $(!moveUp)$ **then**
  $B_1 := A_1'$; $B_2 := A_2'$;
end
if $(moveUp)$ **then**
  $A_1 := B_1'$; $A_2 := B_2'$;
end
$\langle A, B\rangle := Min$;

Algorithm 8: Ленивая спаривающая цепь Маркова

Причина перехода к ленивому варианту кроется в значительном повышении скорости вычислений. Теоретически это следует из теоремы 1.1.

Л.А. Якимова в ее выпускной квалификационной работе бака-

лавра [60], выполненной под руководством автора, продемонстрировала, что выигрыш от такого перехода на реальных данных может достигать очень большой величины, близкой к теоретическому предсказанию.

Л.А. Якимова разработала программу, которая реализует сравнение алгоритмов 4 и 8 на массиве Mushrooms из репозитория данных для тестирования алгоритмов машинного обучения Университета Калифорнии в г. Ирвайн.

В этом массиве (фактически, оцифрованном «Определителе грибов Северной Америки» [71]) содержится описание 8124 грибов, из которых $k = 4208$ являются съедобными, а 3916 ядовитыми. Грибы кодировались битовыми строками длины $n = 124$ бита.

По теореме 1.1 средний выигрыш от применения ленивой схемы вычислений достигает на операциях замыкания

$$\frac{1}{2} \cdot \frac{(k+n)^2}{k \cdot n} \approx 18.$$

Л.А. Якимова запускала программную реализацию алгоритма 6 индуктивного обобщения с использованием спаривающих цепей Маркова в стандартном (алгоритм 4) и ленивом вариантах (алгоритм 8) и сравнивала времена вычисления выбранного числа ВКФ-гипотез. Соотношение этих промежутков времени (чуть более 17 раз) оказалось замечательно согласованным с теоретическим результатом, вычисленным выше. То, что выигрыш оказался несколько меньше, объясняется тем, что кроме операций взятия поляр (замыкания) в этих алгоритмах имеется еще операции сходства (побитового умножения), которые хоть и очень быстрые, тем не менее остаются в неизменном количестве. За подробностями читатель отсылается к [60].

Суммируем программные технологии, используемые при разработке ВКФ-системы:

- Объекты (обучающие примеры, контр-примеры и тестовые примеры, представленные для предсказания целевого свойства) представляются битовыми строками (объектами класса *boost* :: *dynamic_bitset*).

- Полученные наборы битовых строк хранятся в контейнерах типа *std* :: *vector* стандартной библиотеки C++.
- Классы *boost* :: *dll* обеспечивают платформенно независимый доступ к библиотеке разделяемого доступа.
- Программа использует классы *std* :: *random* для датчиков случайных чисел. Это нужно для ленивой спаривающей цепи Маркова (алгоритм 8).
- Для реализации многопоточности используются классы *std* :: *thread*.

ВКФ-система реализована с помощью следующих классов:

1. **AttrSet** - *dynamic_bitset* для представления фрагмента кандидата или фрагмента объекта;
2. **ObjSet** - *dynamic_bitset* для представления списка родителей кандидата;
3. **AOMatrix** - класс для представления формального контекста, хранящий наибольший и наименьший кандидаты для устранения повторных вычислений;
4. **Candidate** - пара объектов классов AttrSet и ObjSet, использующихся для представления кандидатов;
5. **Obstacles** - контейнер объектов класса AttrSet, представляющих контр-примеры;
6. **MCState** - упорядоченная пара объектов класса Candidate, применяемая в алгоритме 4 и реализующая вызовы операций «замыкай-по-одному»;
7. **Hypothesis** - потомок класса Candidate, запускающий алгоритм 8 в своем конструкторе;

8. **Hypotheses** - контейнер для Hypothesis, запускающий алгоритм 6 (индуктивное обобщение) в своем конструкторе, и реализующий алгоритм 7 (абдуктивное уточнение) как специальный метод;

9. **PredictMe** - класс для представления объектов, предъявленных для предсказания целевого свойства;

10. **TestSample** - контейнер для PredictMe, запускающий алгоритм 5 (предсказание по аналогии) в своем конструкторе.

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.

- ВКФ-система вычисляет процедуру абдуктивного уточнения и принятия ВКФ-гипотез тоже в несколько потоков.

- Предсказание свойств по аналогии осуществляется в один поток, так как вычислительная сложность этого шага мала в сравнении с шагом индукции.

- На современных компьютерах загрузка ядер процессора замечательно балансируется по вычислительным потокам (превышает 90% на этапе индуктивного обобщения).

4.4 Экспериментальная апробация

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 CLIP (версия 3), которая реализует обучение покрытию средствами целочисленного программирования. Более поздняя 4 версия программы CLIP достигла точности 86.1%, совпадающей с точностью ВКФ-системы

Второй массив Mushrooms - данные из определителя грибов Северной Америки [71], 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.
- 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, описывающие различные характеристики грибов (цвет шляпки, форма шляпки, запах, форма ножки, ..., цвет спор, места произрастания,частота встречаемости и т.п.). Эти признаки - номинальные, принимающие одно из нескольких значений.
- ВКФ-система закодировала (с использованием алгоритма 1) 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).

- Время работы ВКФ-системы с абдуктивным уточнением (по 80 ВКФ-гипотез дают 100% точность) превышает время работы без нее (80/150 ВКФ-гипотез, соответственно) более, чем на два порядка.

Основные выводы

1. Познавательные процедуры логико-комбинаторного ДСМ-метода (индуктивного обобщения, абдукции и предсказания по аналогии), основанные на сходстве, для поиска причинно-следственных зависимостей в сложно-структурированных данных допускают вероятностные варианты.

2. Абдукция не нуждается в предварительном расширении множества вероятностно порожденных ВКФ-гипотез с помощью алгоритма 7 абдуктивного уточнения, так как может быть успешно заменена увеличенным числом порождаемых ВКФ-гипотез на этапе индукции.

3. Оценка необходимого числа ВКФ-гипотез для надежного предсказания важных объектов (теорема 4.1) превращает ВКФ-метод в метод статистического машинного обучения (с указанием степени надежности выводов).

4. Программная ВКФ-система, реализующая метод машинного обучения, основанного на теории решеток, великолепно распараллеливается.

5. Применение ВКФ-системы к массиву SPECT Hearts из репозитория данных для тестирования алгоритмов машинного обучения продемонстрировало преимущества нашего подхода над

некоторыми другими алгоритмами комбинаторного анализа данных.

6. Успешное применение ВКФ-системы к массиву Mushrooms (8124 объекта) из этого же репозитория подтверждает возможность ее применения к обучающим выборкам большого размера.

Заключение

Результаты, выносимые на защиту

1. Оценка эффективности ленивых вычислений на шаге индукции в теореме 1.2.

2. Оценка (теорема 2.2) асимптотической вероятности появления фантомного сходства при наличии контр-примеров. Доказательство этой теоремы оценивает скорость сходимости к пределу.

3. Явный вид производящих функций (теоремы 2.3 и 2.4) для вероятности возникновения фантомного сходства при фиксированном и произвольном числе контр-примеров.

4. Теорема 3.3 об изменении вероятностей множеств эргодических состояний для спаривающей цепи Маркова, остановленной с верхней границей по r предварительным прогонам.

5. Теорема 3.4 о среднем времени склеивания и теорема 3.5 о сильной концентрации времени склеивания около его среднего для случая Булеана (множества всех подмножеств признаков).

6. Верхняя оценка (3.15) (из теоремы 3.8) времени перемешивания и теорема 3.9 об асимптотической точности этой оценки для случая Булеана.

7. Теорема 4.1 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 вероятностно-комбинаторного метода обучения, основанного на операции сходства.

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. Абдукция не нуждается в расширении посредством процедуры абдуктивного уточнения множества ВКФ-гипотез, так как с успехом может быть заменена увеличением числа порождаемых гипотез на этапе индукции.

11. Оценка необходимого числа ВКФ-гипотез для надежного предсказания важных объектов превращает ВКФ-метод в метод статистического машинного обучения.

12. Программная ВКФ-система продемонстрировала хорошую балансировку нагрузки по вычислительным узлам при многопоточной реализации.

13. 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.

Направления дальнейших исследований

Теперь мы сформулируем открытые проблемы, разрешение которых позволит улучшить понимание вероятностных когнитивных процедур, основанных на операции сходства, для обнаружения причинно-следственных зависимостей в сложно-структурированных данных:

- Получить оценку среднего времени склеивания для спаривающей цепи Маркова в случае произвольного контекста. Полезно указать, что метрика Хэмминга между верхним и нижним кандидатами не является функцией Ляпунова (может возрастать). Соответствующий пример приводится в параграфе 3.3.

- Исследовать вопрос о времени перемешивания для монотонной цепи Маркова в случае произвольного контекста. Следует отметить, что в частном случае Булеана подобный результат был доказан нами в параграфе 3.5.

- Исследовать асимптотическую вероятность возникновения фантомного сходства, когда число контр-примеров растет, а число признаков сохраняется. Автор надеется, что изложенные в параграфе 2.3 главы 2 производящие функции окажутся при этом полезными.

Список сокращений и условных обозначений

АФП - анализ формальных понятий

ДСМ - Джон Стьюарт Милль (английский философ, экономист и логик)

ИАД - интеллектуальный анализ данных

п.ф. - производящая функция (последовательности чисел)

р.в. - распределение вероятностей

с.в. - случайная величина

ц.с.в. - целочисленная с.в.

CbODown - операция «Замыкай-по-одному-вниз»

CbOUp - операция «Замыкай-по-одному-вверх»

$\boldsymbol{D}$ - дисперсия с.в.

$\boldsymbol{E}$ - математическое ожидание с.в.

$\boldsymbol{P}$ - вероятность (включая условную) случайного события

Словарь терминов

ВКФ-гипотеза - кандидат, фрагмент которого не включается ни в один контр-пример

ВКФ-метод - вероятностно-комбинаторный формальный метод машинного обучения, назван так в честь В.К. Финна

ВКФ-система - программная система, реализующая вероятностные алгоритмы ВКФ-метода

ДСМ-гипотеза - кандидат, имеющий не менее двух родителей (в разных вариантах должен удовлетворять дальнейшим логическим условиям, например, запрету контр-примеров)

ДСМ-метод - логико-комбинаторный метод ИАД, созданный группой исследователей под руководством проф. В.К. Финна

ДСМ-система - программная система, реализующая синтез познавательных процедур ДСМ-метода

запрет контр-примеров - условие, что фрагмент никакого кандидата не может включаться в описание никакого контр-примера

кандидат - пара, состоящая из списка родителей и фрагмента - то же самое, что и формальное понятие в АФП

контр-пример - объект, не обладающий целевым свойством

обучающий пример - объект, обладающий целевым свойством

родители кандидата - множество всех обучающих примеров, содержащих фрагмент этого кандидата

формальный контекст - множество всех обучающих примеров, заданных битовыми строками

фрагмент кандидата - множество всех общих признаков у всех объектов-родителей этого кандидата

Литература

1. **Айзерман М.А., Браверманн Э.М., Розоноэр Л.И.** *Метод потенциальных функций в теории обучения машин.* – М.: Наука. – 1970. – 384 с.

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

3. **Аншаков 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

4. **Блинова В.Г.** О результатах применения ДСМ-метода порождения гипотез к задачам анализа связи «структура химических соединений => биологическая активность» // *Научная и техническая информация, Сер. 2.* – 1995. – № 5. – С. 14–17

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

6. **Вапник В.Н.** *Восстановление зависимостей по эмпирическим данным.* – М.: Наука, Гл. ред. физ.–мат. лит. – 1979. – 448 с.

7. **Вапник В.Н., Червоненкис А.Я.** О равномерной сходимости частот появления событий к их вероятностям // *Теория вероятностей и ее применения.* Т. 16, Вып. 2. – 1971. – С. 264–279

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

9. **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

10. **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

11. **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

12. **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

13. **Виноградов Д.В.** Вероятностно-комбинаторный подход к автоматическому порождению гипотез // В кн.: *Гуманитарные чтения РГГУ – 2014.* – М.: РГГУ. – 2015. – С. 771-775

14. **Виноградов Д.В.** Вероятность порождения случайного ДСМ-сходства при наличии контр-примеров // *Научная и техническая информация,* Сер. 2. – 2015. – № 3. – С. 1–5

15. **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

16. **Виноградов Д.В.** Предельная вероятность порождения случайного сходства при наличии контр-примеров // *Научная и техническая информация, Сер. 2.* – 2017. – № 2. – С. 17–19

17. **Виноградов Д.В.** Эффективность ленивых вычислений для поиска сходств в ВКФ-системе // *Научная и техническая информация*, Сер. 2. – 2017. – № 4. – С. 19–23

18. **Виноградов Д.В.** Анализ результатов применения ВКФ-системы: успехи и открытая проблема // *Научная и техническая информация*, Сер. 2. – 2017. – № 5. – С. 1–4

19. **Виноградов Д.В.** ВКФ-метод интеллектуального анализа данных: обзор результатов и открытых проблем // *Искусственный интеллект и принятие решений.* – 2017. – № 2. – С. 9–16

20. **Виноградов Д.В.** Надежность предсказания по аналогии // *Научная и техническая информация*, Сер. 2. – 2017. – № 7. – С. 11–15

21. **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

22. **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

23. **Виноградов Д.В.** Учет предварительных оценок скорости порождения сходств спаривающей цепью Маркова // *Информатика и ее применения.* – 2018. – № 1. – С. 50–55

24. **Виноградов Д.В.** О представлении объектов битовыми строками для ВКФ-метода // *Научная и техническая информация, Сер. 2.* – 2018. – № 5. – С. 1–4

25. **Винокурова Л.В., Агафонов М.А., Варванина Г.Г., Финн В.К., Панкратова Е.С., Добрынин Д.А.** Применение интеллектуальной системы типа ДСМ для анализа клинических данных // *Российский биотерапевтический журнал.* – 2014. – № 9. – С. 57–60

26. **Вьюгин В.В.** *Математические основы теории машинного обучения и прогнозирования.* – М.: МЦНМО, 2013. – 390 с.

27. **Голубов Б.И., Ефимов А.В., Скворцов В.А.** *Ряды и преобразования Уолша: теория и применения.* – М.: Наука, 1987. – 346 с.

28. **Грин Д., Кнут Д.** *Математические методы анализа алгоритмов.* Пер. с англ. – М.: Мир, 1987. – 120 с.

29. **Гусакова С.М.** Подход к решению задачи атрибуции исторических источников с помощью ДСМ-метода // *Новости искусственного интеллекта.* – 2004. – № 3. – С. 42–49

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

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

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

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

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

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

36. **Кемени Дж., Снелл Дж.** *Конечные цепи Маркова.* Пер. с англ. – М.: Наука, Гл. ред. физ.–мат. лит. – 1970. – 272 с.

37. **Кемени Дж., Снелл Дж., Кнепп А.** *Счетные цепи Маркова.* Пер. с англ. – М.: Наука, Гл. ред. физ.–мат. лит. – 1987. – 416 с.

38. **Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К.** *Алгоритмы: построение и анализ, 2-е изд.* Пер. с англ. – М.: Вильямс, 2005. – 1296 с.

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

40. **Кузнецов С.О.** Быстрый алгоритм построения всех пересечений объектов из нижней полурешетки // *Научная и техническая информация, Сер. 2.* – 1993. – № 1. – С. 17–20

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

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

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

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

45. **Минский М., Пейперт С.** *Персептроны.* Пер. с англ. – М.: Мир, 1971. – 261 с.

46. **Михеенкова М.А.** ДСМ-метод правдоподобных рассуждений как средство анализа социального поведения // *Известия РАН, Сер. «Теория и системы управления».* – 1997. – № 5. – С. 62–70

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

48. **Опарышева А.С.** *Поиск случайных сходств в реальных массивах данных.* Выпускная квалификационная работа бакалавра по направлению подготовки 45.03.04 (Науч.рук.: **Виноградов Д.В.**). – М.: РГГУ, 2018. – 33 с.

49. **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

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

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

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

53. **Серр Ж.-П.** *Линейные представления конечных групп.* Пер. с франц. – М.: Мир, 1970. – 132 с.

54. **Сидорова Е.Ю.** *Экспериментальная реализация вероятностных алгоритмов для поиска ДСМ-сходств.* Дипломная работа по направлению подготовки 036000 (Науч.рук.: **Виноградов Д.В.**). – М.: РГГУ, 2013. – 41 с.

55. **Феллер В.** *Введение в теорию вероятностей и ее приложения.* В 2–х томах. Т. 1: Пер. с англ. – М.: Мир, 1984. – 528 с.

56. **Финн 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

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

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

59. **Шлезингер М.И., Главач В.** *Десять лекций по статистическому и структурному распознаванию.* – Киев: Наукова думка, 2004. – 545 с. [Accessed: September-15-2018] `http://irtc.org.ua/image/Files/Schles/esh10_full.pdf`

60. **Якимова Л.А.** *Реализация ленивой схемы вычислений сходств в ВКФ-методе.* Выпускная квалификационная работа бакалавра по направлению подготовки 45.03.04 (Науч.рук.: **Виноградов Д.В.**). – М.: РГГУ, 2018. – 33 с.

61. **Anshakov, O.M., V.K. Finn, and D.V. Vinogradov.** Logical means for plausible reasoning of JSM-type // В кн.: *Многозначные логики и их применения* (ред. В.К. Финн). Т. 2: *Логики в системах искусственного интеллекта.* – М.: Эдиториал УРСС, – 2008. – С. 226–235

62. **Davey, B.A. and H.A. Priestley.** *Introduction to Lattices and Order.* 2nd eds. – Cambridge: Cambridge University Press, 2002. – 298 pp.

63. **Diaconis, Persi.** *Group representations in probability and statistics.* IMS Lecture Notes – Monograph Series Vol. 11.– Hayward (CA): Institute of Mathematical Statistics, 1988. – 198 pp.

64. **Ehrenfest, Paul and Tatiana Ehrenfest.** *The Conceptual Foundations of the Statistical Approach in Mechanics.*– NY: Cornell University Press, 1959. – 128 pp.

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

66. **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

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

68. **Häggström, Olle.** *Finite Markov Chains and Algorithmic Applications.* – Cambridge: Cambridge University Press, 2002. – 124 pp.

69. **den Hollander, Frank.** *Probability Theory: the Coupling Method.* 3rd Draft. – Leiden, 2012.– 73 pp. [Accessed: September-15-2018] `http://websites.math.leidenuniv.nl/probability/lecturenotes/CouplingLectures.pdf`

70. **Levin, David A., Yuval Peres, and Elizabeth L. Wilmer.** *Markov Chains and Mixing Times.* – Providence (RI): AMS, 2009.– 387 pp. [Accessed: September-15-2018] `http://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf`

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

72. **Montenegro, Ravi and Prasad Tetali.** Mathematical Aspects of Mixing Times in Markov Chains. // *Foundations and Trends in Theoretical Computer Science*, Vol. 1, – Issue 3. – 2001. – p. 1–121

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

74. **Sinclair, A.J.** *Algorithms for random generation and counting.* Advances in Theoretical Computer Science. – Boston: Birkäuser, 1993. – 284 pp.

75. **Valiant, L.G.** A theory of learnable // *Communications of the ACM*, Vol. 27, Issue 11. – 1984. – p. 1134–1142

76. **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

77. **Vinogradov, D.V.** VKF-method of hypotheses generation // *Communications in Computer and Information Science*, Vol. 436. – 2014. – p. 237–248

78. **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

79. **Vinogradov, D.V.** Machine learning based on similarity operation // *Communications in Computer and Information Science*, Vol. 934. – 2018. – p. 46–59

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