Федеральный исследовательский центр
«Информатика и Управление»
Российской академии наук

На правах рукописи

Ишкина Шаура Хабировна

Комбинаторные оценки переобучения пороговых решающих правил

1.2.1 – Искусственный интеллект и машинное обучение

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
д. ф.-м. н., проф. РАН
К. В. Воронцов

Москва – 2025

Оглавление

Введение 4

Глава 1. Достигаемые верхние оценки обобщающей способности прямых последовательностей классификаторов 15

1.1. Основные определения 16

1.2. Прямые последовательности классификаторов 18

1.3. Постановка задачи 23

1.4. Финитный метод обучения 23

1.5. Переобучение произвольного семейства 26

1.6. Переобучение прямой последовательности 30

1.7. Алгоритм вычисления оценок обобщающей способности прямой последовательности 41

1.8. Выводы к первой главе 42

Глава 2. Исследование завышенности существующих оценок обобщающей способности пороговых решающих правил 44

2.1. Обзор известных оценок вероятности переобучения 44

2.2. Обзор известных оценок полного скользящего контроля 46

2.3. Оценка частоты ошибок на контрольной выборке 47

2.4. Вычислительные эксперименты 49

2.5. Выводы ко второй главе 52

Глава 3. Применение комбинаторных оценок при планировании трассерных исследований в нефтегазовых месторождениях . . 54

3.1. Планирование трассерных исследований с применением методов машинного обучения 55

3.2. Явление переобучения в деревьях решений 62

3.3. Постановка задачи 64

3.4. Предлагаемый критерий 64
3.5. Псевдокод алгоритма 65
3.6. Тестирование подхода 66
3.7. Выводы к третьей главе 70

Глава 4. Суррогатное моделирование для вычисления оценок обобщающей способности пороговых решающих правил 74
4.1. Вклад порогового классификатора в переобучение семейства . . 75
4.2. Суррогатное моделирование 77
4.3. Анализ суррогатных моделей 82
4.4. Обсуждение результатов 87
4.5. Выводы к четвертой главе 90

Заключение 91

Список иллюстраций 93

Список таблиц 95

Список обозначений 96

Публикации автора по теме диссертации 98

Список литературы 101

Приложение А. Акт внедрения 108

Введение

Актуальность темы исследования. Диссертация посвящена теме построения верхних оценок обобщающей способности одномерных пороговых решающих правил.

При решении задачи обучения на основании обучающей выборки объектов, часто называемой обучением по прецедентам, строится алгоритм, восстанавливающий зависимость выходных переменных от входных на объектах из обучающей выборки. В задаче классификации выходная переменная одна и принимает бинарные значения, а алгоритмы называются классификаторами. Для успешного применения построенного классификатора он должен иметь высокую обобщающую способность, то есть хорошо работать на произвольных объектах, не обязательно входящих в обучение. Если же качество классификатора на независимой выборке, называемой контрольной, оказывается значительно хуже, чем на обучающей выборке, то говорят, что произошло переобучение.

Получение оценок обобщающей способности семейства классификаторов на основе информации об обучающей выборке и структуре семейства с публикации [66] остается одной из основных задач теории статистического обучения. Завышенность полученных оценок может приводить к неоптимальному выбору структурных параметров [27, 35, 42]. Кроме того, завышенные оценки не дают возможности исследовать явление переобучения, оценивать и контролировать его значения при решении реальных задач.

Степень разработанности темы исследования. В конце 70-х гг. XX в. советские ученые В. Н. Вапник и А. Я. Червоненкис сформулировали основные статистические проблемы обучения в терминах проблемы минимизации среднего риска, т. е. вероятности ошибки классификатора на новом объекте, и предложили методы оценки среднего риска по эмпирическим данным [65, 67]. Вапник и Червоненкис получили равномерные по семействам классификаторов оценки, связывающие вероятность уклонения среднего риска от эмпирического с

длиной обучающей выборки и сложностью семейства, над которыми минимизируется средний риск. Этот фундаментальный результат активно используется и сегодня.

Однако оценки Вапника–Червоненкиса являются завышенными. В работе [55] показано, что они бывают завышены на 6–12 порядков и плохо согласуются с результатами экспериментов. В этой же работе исследуются причины завышенности оценок, из которых основной является независимость оценок от конкретной выборки. Оценка Вапника–Червоненкиса универсальна и, следовательно, является оценкой худшего случая.

Теория статистического обучения продолжает активно развиваться, последователи теории занимаются повышением точности равномерных оценок с учетом особенностей данных и конкретных алгоритмов классификации [19, 20, 51, 54]. Получены более тонкие оценки, которые зависят от свойств отношения частичного порядка на множестве вектор-столбцов матрицы ошибок [31]. Среди плодотворных подходов можно выделить оценки, адаптирующиеся к данным и использующие понятие Радемахеровской сложности, предложенной в 1999 г. В. Колчинским [40].

В качестве характеристик обобщающей способности используются функционалы вероятности переобучения и полного скользящего контроля [38].

В комбинаторной теории переобучения [68, 69], предложенной К. В. Воронцовым, вероятностью переобучения называют долю разбиений конечного множества объектов на обучающую и контрольную выборки фиксированной длины, при которых произошло переобучение. Данное определение ранее появлялось в [31] для частного случая контрольной выборки, состоящей из одного объекта.

Точность эмпирических оценок функционалов обобщающей способности, полученных методом Монте–Карло, зависит от числа случайных разбиений. Вычисление оценок по определению требует экспоненциального по общему количеству объектов перебора всех возможных разбиений. Но для некоторых модель-

ных семейств классификаторов удается аналитически вычислить достигаемые верхние оценки вероятности переобучения. К настоящему времени достигаемые верхние оценки получены для слоев и интервалов булева куба, многомерных сетей [18], хэмминговых шаров и некоторых их разреженных подмножеств [80]. Разработан теоретико-групповой подход [26], который позволяет получать достигаемые верхние оценки для семейств с произвольными симметриями.

В [62] предложен способ аппроксимации вероятности переобучения стандартных методов классификации (нейронных сетей, решающих деревьев, ближайшего соседа) на реальных задачах с помощью монотонных сетей подходящей размерности. Оценки переобучения могут использоваться в качестве критерия отбора признаков при построении элементарных конъюнкций в логических алгоритмах классификации [59] или в качестве критерия ветвления в решающих деревьях [18].

В комбинаторной теории для вероятности переобучения была получена оценка расслоения–связности [59], учитывающая особенности способа построения классификатора по обучающей выборке, а также локальные свойства семейства классификаторов – эффекты расслоения и связности [56]. Благодаря расслоению, классификаторы с высокой вероятностью ошибки вносят пренебрежимо малый вклад в переобучение. Благодаря связности, у классификаторов с близкими векторами ошибок резко снижается вклад в переобучение.

В [73] получены условия, при которых оценка расслоения–связности является точной. Им удовлетворяют, в частности, монотонные и унимодальные цепи классификаторов [57]. В практических задачах статистического обучения такие цепи могут порождаться элементарными пороговыми правилами, используемых в таких алгоритмах классификации, как решающие деревья, логические закономерности [74], алгоритмы вычисления оценок [75], а также при построении линейных классификаторов методом покоординатной оптимизации. Но при этом делается предположение о существовании безошибочного правила, практически не выполнимое в реальных задачах. В общем случае пороговые правила

порождают семейства классификаторов, называемые прямыми последовательностями.

Ранее для них были известны лишь верхние оценки ожидаемой частоты ошибок на контрольной выборке [72] в частном случае, когда признак принимает попарно различные значения на объектах. Различные уточнения оценок расслоения–связности, например, учитывающие попарную конкуренцию между классификаторами [70] или послойную кластеризацию множества классификаторов [83, 84], также остаются завышенными для прямых последовательностей. Однако завышенность верхних оценок остается неизученной.

Цель диссертационной работы. Построение достигаемых верхних оценок обобщающей способности одномерных пороговых решающих правил в рамках комбинаторной теории переобучения, где в качестве характеристик обобщающей способности рассматриваются функционалы вероятности переобучения, полного скользящего контроля и ожидаемой переобученности. Исследование завышенности известных оценок обобщающей способности. Применение полученных оценок в практических задачах.

Научная новизна. Рассмотрены методы минимизации эмпирического риска и максимизации переобученности и показано, что они обладают свойством финитности. Для финитного метода обучения и произвольного семейства классификаторов доказаны теоремы о представлении достигаемых верхних оценок обобщающей способности в виде произведения числа разбиений двух непересекающихся множеств объектов генеральной совокупности.

Для прямых последовательностей классификаторов, порождаемых элементарными пороговыми правилами при варьировании параметра порога, доказаны теоремы и реализован алгоритм полиномиальной сложности для вычисления достигаемых верхних оценок обобщающей способности. Алгоритм основан на рекуррентном подсчете числа допустимых траекторий при блуждании по трехмерной сетке между двумя заданными точками с ограничениями специального вида.

Получен новый алгоритм построения дерева решений, в котором в качестве критерия выбора атрибута для разделения узла дерева решений используются достигаемые верхние оценки полного скользящего контроля и ожидаемой переобученности пороговых решающих правил.

Построена суррогатная модель для быстрого вычисления приближенных оценок обобщающей способности семейства пороговых решающих правил с высокой точностью.

Теоретическая и практическая значимость. Доказаны теоремы о вычислении достигаемых верхних оценок обобщающей способности прямых последовательностей классификаторов, порождаемых пороговыми правилами над одномерным признаком при варьировании параметра порога. В рамках комбинаторного подхода до сих пор не удавалось получать достигаемые верхние оценки обобщающей способности для данного семейства в общем случае. Достигаемые верхние оценки были известны только для частных случаев задач классификации, где значения одномерного признака на классифицируемых объектах были попарно различны.

Предложенные в работе методы вычисления оценок обобщающей способности применимы в качестве критерия отбора признаков при построении алгоритмов классификации, в частности, в решающих деревьях, логических закономерностях, и при построении линейных классификаторов методом покоординатной оптимизации. Предложенный в работе способ построения программы трассерных исследований применим для повышения эффективности трассерных исследований в нефтегазовых месторождениях.

Положения, выносимые на защиту:

1. Доказаны теоремы о представлении достигаемых верхних оценок обобщающей способности произвольного семейства классификаторов в виде произведения числа разбиений двух непересекающихся множеств объектов генеральной совокупности для финитного метода обучения.

2. Доказаны теоремы и разработан алгоритм полиномиальной сложности для вычисления достигаемых верхних оценок обобщающей способности прямых последовательностей классификаторов, порождаемых одномерными пороговыми решающими правилами при варьировании параметра порога, для финитного метода обучения.

3. Разработан алгоритм для построения программы трассерных исследований с применением деревьев решений.

4. Разработан алгоритм построения дерева решений с использованием полученных достигаемых верхних оценок полного скользящего контроля и ожидаемой переобученности в качестве критерия выбора атрибута в узле.

5. Разработан алгоритм вычисления приближенных оценок обобщающей способности одномерных пороговых решающих правил с использованием суррогатных моделей.

Степень достоверности и апробация результатов. Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных методов на прикладной задаче классификации пар скважин при составлении программы трассерных исследований в нефтегазовых месторождениях; публикациями результатов исследования в рецензируемых научных изданиях, в том числе рекомендованных ВАК, регистрацией патента на изобретение и актом внедрения основных результатов (см. Приложение А).

Основные результаты диссертации докладывались на следующих конференциях:

1. Международная школа-конференция «Фундаментальная математика и ее приложения в естествознании», 2023. [5]

2. Международная научно-практическая конференция «Цифровая трансформация в нефтегазовой отрасли», 2023. [6]

3. Межрегиональная школа-конференция «Теоретические и экспериментальные исследования нелинейных процессов в конденсированных средах», 2021. [16]

4. Всероссийская молодежная научно-практическая конференция «Геолого-геофизические исследования нефтегазовых пластов», 2021.[2]

5. Международная конференция «Управление развитием крупномасштабных систем», 2016. [3]

6. Международная конференция «Intelligent Data Processing», 2016. [7]

7. Всероссийская конференция «Математические методы распознавания образов», 2015. [8]

8. Всероссийская конференция «Математические методы распознавания образов», 2013. [13]

Публикации. Результаты диссертации содержатся в 16 публикациях. В изданиях из списка ВАК представлено 8 публикаций, в том числе 1 патент на изобретение [1, 4, 9–12, 14, 15]. Работы [1, 4, 9, 10, 15] индексируются SCOPUS, Web of Science. Отдельные результаты включались в отчёты по проектам РФФИ (№ 15-37-50350 мол_нр и № 14-07-00847), Правительства РФ (№ 075-15-2019-1926). Список публикаций приведен в конце автореферата и диссертации.

Личный вклад автора. Результаты получены самостоятельно под научным руководством д.ф.-м.н. К. В. Воронцова. Личный вклад автора в работы, выполненные совместно с соавторами, заключается в следующем:

- в работе [1] сформулирована и доказана теорема о вычисления оценки функционала ожидаемой переобученности семейства и оценки частоты ошибок метода минимизации эмпирического риска на контрольной выборке для семейства одномерных пороговых решающих правил, проведены вычислительные эксперименты;

- в работе [14] разработан алгоритм построения программы трассерных исследований с использованием методов машинного обучения, проведены вычислительные эксперименты;

- в работе [11] реализован алгоритм построения дерева решений с использованием комбинаторных оценок для выбора атрибута в узле дерева, проведены вычислительные эксперименты и доказана статистическая значимость результатов.

- в работе [12] разработан алгоритм интерпретации исследований скважин методом эхометрирования с применением методов машинного обучения.

- в работе [4] разработан алгоритм интерпретации исследований скважин на неустановившихся режимах с применением методов машинного обучения, проведено тестирование алгоритма.

- в работе [15] разработан алгоритм «виртуального расходомера» на основе стекинга моделей машинного обучения, проведены вычислительные эксперименты.

Соответствие паспорту специальности. Результаты диссертационного исследования соответствуют паспорту специальности 1.2.1 «Искусственный интеллект и машинное обучение», а именно: пункту 1 «Естественно-научные основы и методы искусственного интеллекта», пункту 2 «Исследования в области оценки качества и эффективности алгоритмических и программных решений для систем искусственного интеллекта и машинного обучения. Методики сравнения и выбора алгоритмических и программных решений при многих критериях».

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, списка иллюстраций, списка таблиц, списка литературы и приложения. Общий объем диссертации составляет 108 страниц, из них

92 страницы текста, включая 15 рисунков и 6 таблиц. Библиография включает 85 наименований на 10 страницах.

Краткое содержание работы по главам:

В первой главе проводится теоретическое исследование и доказательство теорем для вычисления достигаемых верхних оценок обобщающей способности в семействах классификаторов в рамках комбинаторной теории переобучения. Вводится понятие финитного метода обучения, для которого в случае произвольного семейства классификаторов доказываются теоремы о представлении достигаемых верхних оценок в виде произведения числа разбиений двух непересекающихся множеств объектов генеральной совокупности. Доказывается, что свойством финитности обладают рассматриваемые в данной методы минимизации эмпирического риска и максимизации переобученности.

Исследуется явление переобучения одномерных пороговых решающих правил при выборе порога. Проводятся вычислительные эксперименты, которые показывают, что переобучение семейства зависит от формы графика числа ошибок при варьировании порогового значения. Доказываются теоремы о вычислении достигаемых верхних оценок обобщающей способности семейства одномерных пороговых решающих правил. Приводится псевдокод алгоритма вычисления достигаемых оценок данного семейства и доказывается его полиномиальная вычислительная сложность.

Результаты первой главы опубликованы в работах [1] и [9].

Во второй главе исследуется завышенность известных оценок обобщающей способности для пороговых решающих правил по сравнению с достигаемыми верхними оценками, рассчитанными с помощью алгоритма, описанного в первой главе. Оценки вероятности переобучения (Вапника–Червоненкиса, расслоения–связности и Соколова) и оценки частоты ошибок на контрольной выборке на основе Радемахеровской сложности оказываются завышены. Показано, что оценки Гуза для величины полного скользящего контроля обладают высокой точностью, откуда следует вывод о применимости данных оценок в

прикладных задачах в частных случаях.

Результаты второй главы опубликованы в работе [1].

Третья глава посвящена задаче применения комбинаторных оценок в прикладной задаче планирования трассерных исследований в нефтегазовых месторождениях. Предлагается алгоритм построения программы исследований, согласно которому пара скважин включается в программу на основе ответа классификатора дерева решений. Ставится задача повышения обобщающей способности дерева решений. Исследуются причины возникновения переобучения классификатора, одной из которых является смещенность существующих критериев выбора атрибута при построении разбиения в узле. Предлагается модификация алгоритма построения дерева, согласно которой в качестве критерия используются достигаемые верхние оценки переобучения пороговых решающих правил. Для вычисления критерия применяется алгоритм, разработанный в первой главе. Проводятся вычислительные эксперименты на промысловых данных, которые показывают статистически значимое повышение обобщающей способности дерева решений.

Результаты третьей главы опубликованы в работах [11] и [14].

Для устранения ограничения алгоритма, описанного в первой главе, связанного с большой вычислительной сложностью, которое не дает использовать его на выборках большого объема, **в четвертой главе** решается задача разработки алгоритма для быстрого вычисления приближенных оценок путем построения суррогатной модели. Описывается процесс сбора обучающей выборки для модели, которая состоит из пар «объект, ответ», и каждым объектом является семейство одномерных пороговых решающих правил, ответом – достигаемая верхняя оценка обобщающей способности семейства. На основе имеющихся исследований оценок обобщающей способности, проведенных в рамках комбинаторной теории переобучения, формируется перечень признаков, которые описывают объекты выборки. Рассматриваются модели различной структуры, наилучшей по результатам тестирования выбрана модель нейронной се-

ти с MAPE=2.8%. По итогам анализа значимости признаков показано, что при построении оценок переобучения недостаточно учитывать только количество классификаторов и минимальное число ошибок классификаторов, необходимо использовать внутреннюю структуру семейства (расслоение по числу ошибок) и взаимосвязь между классификаторами (связность). Показано, что использование модели позволяет сократить время вычисления оценок обобщающей способности на несколько порядков с $O(L^5)$ до $O(L^2)$ по сравнению с алгоритмом, описанным в первой главе, откуда следует вывод о практической значимости разработанного подхода в задачах отбора признаков при построении деревьев решений, нейронных сетей и в алгоритмах бустинга для контроля переобучения.

Результаты четвертой главы опубликованы в работе [10].

Благодарность. Автор признателен научному руководителю, профессору РАН Воронцову Константину Вячеславовичу, за постановки и обсуждение задач и внимание к работе, профессору Стрижову Вадиму Викторовичу за ценные замечания при подготовке текста диссертационной работы и руководителю в ООО «РН-БашНИПИнефть», начальнику управления по моделированию и анализу исследований скважин и пластов, Давлетбаеву Альфреду Ядгаровичу и коллегам за помощь в реализации разработанных алгоритмов в прикладных задачах нефтегазовой отрасли.

Глава 1

Достигаемые верхние оценки обобщающей способности прямых последовательностей классификаторов

Математическая модель задачи классификации как задачи принятия решений в условиях неполноты информации формулируется следующим образом. Дана бинарная матрица, строки которой соответствуют объектам, столбцы — классификаторам, называемым также правилам принятия решений или гипотезами. В ячейке матрицы находится единица тогда и только тогда, когда данный классификатор ошибается на данном объекте. Из множества $\mathbb{X}$ всех строк матрицы случайно и равновероятно выбирается наблюдаемая обучающая выборка — подмножество $X \subset \mathbb{X}$ фиксированной мощности. Затем из множества $\mathbb{A}$ всех столбцов матрицы выбирается классификатор с минимальной частотой ошибок на X.

В данной главе рассматриваются бинарные матрицы со следующим свойством: все строки, по которым отличаются соседние столбцы, различны. Матрица с указанным свойством однозначно определяет семейство классификаторов, называемое прямой последовательностью.

Доказывается теорема о бинарном соответствии между семействами прямых последовательностей и одномерными пороговыми классификаторами.

Решается задача построения достигаемых верхних оценок обобщающей способности данного семейства классификаторов. Предлагается алгоритм полиномиальной сложности для вычисления вероятности переобучения произвольной прямой последовательности при выборе классификатором описанным выше способом минимизации ошибок. Алгоритм основан на рекуррентном подсчете числа допустимых траекторий при блуждании по трехмерной сетке между дву-

мя заданными точками с ограничениями специального вида.

1.1. Основные определения

Задано конечное множество $\mathbb{X} = \{x_1, \ldots, x_L\}$, элементы которого называются *объектами*, и конечное множество $\mathbb{A}$, элементы которого называются *классификаторами*. Множество $\mathbb{A}$ называется *семейством классификаторов*.

Задана функция $I\colon \mathbb{A} \times \mathbb{X} \to \{0, 1\}$, называемая *индикатором ошибки*. Если $I(a, x) = 1$, то говорят, что классификатор a *допускает ошибку* на объекте x. Бинарная матрица $\big(I(a, x)\colon x \in \mathbb{X}, a \in \mathbb{A}\big)$ размера $|\mathbb{X}|\times|\mathbb{A}|$ называется *матрицей ошибок*.

Предполагается, что каждому классификатору $a \in \mathbb{A}$ взаимно однозначно соответствует его вектор ошибок $(I(a, x_i))_{i=1}^{L}$, то есть в матрице ошибок не может быть двух равных столбцов. Будем считать, что порядок строк в матрице ошибок не важен. Договоримся обозначать через a как классификатор, так и его вектор ошибок.

Числом ошибок классификатора a на выборке $X \subset \mathbb{X}$ называется величина

$$n(a, X) = \sum_{x \in X} I(a, x).$$

Частотой ошибок классификатора a на выборке $X \subset \mathbb{X}$ называется величина

$$\nu(a, X) = n(a, X)/|X|,$$

где через $|X|$ обозначен объем выборки X.

Обозначим через $[\mathbb{X}]^{\ell}$ множество всех подмножеств $\mathbb{X}$ мощности $\ell < L$. Подмножества $X \in [\mathbb{X}]^{\ell}$ будем называть *обучающими выборками*, а их дополнения $\bar{X} = \mathbb{X}\backslash X$ — *контрольными выборками*. Введем на множестве $[\mathbb{X}]^{\ell}$ равномерное распределение вероятностей:

$$\mathsf{P}(X) = 1/C_L^{\ell}, \qquad X \in [\mathbb{X}]^{\ell}.$$

Переобученностью классификатора a на разбиении $(X, \bar{X})$ называется величина

$$\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X).$$

Если $\delta(a, X) > \varepsilon$, то будем говорить, что классификатор a переобучен на X.

Методом обучения называется отображение $\mu\colon [\mathbb{X}]^\ell \to \mathbb{A}$, которое каждой обучающей выборке X ставит в соответствие классификатор $a = \mu X$ из семейства $\mathbb{A}$.

Для фиксированного метода обучения μ, семейства классификаторов $\mathbb{A}$, множества $\mathbb{X}$ и объема обучающей выборки ℓ *вероятностью переобучения* называется функционал

$$Q_\varepsilon(\mu, \mathbb{A}, \mathbb{X}, \ell) = \mathsf{P}[\delta(\mu X, X) \geqslant \varepsilon] = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} [\delta(\mu X, X) \geqslant \varepsilon].$$

Здесь и далее квадратные скобки будут использоваться для преобразования логического условия в числовое значение по правилу $[\text{истина}] = 1$, $[\text{ложь}] = 0$.

Полным скользящим контролем (complete cross-validation, CCV) называется функционал, равный математическому ожиданию числа ошибок на контрольной выборке:

$$\mathsf{CCV}(\mu, \mathbb{A}, \mathbb{X}, \ell) = \mathsf{E}\nu(\mu X, \bar{X}) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \nu(\mu X, \bar{X}).$$

Ожидаемой переобученностью называется функционал, равный математическому ожиданию переобученности классификатора, выбранного методом обучения:

$$\mathsf{EOF}(\mu, \mathbb{A}, \mathbb{X}, \ell) = \mathsf{E}\delta(\mu X, \bar{X}) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \nu(\mu X, \bar{X}) - \nu(\mu X, X).$$

Для краткости параметры, от которых зависят данные величины, опускаются.

В данной работе рассматривается метод обучения, *минимизирующий эмпирический риск* (МЭР)

$$\mu X \in M(X) = \operatorname{Arg}\min_{a\in\mathbb{A}} n(a, X).$$

Для получения верхних оценок Q_ε и CCV вводится понятие метода *пессимистичной минимизации эмпирического риска* (ПМЭР)

$$\mu X = \arg\max_{a\in M(X)} n(a, \mathbb{X}).$$

Это метод МЭР, который в случае неоднозначности среди $M(X)$ выбирает классификатор с наибольшим числом ошибок на множестве $\mathbb{X}$ [57].

Другим методом обучения, рассмотренным в данной работе для получения верхних оценок EOF, является метод *максимизации переобученности* (МП):

$$\mu X = \arg\max_{a\in\mathbb{A}} \nu(a, X).$$

Метод МП возникает в задаче комбинаторного вычисления рademахеровской сложности класса решающих правил [39].

Будем считать, что в случае неоднозначности определенные выше методы обучения выбирают классификатор с наибольшим номером. Данное ограничение не влияет на оценку рассматриваемых функционалов обобщающей способности, но далее оно позволит точно вычислить искомые значения.

1.2. Прямые последовательности классификаторов

Рассмотрим множества объектов, по которым различаются соседние классификаторы семейства $\mathbb{A} = \{a_0, \ldots, a_P\}$:

$$G_p = \big\{x \in \mathbb{X} \mid I(a_p, x) \neq I(a_{p+1}, x)\big\}, \quad p = 0, \ldots, P-1. \tag{1.1}$$

Определение 1.1. *Семейство классификаторов называется прямой последовательностью, если множества G_p попарно не пересекаются.*

Заметим, что из определения следует, что порядок классификаторов важен. Действительно, рассмотрим два семейства классификаторов, первое из которых является прямой последовательностью $\mathbb{A} = \{a_0, \ldots, a_P\}$, а второе получается из первого перестановкой классификаторов a_p и a_{p+1} для некоторого p: $\mathbb{A}' = \{a_0, \ldots, a_{p-1}, a_{p+1}, a_p, a_{p+2}, \ldots, a_P\}$. Определим множества G_p по (1.1). Тогда семейство $\mathbb{A}'$ не является прямой последовательностью, поскольку соседние классификаторы a_{p-1} и a_{p+1} различаются по множеству объектов $G_{p-1} \sqcup G_p$, а классификаторы a_{p+1} и a_p – по множеству объектов G_p, то есть эти множества пересекаются.

Определение 1.2. *Прямая последовательность* $\mathbb{A} = \{a_0, \ldots, a_P\}$ *называется прямой цепью, если каждая пара соседних классификаторов различается по одному объекту:* $|G_p| = 1$, $p = 0, \ldots, P-1$. *Число* P *называется длиной прямой цепи* $\mathbb{A}$.

Определение 1.3. *Одномерным пороговым классификатором над множеством* $\mathbb{X} \subset \mathbb{R}$ *называется семейство пороговых правил* $a(x, \theta) = [x \geqslant \theta]$, *где* $\theta \in \mathbb{R}$ – *параметр, называемый порогом.*

Согласно следующей теореме, понятия прямой последовательности и одномерного порогового классификатора являются синонимами.

Теорема 1.1. *Определим множество* V *прямых последовательностей* $\mathbb{A} = \{a_0, \ldots, a_P\}$, *таких, что* $\sum_{p=0}^{P-1} |G_p| = L$, *где* G_p *определены по* (1.1), *и множество* U *одномерных пороговых классификаторов над множеством* $\mathbb{X} = \{x_1, \ldots, x_L\}$ *точек числовой оси, таким, что каждому* x_i *соответствует истинная метка класса* $y_i \in \{0, 1\}$. *Тогда между этими множествами имеется биекция.*

Доказательство. Во множествах V и U объекты определены с точностью до переименования объектов множества $\mathbb{X}$.

Каждый объект $u \in U$ однозначно определяется распределением объектов двух классов $\{0,1\}$ на числовой оси, то есть расположением точек множества $\mathbb{X}$ на оси $\mathbb{R}$ и набором правильных ответов $\{y_1,\ldots,y_L\}$. Значения порогов выбираются так, чтобы они всеми возможными различными способами разбивали множество $\mathbb{X}$ на два класса.

Каждый объект множества V однозначно определяется количеством единиц в векторе a_0, то есть $n(a_0,\mathbb{X})$, и последовательностью пар $(n_0^p, n_1^p)_{p=0}^{P-1}$, где n_0^p – количество нулей в векторе a_p, являющихся единицами в a_{p+1}, и n_1^p – количество единиц в векторе a_p, являющихся нулями в a_{p+1}. При наличии данной информации матрица ошибок $\{a_0,\ldots,a_P\}$ строится следующим образом. Вектор a_0 задается так, что на первых $n(a_0,\mathbb{X})$ позициях стоят единицы, затем нули. Для каждого p последовательно, начиная с $p=0$, вектор a_{p+1} получается из вектора a_p путем инвертирования n_0^p нулей и n_1^p единиц.

Построим отображение $f: U \to V$ следующим образом. Пусть дан объект $u \in U$, то есть набор точек $x_1 \leq \cdots \leq x_L$ и правильных ответов $y_1,\ldots,y_L$. Поставим ему в соответствие прямую последовательность $v = f(u) \in V$.

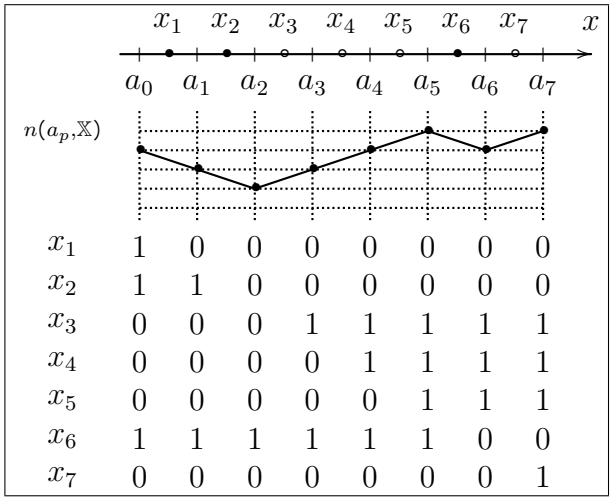
Для этого введём индикатор ошибки $I(a,x_i) = \big[a(x_i,\theta) \neq y_i\big]$. Варьирование θ порождает не более $L+1$ классификаторов с попарно различными векторами ошибок. Они образуют прямую последовательность. Если все объекты x_i попарно различны, $x_1 < x_2 < \cdots < x_L$, то прямая последовательность является прямой цепью.

Отображение f однозначно определяет прямую последовательность по семейству пороговых правил, то есть оно является инъекцией. Докажем, что оно является сюръекцией.

Пусть дана прямая последовательность $v \in V$, то есть величина $n(a_0,\mathbb{X})$ и набор пар $(n_0^p, n_1^p)_{p=0}^{P-1}$. Построим матрицу ошибок $\{a_0,\ldots,a_P\}$. Определим семейство пороговых правил $u \in U$ следующим образом. Поставим в соответствие каждому множеству G_p точки $x_p^1 = \cdots = x_p^{|G_p|}$ и положим, что $x_0^1 < x_1^1 < \cdots < x_{P-1}^1$. Положим $y_p^i = 1$, если $I(a_p, x_p^i) = 0$, и $y_p^i = 0$ в противном случае. Легко

проверить, что построенное семейство u является прообразом v при отображении f, то есть $v = f(u)$. Таким образом, отображение f является биекцией.

Пример 1. *На рис. 1.1 показан пример прямой цепи. По оси x отложены объекты x_i. Правильные решения y_i показаны точками $\circ$ и $\bullet$. Пороги θ выбраны посередине между соседними объектами. Ниже показан график числа ошибок классификаторов и матрица ошибок.*



	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7
x_1	1	0	0	0	0	0	0	0
x_2	1	1	0	0	0	0	0	0
x_3	0	0	0	1	1	1	1	1
x_4	0	0	0	0	1	1	1	1
x_5	0	0	0	0	0	1	1	1
x_6	1	1	1	1	1	1	0	0
x_7	0	0	0	0	0	0	0	1

Рис. 1.1. Пример прямой цепи

Определение 1.4. *Прямая цепь $\mathbb{A} = \{a_0, \ldots, a_P\}$ называется возрастающей (убывающей), если каждый классификатор a_p допускает $m+p$ (соответственно, $m-p$) ошибок на множестве $\mathbb{X}$ при некотором значении m. Прямую цепь $\mathbb{A}$ будем называть монотонной, если она является убывающей или возрастающей.*

Прямая цепь $\mathbb{A}$ может состоять из нескольких участков монотонности. Например, в цепи, показанной на рис. 1.1, имеется четыре участка монотонности: $\{a_0, a_1, a_2\}$ и $\{a_5, a_6\}$ — убывающие, $\{a_2, a_3, a_4, a_5\}$ и $\{a_6, a_7\}$ — возрастающие.

Покажем, что переобучение цепи зависит от геометрической структуры классов.

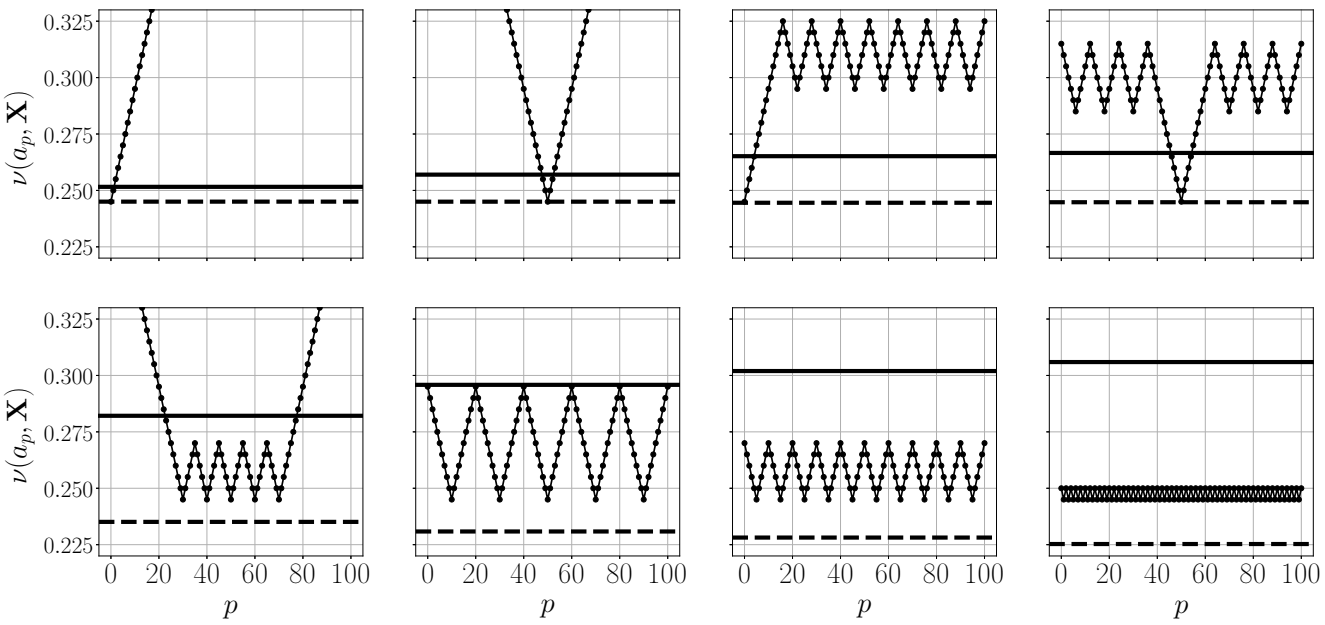


Рис. 1.2. Сравнение переобучения прямых цепей различной формы. По горизонтали отложены номера p классификаторов цепи. Условия эксперимента: $L = 200$, $\ell = 150$, $\varepsilon = 0,05$. Минимальная частота ошибок равна $0,245$.

Эмпирической оценкой функции $\phi(X, \bar{X})$, не зависящей от порядка элементов в выборках X и $\bar{X}$, называется величина $\hat{\mathsf{E}}\phi(X, \bar{X})$, полученная методом Монте–Карло путем усреднения по некоторому случайному подмножеству выборок $N \subset [\mathbb{X}]^\ell$

$$\hat{\mathsf{E}}\phi(X, \bar{X}) = \frac{1}{|N|} \sum_{X \in N} \phi(X, \bar{X}).$$

На рис. 1.2 изображены прямые цепи различной формы. График частоты ошибок классификаторов на полном множестве изображен линией с точкой. Цепи упорядочены по возрастанию количества классификаторов в нижних слоях, где слои определяются по числу ошибок. Пилообразные участки соответствуют шуму в данных, где объекты из разных классов чередуются друг с другом. Чем чаще пила, тем выше уровень шума. Рассматриваются случаи, когда пилообразные участки расположены вдали от границы классов (верхний ряд) и вблизи от границы (нижний ряд).

Горизонтальными линиями показаны эмпирические оценки частоты ошибок метода ПМЭР, вычисленные методом Монте–Карло по 10^5 разбиениям.

Пунктирной линией изображена оценка на обучающей выборке $\hat{\mathsf{E}}\nu(\mu X, X)$ и сплошной линией – оценка на контрольной выборке $\hat{\mathsf{E}}\nu(\mu X, \bar{X})$. Чем больше расстояние между ними, равное $\hat{\mathsf{E}}\delta(\mu X, X)$, тем сильнее переобучается семейство.

Данный эксперимент показывает, что одни семейства переобучаются значительно сильнее, чем другие: переобучение тем выше, чем больше классификаторов находится в нижних слоях семейства и чем более они различны. Эффективное вычисление Q_ε, CCV и EOF непосредственно по определению возможно только при малых $|X| = \ell$. Если ℓ близко к $L/2$, то число слагаемых экспоненциально по L.

Вследствие этого ставится следующая задача.

1.3. Постановка задачи

Для прямой последовательности $\mathbb{A}$ общего вида, методов обучения ПМЭР и МП вычислить достигаемые верхние оценки вероятности переобучения Q_ε, полного скользящего контроля CCV и ожидаемой переобученности EOF за полиномиальное по L время.

1.4. Финитный метод обучения

Пусть дано произвольное подмножество $\mathbb{D} \subseteq \mathbb{X}$ множества $\mathbb{X}$. Каждое разбиение $(X, \bar{X})$ множества $\mathbb{X} = X \sqcup \bar{X}$ индуцирует разбиение $(X \cap \mathbb{D}, \bar{X} \cap \mathbb{D})$ подмножества $\mathbb{D}$. Также любая пара разбиений $(D', \bar{D}')$ и $(D'', \bar{D}'')$ подмножеств $\mathbb{D}' \subseteq \mathbb{X}$ и $\mathbb{D}'' = \mathbb{X} \setminus \mathbb{D}'$ соответственно определяет разбиение $(X, \bar{X})$ множества $\mathbb{X}$ по правилу $X = D' \cup D''$ и $\bar{X} = \bar{D}' \cup \bar{D}''$.

Рассмотрим произвольное семейство классификаторов $\mathbb{A} = \{a_0, \ldots, a_P\}$. Назовем пару классификаторов a и a' *неразличимыми на множестве* $\mathbb{X}' \subset \mathbb{X}$, если $I(a, x) = I(a', x)$ для всех $x \in \mathbb{X}'$.

Пусть на множестве $\mathbb{A} \times \mathbb{A} \times [\mathbb{X}]^\ell$ имеется отношение строгого порядка $a \succ_X a'$. Назовем его *финитным*, если для любых классификаторов $a, a' \in \mathbb{A}$, неразличимых на множестве $\mathbb{X}' \subset \mathbb{X}$, отношение $a \succ_X a'$ не зависит от выбора разбиения множества $\mathbb{X}'$.

Лемма 1.1. *Определенные по следующим правилам отношения порядка являются финитными:*

1. $a_p \succ_X a_i \iff n(a_p, X) \leq n(a_i, X)$;
2. $a_p \succ_X a_i \iff \delta(a_p, X) \geq \delta(a_i, X)$.

Если $i > p$, то неравенства строгие.

Доказательство. Действительно, для любого $X \in [\mathbb{X}]^\ell$ и для любого $\mathbb{X}'$ справедливо равенство $n(a, X) = n(a, X \cap \mathbb{X}') + n(a, X \setminus \mathbb{X}')$. Если классификаторы a и a' неразличимы на множестве $\mathbb{X}'$, то $n(a, \mathbb{X}' \cap X) = n(a', \mathbb{X}' \cap X)$, откуда следует финитность отношения 1.

Для доказательства финитности отношения 2 перепишем переобученность как $\delta(a, X) = \frac{1}{L-\ell} n(a, \mathbb{X}) - \frac{L}{(L-\ell)\ell} n(a, X)$. Тогда утверждение следует из первого пункта.

Запасом ошибок классификатора a относительно a_p на выборке X назовем величину

$$\Delta_p(a, X) = n(a, X) - n(a_p, X). \tag{1.2}$$

Аналогично доказывается справедливость следующей леммы:

Лемма 1.2. *Определенное по следующему правилу отношение порядка является финитным: $a_p \succ_X a_i$ тогда и только тогда, когда выполнено одно из условий:*

1. $\Delta_p(a_i, X) > 0$;

2. $\Delta_p(a_i, X) = 0$ *и* $i < p$ *и* $n(a_i, \mathbb{X}) \leqslant n(a_p, \mathbb{X})$;

3. $\Delta_p(a_i, X) = 0$ *и* $i > p$ *и* $n(a_i, \mathbb{X}) < n(a_p, \mathbb{X})$.

Будем говорить, что на выборке X классификатор a *лучше, чем* a', если $a \succ_X a'$.

Назовем метод обучения $\mu\colon [\mathbb{X}]^\ell \to \mathbb{A}$ *финитным*, если результатом обучения является лучший с точки зрения финитного отношения $\succ_X$ классификатор:

$$a = \mu X \quad \Leftrightarrow \quad a \succ_X a', \quad \forall a' \in \mathbb{A} \setminus \{a\}. \tag{1.3}$$

Теорема 1.2. *Методы минимизации эмпирического риска (МЭР), максимизации переобученности (МП) и пессимистичной минимизации эмпирического риска (ПМЭР) являются финитными.*

Доказательство. Утверждение для МЭР и МП следует из леммы 1.1. Отношение порядка, определенное в лемме 1.2, соответствует способу выбора классификатора по обучающей выборке на основе метода обучения ПМЭР, откуда следует утверждение теоремы для ПМЭР.

Таким образом, рассматриваемые в данной работе методы МП и ПМЭР являются финитными. Далее будет показано, что для финитных методов обучения при вычислении функционалов обобщающей способности достаточно рассмотреть некоторое подмножество объектов генеральной совокупности. По аналогии с финитными функциями, это подмножество можно назвать *носителем* для метода обучения.

Заметим также, что из определения финитного отношения вытекает следующее свойство:

Лемма 1.3. *Пусть классификаторы семейства* $\mathbb{A}' \subseteq \mathbb{A}$ *неразличимы на множестве* $\mathbb{N}'$. *Тогда для любого* $a \in \mathbb{A}'$ *выполнение финитного отношения* $a \succ_X a'$ *одновременно для всех* $a' \in \mathbb{A}' \setminus \{a\}$ *не зависит от выбора разбиения множества* $\mathbb{N}'$.

Данное свойство финитного отношения позволит далее рассмотреть прямую последовательность как объединение двух последовательностей – левой и правой – и решить поставленную задачу независимо на каждом семействе.

1.5. Переобучение произвольного семейства

Пусть дано произвольное семейство классификаторов $\mathbb{A} = \{a_0, \ldots, a_P\}$. Определим множество $\mathbb{D}$ объектов, по которым классификаторы различимы:

$$\mathbb{D} = \big\{x \in \mathbb{X} \mid \exists\, a, a' \in \mathbb{A}\colon I(a,x) \neq I(a',x)\big\}. \tag{1.4}$$

Легко проверить, что множество $\mathbb{D}$ представимо в виде $\mathbb{D} = G_0 \cup \cdots \cup G_{P-1}$, где множества G_p определяются согласно (1.1).

Объекты множества $\mathbb{N} = \mathbb{X} \setminus \mathbb{D}$ назовем *нейтральными*. На множестве $\mathbb{N}$ классификаторы семейства неразличимы и допускают одинаковое число ошибок m. Через m_p обозначим число ошибок классификатора a_p на множестве $\mathbb{D}$:

$$\begin{aligned} m &= n(a, \mathbb{N}), \quad \forall a \in \mathbb{A}; \\ m_p &= n(a_p, \mathbb{D}). \end{aligned} \tag{1.5}$$

Будем обозначать через t число объектов из $\mathbb{D}$, попавших в обучающую выборку X, а через e — число ошибок классификатора a_p на этих объектах. Введём две функции от t и e: число разбиений множества $\mathbb{N}$, таких, что классификатор a_p переобучен на X

$$N_p(t,e) = \#\big\{(X \cap \mathbb{N}, \bar{X} \cap \mathbb{N}) \;\big|\; \delta(a_p, X) \geqslant \varepsilon,\ t = |X \cap \mathbb{D}|,\ e = n(a_p, X \cap \mathbb{D})\big\},$$

и число разбиений множества $\mathbb{D}$, таких, что a_p является результатом обучения:

$$D_p(t,e) = \#\big\{(X \cap \mathbb{D}, \bar{X} \cap \mathbb{D}) \;\big|\; \mu X = a_p,\ t = |X \cap \mathbb{D}|,\ e = n(a_p, X \cap \mathbb{D})\big\}.$$

Введём *гипергеометрическую функцию распределения*

$$H_L^{\ell,m}(s) = \frac{1}{C_L^\ell} \sum_{i=0}^{\min\{\lfloor s \rfloor, \ell, m\}} C_m^i C_{L-m}^{\ell-i},$$

где $\lfloor x \rfloor$ — целая часть x, то есть наибольшее целое число, не превосходящее x. Гипергеометрическая функция распределения $H_L^{\ell,m}(s)$ для данного множества $\mathbb{X}$ мощности L и выборки $X_0 \subset \mathbb{X}$ объема m равна доле выборок множества $\mathbb{X}$ объема ℓ, содержащих не более s элементов из X_0. Будем полагать $C_n^i = 0$ при невыполнении условия $0 \leqslant i \leqslant n$.

Теорема 1.3. *Для произвольного семейства классификаторов $\mathbb{A} = \{a_0, \ldots, a_P\}$, финитного метода обучения μ, множества $\mathbb{X}$ мощности L, объема обучающей выборки ℓ, точности $\varepsilon \in (0, 1)$ вероятность переобучения имеет вид*

$$Q_\varepsilon = \frac{1}{C_L^\ell} \sum_{p=0}^{P} \sum_{(t,e)\in\Psi_p} D_p(t,e) N_p(t,e), \tag{1.6}$$

где множество $\mathbb{D}$, параметры m_p и m определяются по (1.4) и (1.5)

$$\Psi_p = \big\{(t,e) \mid 0 \leqslant t \leqslant \min\{\ell, |\mathbb{D}|\},\ 0 \leqslant e \leqslant \min\{t, m_p\}\big\}; \tag{1.7}$$

$$N_p(t,e) = C_{L-|\mathbb{D}|}^{\ell-t}\, H_{L-|\mathbb{D}|}^{\ell-t,m}(s_p(e)); \tag{1.8}$$

$$s_p(e) = \frac{\ell}{L}(n(a_p, \mathbb{X}) - \varepsilon(L-\ell)) - e.$$

Доказательство. Представим вероятность переобучения в виде

$$Q_\varepsilon = \sum_{p=0}^{P} \mathsf{P}\big[\mu X = a_p \text{ и } \delta(a_p, X) \geqslant \varepsilon\big].$$

Рассмотрим множество разбиений $(X, \bar{X})$ с фиксированными значениями t и e:

$$t = |X \cap \mathbb{D}|, \quad e = n(a_p, X \cap \mathbb{D}). \tag{1.9}$$

Множество допустимых значений (t, e) есть Ψ_p, согласно (1.7).

Для таких разбиений выполнение условия $\delta(a_p, X) \geqslant \varepsilon$ не зависит от выбора разбиения множества $\mathbb{D}$, а выполнение условия $\mu X = a_p$ по лемме 1.3 не зависит от выбора разбиения множества $\mathbb{N}$, поскольку классификаторы неразличимы на множестве $\mathbb{N}$. Поэтому для каждой тройки параметров p, t, e число

разбиений множества $\mathbb{X}$, таких, что одновременно выполнены условия $\mu X = a_p$ и $\delta(\mu X, X) \geqslant \varepsilon$, равно произведению $N_p(t,e)D_p(t,e)$.

Докажем (1.8). Пусть $n(a_p, X \cap \mathbb{N}) = s$, тогда $n(a_p, X) = e + s$. Условие $\delta(a_p, X) \geqslant \varepsilon$ эквивалентно условию $n(a_p, X) \leqslant \frac{\ell}{L}(n(a_p, \mathbb{X}) - \varepsilon(L-\ell))$, значит, $s \leq s_p(e)$. Число разбиений множества $\mathbb{N}$ при данных t и s равно $C_m^s C_{L-|\mathbb{D}|-m}^{\ell-t-s}$, откуда следует

$$N_p(t,e) = \sum_{s=0}^{s_p(e)} C_m^s C_{L-|\mathbb{D}|-m}^{\ell-t-s} = C_{L-|\mathbb{D}|}^{\ell-t} \frac{1}{C_{L-|\mathbb{D}|}^{\ell-t}} \sum_{s=0}^{s_p(e)} C_m^s C_{L-|\mathbb{D}|-m}^{\ell-t-s} = C_{L-|\mathbb{D}|}^{\ell-t}\, H_{L-|\mathbb{D}|}^{\ell-t,m}(s_p(e)).$$

Для функционала полного скользящего контроля имеет место аналогичная теорема.

Теорема 1.4. *Для произвольного семейства классификаторов $\mathbb{A} = \{a_0, \ldots, a_P\}$, финитного метода обучения μ, множества $\mathbb{X}$ мощности L, объема обучающей выборки ℓ, функционал полного скользящего контроля имеет вид*

$$\mathsf{CCV} = \frac{1}{(L-\ell)C_L^\ell} \sum_{p=0}^{P} \sum_{(t,e)\in\Psi_p} D_p(t,e)F_p(t,e), \tag{1.10}$$

где

$$F_p(t,e) = \sum_{s=0}^{\min\{\ell-t,m\}} C_m^s C_{L-|\mathbb{D}|-m}^{\ell-t-s}\big(n(a_p,\mathbb{X}) - s - e\big), \tag{1.11}$$

множества $\mathbb{D}$ и Ψ_p определяются по (1.4) и (1.7), параметры m_p и m определяются по (1.5).

Доказательство. Запишем формулу полного скользящего контроля и переставим в ней знаки суммирования:

$$\mathsf{CCV} = \frac{1}{C_L^\ell} \sum_{X\in[\mathbb{X}]^\ell} \sum_{p=0}^{P} [\mu X = a_p]\, \nu(a_p, \bar{X}) = \frac{1}{C_L^\ell} \sum_{p=0}^{P} \sum_{X\in[\mathbb{X}]^\ell} [\mu X = a_p]\, \nu(a_p, \bar{X}).$$

Выполнение условия $\mu X = a_p$ по лемме 1.3 не зависит от выбора разбиения множества $\mathbb{N}$.

Представим число ошибок классификатора a_p на контрольной выборке в виде

$$n(a_p, \bar{X}) = n(a_p, \mathbb{X}) - n(a_p, X) = n(a_p, \mathbb{X}) - n(a_p, X \cap \mathbb{D}) - n(a_p, X \cap \mathbb{N}).$$

Определим параметры t и e по формулам (1.9). Обозначим $s = n(a_p, X \cap \mathbb{N})$. Из ограничений $s + t \leqslant l$ и $s \leqslant m$ следует верхняя оценка параметра s в (1.11).

Легко проверить, что число разбиений множества $\mathbb{N}$ при данных t и s равно $C_m^s C_{L-|\mathbb{D}|-m}^{\ell-t-s}$, откуда следует утверждение теоремы.

Аналогичная теорема имеет место для функционала ожидаемой переобученности.

Теорема 1.5. *Для финитного метода обучения μ, произвольной прямой последовательности классификаторов $\mathbb{A} = \{a_0, \ldots, a_P\}$, множества $\mathbb{X}$ мощности L, объема обучающей выборки ℓ выражение для ожидаемой переобученности имеет вид*

$$\mathsf{EOF} = \frac{1}{C_L^\ell} \sum_{p=0}^{P} \sum_{(t,e)\in\Psi_p} D_p(t,e) K_p(t,e), \tag{1.12}$$

где множества $\mathbb{D}$ и Ψ_p определяются по (1.4) и (1.7), параметры m_p и m определяются по (1.5) и

$$K_p(t,e) = \sum_{s=0}^{\min\{\ell-t, m\}} C_m^s C_{L-|\mathbb{D}|-m}^{\ell-t-s} \Big(\tfrac{1}{L-\ell} \big(n(a_p, \mathbb{X}) - (s+e) \big) - \tfrac{1}{\ell}(s+e) \Big). \tag{1.13}$$

Доказательство. Запишем формулу переобученности и переставим в ней знаки суммирования:

$$\mathsf{EOF} = \frac{1}{C_L^\ell} \sum_{X\in[\mathbb{X}]^\ell} \sum_{p=0}^{P} [\mu X = a_p]\, \delta(a_p, X) = \frac{1}{C_L^\ell} \sum_{p=0}^{P} \sum_{X\in[\mathbb{X}]^\ell} [\mu X = a_p]\, \delta(a_p, X).$$

Рассмотрим множество разбиений $(X, \bar{X})$ с фиксированными значениями t и e:

$$t = |X \cap \mathbb{D}|, \quad e = n(a_p, X \cap \mathbb{D}).$$

Множество допустимых значений (t, e) есть Ψ_p согласно (1.7).

Обозначим $s = n(a_p, X \cap \mathbb{N})$. Из ограничений $s + t \leqslant l$ и $s \leqslant m$ следует верхняя оценка параметра s в (1.13).

Поскольку число ошибок классификатора a_p на контрольной выборке равно

$$n(a_p, \bar{X}) = n(a_p, \mathbb{X}) - n(a_p, X) = n(a_p, \mathbb{X}) - n(a_p, X \cap \mathbb{D}) - n(a_p, X \cap \mathbb{N}),$$

переобученность классификатора a_p для данных s и e представляется в виде

$$\delta(a_p, X) = \tfrac{1}{L-\ell} n(a_p, \bar{X}) - \tfrac{1}{\ell} n(a_p, X) = \tfrac{1}{L-\ell}\big(n(a_p, \mathbb{X}) - (s+e)\big) - \tfrac{1}{\ell}(s+e).$$

Аналогично доказательству теоремы (1.4), из того, что выполнение условия $\mu X = a_p$ не зависит от выбора разбиения множества $\mathbb{N}$ и количество разбиений множества $\mathbb{N}$ для данных t и s равно $C_m^s C_{L-P-m}^{\ell-t-s}$, следует утверждение теоремы.

Таким образом, задача сводится к вычислению для каждого p значений $D_p(t, e)$, то есть носителем финитного метода обучения для произвольного семейства является множество $\mathbb{D}$ объектов, по которым классификаторы различимы. Для случая прямой последовательности далее будет описан рекуррентный алгоритм вычисления $D_p(t, e)$ для всех $(t, e) \in \Psi_p$.

1.6. Переобучение прямой последовательности

Пусть теперь семейство $\mathbb{A} = \{a_0, \ldots, a_P\}$ является прямой последовательностью. Объекты множества $\mathbb{D}$ будем называть *ребрами прямой последовательности* $\mathbb{A}$.

1.6.1. Сведение к задачам на левой и правой последовательностях

Рассмотрим классификатор a_p и зафиксируем точку $(t, e) \in \Psi_p$. Относительно a_p прямая последовательность $\mathbb{A}$ разбивается на две: левую $a_0, a_1, \ldots, a_p$ и правую $a_p, a_{p+1}, \ldots, a_P$.

Сведем задачу вычисления $D_p(t,e)$ к нахождению числа разбиений множества ребер левой и правой последовательностей с некоторыми ограничениями.

Теорема 1.6. *Пусть μ – финитный метод обучения. Для каждого p для всех $(t,e) \in \Psi_p$ число разбиений множества $\mathbb{D}$, таких, что $t = |X \cap \mathbb{D}|$, $e = n(a_p, X \cap \mathbb{D})$ и $\mu X = a_p$, равно*

$$D_p(t,e) = \sum_{t'+t''=t} \sum_{e'+e''=e} L_p(t',e') R_p(t'',e''), \tag{1.14}$$

где

$$L_p(t',e') = \#\left\{ (X \cap \mathbb{L}_p, \bar{X} \cap \mathbb{L}_p) \;\middle|\; \begin{array}{l} \forall d = 0,\ldots,p, \;\; a_p \succ_X a_d, \\ t' = |X \cap \mathbb{L}_p|, \;\; e' = n(a_p,\, X \cap \mathbb{L}_p) \end{array} \right\}, \tag{1.15}$$

$$R_p(t'',e'') = \#\left\{ (X \cap \mathbb{R}_p, \bar{X} \cap \mathbb{R}_p) \;\middle|\; \begin{array}{l} \forall d = p+1,\ldots,P, \;\; a_p \succ_X a_d, \\ t'' = |X \cap \mathbb{R}_p|, \;\; e'' = n(a_p,\, X \cap \mathbb{R}_p) \end{array} \right\}, \tag{1.16}$$

множества $\mathbb{L}_p$ и $\mathbb{R}_p$ – множества ребер левой и правой последовательностей соответственно, точки (t',e') и (t'',e'') являются элементами множеств Ψ'_p и Ψ''_p соответственно, где

$$\Psi'_p = \left\{ (t',e') \;\middle|\; 0 \leqslant t' \leqslant \min\{\ell, |\mathbb{L}_p|\}, 0 \leqslant e' \leqslant \min\{t',\, n(a_p,\mathbb{L}_p)\} \right\}, \tag{1.17}$$

$$\Psi''_p = \left\{ (t'',e'') \;\middle|\; 0 \leqslant t'' \leqslant \min\{\ell, |\mathbb{R}_p|\}, 0 \leqslant e'' \leqslant \min\{t'',\, n(a_p,\mathbb{R}_p)\} \right\}. \tag{1.18}$$

Доказательство. Множества $\mathbb{L}_p$ и $\mathbb{R}_p$ не пересекаются, значит, классификаторы левой последовательности неразличимы на $\mathbb{R}_p$, классификаторы правой последовательности неразличимы на $\mathbb{L}_p$. Тогда выполнение условия (1.3) для всех классификаторов левой последовательности по лемме 1.3 не зависит от выбора разбиения множества $\mathbb{R}_p$. Аналогично, выполнение условия (1.3) для всех классификаторов правой последовательности не зависит от выбора разбиения множества $\mathbb{L}_p$. Значит, общее число разбиений множества $\mathbb{D}$, в которых метод обучения выбирает a_p, является произведением числа разбиений множеств $\mathbb{L}_p$ и $\mathbb{R}_p$, в которых a_p лучше всех классификаторов левой и правой

последовательностей соответственно. Параметры t', t'', e', e'' необходимы для выполнения условий, задаваемых параметрами t и e.

Назовем разбиения множеств $\mathbb{L}_p$ и $\mathbb{R}_p$, удовлетворяющие условиям (1.15) и (1.16) соответственно, *допустимыми*.

Поскольку методы ПМЭР и МП, согласно теореме 1.2, являются финитными, то для них справедливы теоремы 1.3 — 1.6 и для каждого p задача сводится к вычислению числа допустимых разбиений $L_p(t', e')$ и $R_p(t'', e'')$ для всех точек множеств Ψ'_p и Ψ''_p.

Далее рассматривается случай, когда прямая последовательность $\mathbb{A}$ является прямой цепью. Тогда левая и правая последовательности также являются цепями. Рассматривается метод ПМЭР μ с определенным по лемме 1.2 отношением порядка $\succ_X$. Для метода МП рассуждения аналогичны.

1.6.2. Нахождение числа допустимых разбиений множества ребер левой цепи

Найдём $L_p(t', e')$ для каждого p в каждой точке $(t', e') \in \Psi'_p$. Заметим, что при $p = 0$ решение задачи тривиально: множество Ψ'_0 состоит из одной точки $(0, 0)$ и $L_0(0, 0) = 1$. Всюду далее считаем $1 \leqslant p \leqslant P$.

Перенумеруем классификаторы так, чтобы последовательность начиналась в a_p и заканчивалась в a_0. Обозначим $\{b_0, \ldots, b_p\}$, где $b_d = a_{p-d}$ для каждого $d = 0, \ldots, p$. Запас ошибок относительно a_p, определенный по (1.2), запишем как $\Delta_0(b_d, X) = \Delta_p(a_{p-d}, X)$ для каждого d.

Левая цепь составлена из возрастающих и убывающих монотонных участков. Обозначим множество всех ребер возрастающих монотонных участков цепи через $\mathbb{C}_p$, убывающих монотонных участков цепи — через $\mathbb{I}_p$. Верно, что $\mathbb{C}_p \sqcup \mathbb{I}_p = \mathbb{L}_p$.

Цепь прямая, следовательно, b_0 не ошибается на всех объектах $\mathbb{C}_p$, то есть

$$\begin{aligned} \mathbb{C}_p &= \{x \in \mathbb{L}_p \colon I(b_0, x) = 0\}, \\ \mathbb{I}_p &= \{x \in \mathbb{L}_p \colon I(b_0, x) = 1\}. \end{aligned} \tag{1.19}$$

Тогда верно, что $e' = |X \cap \mathbb{I}_p|$, а $|X \cap \mathbb{C}_p| = t' - e'$.

Заметим, что, поскольку классификаторы левой цепи различимы только на объектах множества $\mathbb{L}_p$, то для любого классификатора b из левой цепи верно

$$\Delta_0(b, X) = \Delta_0(b, X \cap \mathbb{L}_p), \quad \forall X \subseteq \mathbb{X}.$$

Отсюда следует, что, зафиксировав разбиение множества $\mathbb{L}_p$, мы определим запас ошибок на всех соответствующих обучающих выборках X.

Введём трехмерную сетку $\Omega_p = \{0, \ldots, |\mathbb{L}_p|\} \times \{-|\mathbb{L}_p|, \ldots, |\mathbb{L}_p|\} \times \{0, \ldots, |\mathbb{L}_p|\}$.

Определение 1.5. *Определим на Ω_p множество $\mathbb{T}_p$ траекторий, выходящих из точки $(0, 0, 0)$ и образованных переходами трех видов:*

1) из точки (d, Δ, i) в точку $(d + 1, \Delta, i)$ – «вправо»;

2) из точки (d, Δ, i) в точку $(d + 1, \Delta + 1, i)$ – «вправо-вверх»;

3) из точки (d, Δ, i) в точку $(d + 1, \Delta - 1, i + 1)$ – «вправо-вниз»;

причем для каждого d переход из точки (d, Δ, i) удовлетворяет условию: пусть классификаторы b_d и b_{d+1} соединены ребром x, тогда

1) если $x \in \mathbb{C}_p$, то это переход вида «вправо» или «вправо-вверх»;

2) если $x \in \mathbb{I}_p$, то это переход вида «вправо» или «вправо-вниз».

Теорема 1.7. *Между разбиениями множества $\mathbb{L}_p$ и траекториями из множества $\mathbb{T}_p$ имеется взаимно однозначное соответствие. Траектория, соответствующая разбиению $(X \cap \mathbb{L}_p, \bar{X} \cap \mathbb{L}_p)$, проходит через точки (d, Δ, i), где для каждого $d = 0, \ldots, p$ координата $\Delta = \Delta_0(b_d, X)$, а координата i равна числу ребер из $X \cap \mathbb{I}_p$ между b_0 и b_d.*

Доказательство. Пусть классификаторы b_{d-1} и b_d соединены ребром x.

Если $x \in \bar{X}$, то $\Delta_0(b_d, X) = \Delta_0(b_{d-1}, X)$, так как запас ошибок зависит только от X.

Пусть x лежит в X. Если x лежит в возрастающей цепи, то b_{d-1} не ошибается на этом ребре, тогда как b_d ошибается. Тогда $\Delta_0(b_d, X) = \Delta_0(b_{d-1}, X) + 1$. Если же x лежит в $\mathbb{I}_p$, то b_{d-1} ошибается на этом объекте, а b_d — нет. Значит, $\Delta_0(b_d, X) = \Delta_0(b_{d-1}, X) - 1$.

Поставим в соответствие разбиению множества $\mathbb{L}_p$ траекторию по следующему правилу. Пусть траектория проходит через точку (d, Δ, i). При $d = 0$ полагаем, что это точка $(0, 0, 0)$. Из этой точки вдоль траектории выполняется переход вида «вправо», если $x \in \bar{X}$; «вправо-вверх», если $x \in X \cap \mathbb{C}_p$; «вправо-вниз», если $x \in X \cap \mathbb{I}_p$.

Тогда для каждого d координаты Δ и i имеют смысл, указанный в условии теоремы, и при описанных переходах изменяются не более, чем на 1. Значит, траектория действительно целиком лежит на сетке Ω_p и, следовательно, во множестве $\mathbb{T}_p$ и однозначно определена.

По тем же правилам каждой траектории из $\mathbb{T}_p$ можно однозначно поставить в соответствие разбиение множества $\mathbb{L}_p$. Значит, отображение из множества разбиений во множество траекторий $\mathbb{T}_p$ сюръективно и инъективно, то есть оно биективно.

Пример 2. *На рис. 1.3 на нижнем графике изображена цепь, где выделены ребра, попавшие в обучающую выборку. Такому разбиению ребер цепи соответствует траектория, проекция которой на плоскость (d, Δ) изображена на верхнем графике. В данном примере траектория проходит через точки, у которых координата Δ отрицательна. Значит, в цепи имеются классификаторы с отрицательным запасом ошибок. Следовательно, по лемме 1.2 и условию (1.3), при таком разбиении классификатор b_0 не будет выбран методом обучения. Исключив из рассмотрения траектории, не удовлетворяющие*

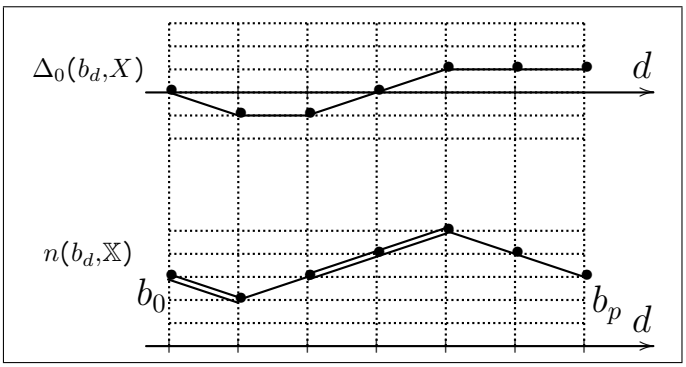


Рис. 1.3. Соответствие разбиения цепи (нижний график) проекции траектории (верхний график). Двойными линиями выделены ребра цепи, попавшие в обучающую выборку

лемме 1.2, мы отбросим и разбиения, не являющиеся допустимыми.

Определим множество

$$\Omega'_p = \left\{ (d,\Delta,i) \in \Omega_p \;\middle|\; \begin{array}{l} 0 \leqslant i \leqslant d \text{ и } |\Delta| \leqslant d \text{ и} \\ \big(\text{либо } \Delta > 0, \text{ либо } (\Delta = 0 \text{ и } n(b_d,\mathbb{X}) \leqslant n(b_0,\mathbb{X}))\big) \end{array} \right\}. \tag{1.20}$$

Лемма 1.4. *Всякая точка (d,Δ,i) траектории из $\mathbb{T}_p$, соответствующей допустимому разбиению множества $\mathbb{L}_p$, принадлежит множеству $\Omega'_p \subseteq \Omega_p$.*

Доказательство. Выполнение первых двух условий из определения (1.20) является следствием теоремы 1.7. Третье условие есть повторение условий леммы 1.2.

Пусть $T_p(d,\,\Delta,\,i)$ есть число траекторий из $\mathbb{T}_p$, соединяющих точку $(0,\,0,\,0)$ с $(d,\,\Delta,\,i)$ и проходящих только через точки множества Ω'_p. Из правил построения траектории по разбиению множества $\mathbb{L}_p$ следует

Лемма 1.5. *В каждой точке $(d,\,\Delta,\,i)$ на трехмерной сетке Ω_p величина $T_p(d,\,\Delta,\,i)$ вычисляется рекуррентно.*

1) Начальное условие $T_p(0,\,0,\,0) = 1$.

2) Если $(d,\,\Delta,\,i) \notin \Omega'_p$, то $T_p(d,\,\Delta,\,i) = 0$.

3) Пусть b_{d-1} и b_d соединены ребром x. Тогда

$$T_p(d,\,\Delta,\,i) = \begin{cases} T_p(d-1,\,\Delta,\,i) + T_p(d-1,\,\Delta-1,\,i), & \text{если } x \in \mathbb{C}_p, \\ T_p(d-1,\,\Delta,\,i) + T_p(d-1,\,\Delta+1,\,i-1), & \text{если } x \in \mathbb{I}_p, \end{cases} \tag{1.21}$$

где множества $\mathbb{C}_p$ и $\mathbb{I}_p$ определяются по (1.19).

Теорема 1.8. *Пусть даны метод ПМЭР μ, множество $\mathbb{X}$ мощности L, объем обучающей выборки ℓ и прямая цепь $\mathbb{A} = \{a_0, \ldots, a_P\}$. Тогда для каждого $p = 1, \ldots, P$ в каждой точке (t', e') множества Ψ'_p, определенного в (1.17), число $L_p(t', e')$ допустимых разбиений множества $\mathbb{L}_p$, определяемое по (1.15), равно*

$$L_p(t',\,e') = T_p(|\mathbb{L}_p|,\,t' - 2e',\,e')$$

и вычисляется рекуррентно по правилам, описанным в лемме 1.5, где $b_d = a_{p-d}$ для каждого d, при краевых условиях $L_0(0,0) = 1$.

Доказательство. Из теоремы 1.7 следует, что

$$\Delta_p(a_0, X) = |X \cap \mathbb{C}_p| - |X \cap \mathbb{I}_p| = t' - 2e'.$$

Между разбиениями множества ребер левой цепи и траекториями из $\mathbb{T}_p$ имеется биекция. Таким образом, число траекторий, проходящих через точку $(p,\,t' - 2e',\,e')$, равно числу разбиений, удовлетворяющих условиям $t' = |X \cap \mathbb{L}_p|$ и $e' = n(a_p, X \cap \mathbb{L}_p)$. Оставив среди них те, которые проходят только через точки множества $\Omega'_p(t', e')$, мы оставим траектории, соответствующие допустимым разбиениям. Их число равно $T_p(|\mathbb{L}_p|,\,t' - 2e',\,e')$.

Замечание 1.1. *Ограничения $i \leqslant e'$ и $\Delta \leqslant t' - e'$, являющиеся следствием теоремы 1.7, выполняются автоматически для тех траекторий, которые соединяют точки $(0,0,0)$ и $(p, t' - 2e', e')$. Действительно, поскольку величины*

i и $\Delta + i$ не возрастают, значит, не превосходят значений в конечной точке, то есть $i \leqslant e'$ и

$$\Delta + i \leqslant t' - 2e' + e' = t' - e'.$$

Координата $i \geqslant 0$, значит, $\Delta \leqslant \Delta + i \leqslant t' - e'$. В силу этого замечания, в определение множества Ω'_p данные ограничения не входят.

Таким образом, мы научились решать задачу для левой цепи.

1.6.3. Нахождение числа допустимых разбиений множества ребер правой цепи

Решаем задачу вычисления $R_p(t'', e'')$ для каждого p в каждой точке $(t'', e'') \in \Psi''_p$. Решение практически повторяет решение задачи для левой цепи после замены $\mathbb{L}_p$ на $\mathbb{R}_p$ и точки (t', e') на (t'', e''). Также имеются краевые условия: при $p = P$ множество $\Psi''_P = \{(0, 0)\}$ и $R_P(0, 0) = 1$. Далее полагаем, что $0 \leqslant p \leqslant P - 1$.

Обозначим классификаторы цепи через $b_d = a_{p+d}$ для каждого $d = 0, \ldots, P - p$. Из леммы 1.2 следует, что для справедливости леммы 1.5 для правой цепи множество Ω'_p необходимо заменить на множество Ω''_p, определяемое следующим образом:

$$\Omega''_p = \left\{ (d, \Delta, i) \in \Omega_p \;\middle|\; \begin{array}{l} 0 \leqslant i \leqslant d \text{ и } |\Delta| \leqslant d \text{ и} \\ \big(\text{либо } \Delta > 0, \text{ либо } (\Delta = 0 \text{ и } n(b_d, \mathbb{X}) < n(b_0, \mathbb{X}))\big) \end{array} \right\}. \tag{1.22}$$

По аналогии с теоремой 1.8, для правой цепи верна следующая теорема.

Теорема 1.9. *Пусть даны метод ПМЭР μ, множество $\mathbb{X}$ мощности L, объем обучающей выборки ℓ и произвольная прямая цепь $\mathbb{A} = \{a_0, \ldots, a_P\}$. Тогда для каждого $p = 0, \ldots, P - 1$ в каждой точке (t'', e'') множества Ψ''_p, определенного в (1.18), число $R_p(t'', e'')$ допустимых разбиений множества $\mathbb{R}_p$, опре-*

деляемое по (1.16), равно

$$R_p(t'', e'') = T_p(|\mathbb{R}_p|, t'' - 2e'', e'')$$

и вычисляется рекуррентно по правилам, описанным в лемме 1.5, с заменой множества Ω'_p *на* Ω''_p *и* b_d *на* a_{p+d} *для каждого* d. *Краевые условия* $R_P(0, 0) = 1$.

1.6.4. Нахождение числа допустимых разбиений множества ребер прямой последовательности

Рассмотрим общий случай прямой последовательности $\mathbb{A} = \{a_0, \ldots, a_P\}$. Сведем задачу вычисления количества допустимых разбиений левой и правой последовательностей к аналогичным задачам для прямых цепей.

Для этого построим прямую цепь $\mathbb{A}_c$, такую, что $\mathbb{A} \subseteq \mathbb{A}_c$ и первый и последний классификаторы семейств совпадают, следующим образом: для каждого i, такого, что $|G_i| > 1$, добавим в последовательность $\mathbb{A}$ прямую цепь $\mathbb{G}_i$

$$\{a_0, \ldots, a_{i-1}\} \cup \mathbb{G}_i \cup \{a_{i+2}, \ldots, a_P\},$$

где прямая цепь $\mathbb{G}_i$ такова, что первым классификатором цепи является a_i, последним — a_{i+1}. Для определенности будем считать, что $\mathbb{G}_i$ строится как прямая цепь, составленная из двух монотонных: убывающей цепи длины n_1 и возрастающей длины n_0, где

$$n_1 = \#\{x \in G_i \mid I(a_i, x) = 1\},$$
$$n_0 = \#\{x \in G_i \mid I(a_i, x) = 0\}.$$

Назовем построенную цепь $\mathbb{A}_c$ *интерполяцией* последовательности $\mathbb{A}$. Ее длина равна $|\mathbb{D}|$.

Для каждого $a_p \in \mathbb{A}$ рассмотрим левую последовательность $\{a_p, \ldots, a_0\} \subseteq \mathbb{A}$ и левую цепь $\{a_p, \ldots, a_0\} \subseteq \mathbb{A}_c$. По построению множества ребер данных семейств совпадают, вследствие чего множества допустимых разбиений левой

цепи и левой последовательности, определенные по (1.15), также совпадают. Вычислим их количество по теоремам 1.8 и 1.9 с единственным отличием.

Согласно (1.3), условие $a_p \succ_X a$ должно быть выполнено только для $a \in \mathbb{A}$. Данное ограничение определяет строение множеств Ω'_p и Ω''_p, задаваемых в (1.20) и (1.22). Переопределим их для случая интерполяции последовательности $\mathbb{A}$.

Заметим, что в приведенных выше рассуждениях для ПМЭР свойство отношения порядка, определенного в лемме 1.2, использовалось только при описании множеств Ω'_p и Ω''_p. Тогда, определив эти множества на основании отношения порядка из леммы 1.1, мы опишем способ вычисления искомого количества разбиений для метода МП.

Лемма 1.6. *Для метода ПМЭР множества Ω'_p и Ω''_p определяются по формулам:*

$$\Omega'_p = \left\{ (d,\Delta,i) \in \Omega_p \;\middle|\; \begin{array}{r} b_d \in \mathbb{A}_c\backslash\mathbb{A} \text{ или } \big(b_d \in \mathbb{A} \text{ и } 0 \leqslant i \leqslant d \text{ и } |\Delta| \leqslant d \\ \text{и } \big(\Delta > 0 \text{ или } (\Delta = 0 \text{ и } n(b_d,\mathbb{X}) \leqslant n(b_0,\mathbb{X}))\big)\big) \end{array} \right\}; \tag{1.23}$$

$$\Omega''_p = \left\{ (d,\Delta,i) \in \Omega_p \;\middle|\; \begin{array}{r} b_d \in \mathbb{A}_c\backslash\mathbb{A} \text{ или } \big(b_d \in \mathbb{A} \text{ и } 0 \leqslant i \leqslant d \text{ и } |\Delta| \leqslant d \\ \text{и } \big(\Delta > 0 \text{ или } (\Delta = 0 \text{ и } n(b_d,\mathbb{X}) < n(b_0,\mathbb{X}))\big)\big) \end{array} \right\}. \tag{1.24}$$

Для метода МП множества Ω'_p и Ω''_p определяются по формулам:

$$\Omega'_p = \left\{ (d,\Delta,i) \in \Omega_p \;\middle|\; \begin{array}{r} b_d \in \mathbb{A}_c\backslash\mathbb{A} \text{ или } \big(b_d \in \mathbb{A} \text{ и } 0 \leqslant i \leqslant d \text{ и } |\Delta| \leqslant d \\ \text{и } \big(\Delta \geq \frac{\ell}{L}(n(b_d,\mathbb{X}) - n(b_0,\mathbb{X}))\big)\big) \end{array} \right\}; \tag{1.25}$$

$$\Omega''_p = \left\{ (d,\Delta,i) \in \Omega_p \;\middle|\; \begin{array}{r} b_d \in \mathbb{A}_c\backslash\mathbb{A} \text{ или } \Big(b_d \in \mathbb{A} \text{ и } 0 \leqslant i \leqslant d \text{ и } |\Delta| \leqslant d \\ \text{и } \big(\Delta > \frac{\ell}{L}\,(n(b_d,\mathbb{X}) - n(b_0,\mathbb{X}))\big)\Big) \end{array} \right\}. \tag{1.26}$$

Доказательство. Для ПМЭР формулы следуют из доказанных ранее лемм. Выведем формулу для МП. Рассмотрим левую последовательность и $b_d \in \mathbb{A}$.

Перепишем переобученность как $\delta(b_d, X) = \frac{1}{L-\ell} n(b_d, \mathbb{X}) - \frac{L}{\ell(L-\ell)} n(b_d, X)$. Тогда

$$\delta(b_d, X) - \delta(b_0, X) = \frac{1}{L-\ell}\left(n(b_d, \mathbb{X}) - n(b_0, \mathbb{X})\right) - \frac{L}{\ell(L-\ell)}\left(n(b_d, X) - n(b_0, X)\right) =$$
$$= \frac{1}{L-\ell}\left(n(b_d, \mathbb{X}) - n(b_0, \mathbb{X})\right) - \frac{L}{\ell(L-\ell)}\Delta_0(b_d, X),$$

где величина $\Delta_0(b_d, X) = \Delta$, откуда на основе леммы 1.1 следует, что в левой последовательности для того, чтобы алгоритм b_0 был выбран методом обучения, необходимо выполнение условия $\delta(b_d, X) - \delta(b_0, X) \geq 0$, что равносильно

$$\Delta \geq \frac{\ell}{L}\left(n(b_d, \mathbb{X}) - n(b_0, \mathbb{X})\right).$$

Для правой последовательности неравенство строгое. Отсюда следует справедливость формул (1.25) и (1.26) и утверждение леммы.

Теорема 1.10. *Пусть даны метод μ ПМЭР или МП, множество $\mathbb{X}$ мощности L, объем обучающей выборки ℓ и прямая последовательность $\mathbb{A} = \{a_0, \ldots, a_P\}$. Пусть прямая цепь $\mathbb{A}_c = \{c_0, \ldots, c_{|\mathbb{D}|}\}$ является интерполяцией последовательности $\mathbb{A}$. Каждому классификатору $a_p \in \mathbb{A}$ соответствует $c_{i_p} \in \mathbb{A}_c$.*

Тогда для каждого $p = 1, \ldots, P$ в каждой точке (t', e') множества Ψ'_p, определенного в (1.17), число $L_p(t', e')$ допустимых разбиений множества $\mathbb{L}_p$, определяемое по (1.15), равно

$$L_p(t', e') = T_p(|\mathbb{L}_p|, t' - 2e', e') \tag{1.27}$$

и вычисляется рекуррентно по правилам, описанным в лемме 1.5, где $b_d = c_{i_p - d}$ для каждого d и множество Ω'_p определено по лемме 1.6. Краевые условия $L_0(0, 0) = 1$.

Для каждого $p = 0, \ldots, P-1$ в каждой точке (t'', e'') множества Ψ''_p, определенного в (1.18), число $R_p(t'', e'')$ допустимых разбиений множества $\mathbb{R}_p$, определяемое по (1.16), равно

$$R_p(t'', e'') = T_p(|\mathbb{R}_p|, t'' - 2e'', e'') \tag{1.28}$$

и вычисляется рекуррентно по правилам, описанным в лемме 1.5, с заменой множества Ω'_p *на* Ω''_p*, определенное по лемме 1.6, и* b_d *на* c_{i_p+d} *для каждого* d*. Краевые условия* $R_P(0,0) = 1$.

1.7. Алгоритм вычисления оценок обобщающей способности прямой последовательности

Итак, в теореме 1.10 описан алгоритм нахождения количества допустимых разбиений множеств ребер правой и левой последовательностей для каждого p. Остается подставить найденные значения в формулы (1.14), (1.6), (1.10) и (1.12). Для сокращения вычислений по теоремам 1.3, 1.4 и 1.5 для каждого p предлагается заранее вычислить $L_p(t', e')$, $R_p(t'', e'')$, $N_p(t, e)$, $F_p(t, e)$ и $B_p(t, e)$, после чего сложить полученные значения. Схема вычислений показана в алгоритме 1.

1.7.1. Сложность алгоритма

Оценим сложность выполнения шагов 5–12 алгоритма 1.

При вычислении $L_p(t', e')$ по теореме 1.8 на шагах 5–6 один раз для всех $(d, \Delta, i) \in \Omega'_p$ вычисляются $T_p(d, \Delta, i)$, затем для каждого $(t', e') \in \Psi'_p$ величина $L_p(t', e')$ полагается равной $T_p(d, t' - 2e', e')$. Множество Ω'_p вложено в куб со стороной $O(|\mathbb{L}_p|)$, поскольку каждая координата ограничена по модулю количеством ребер в левой последовательности. Следовательно, сложность выполнения шагов 5–6 составляет $O(|\mathbb{L}_p|^3)$. Аналогично, сложность выполнения шагов 7–8 составляет $O(|\mathbb{R}_p|)^3)$.

Для нахождения $N_p(t, e)$ и $F_p(t, e)$ необходимо вычислить биномиальные коэффициенты C_m^i и C_{L-P-m}^i при всех возможных i за $O(L)$. Биномиальные коэффициенты для каждого p не пересчитываются. При известных значениях биномиальных коэффициентов искомые $N_p(t, e)$ и $F_p(t, e)$ вычисляются за $O(L)$. Множество Ψ_p вложено в квадрат со стороной L, значит, выполнение шагов 9–12

выполняется за $O(L^3)$. Следовательно, сложность выполнения шагов 5–12 составляет $O(|\mathbb{D}|^3 + L^3) = O(L^3)$ для каждого p.

Множества Ψ'_p и Ψ''_p вложены в квадрат со стороной P, значит, шаги 13–15 выполняются за $O(L^5)$, и сложность алгоритма 1 также составляет $O(L^5)$.

1.8. Выводы к первой главе

В данной главе введено понятие финитного метода обучения. Показано, что финитными являются методы минимизации эмпирического риска (МЭР), пессимистичной минимизации эмпирического риска (ПМЭР) и максимизации переобученности (МП).

Рассмотрен общий случай произвольного семейства классификаторов и финитного метода обучения. Доказана теорема о представлении функционалов вероятности переобучения, полного скользящего контроля и ожидаемой переобученности как произведения числа разбиений множества объектов, по которым классификаторы различимы, и нейтрального множества объектов, на которых классификаторы неразличимы.

Для частного случая прямой последовательности классификаторов, порождаемых элементарными пороговыми правилами при варьировании параметра порога, и методов ПМЭР и МП доказана теорема о вычислении значений функционалов обобщающей способности. Приведен псевдокод алгоритма вычисления достигаемых верхних оценок функционалов обобщающей способности. Показано, что сложность алгоритма является полиномиальной.

Алгоритм 1: Вычисление вероятности переобучения, полного скользящего контроля и ожидаемой переобученности

Вход: матрица ошибок прямой последовательности $\mathbb{A} = \{a_0, \ldots, a_P\}$, параметры ℓ, ε, метод обучения ПМЭР или МП

Выход: вероятность переобучения Q_ε, полный скользящий контроль CCV и ожидаемая переобученность EOF.

построить прямую цепь $\mathbb{A}_c$ — интерполяцию последовательности $\mathbb{A}$;
определить m по (1.5);
для всех $p = 0, \ldots, P$
    разделить цепь $\mathbb{A}_c$ на две — левую $\{a_p, \ldots, a_0\}$ и правую $\{a_p, \ldots, a_P\}$;
    для всех *точек* (t', e') *множества* Ψ'_p, *определенного по* (1.17)
        найти $L_p(t', e')$ по формулам (1.27), (1.21), (1.23) и (1.25);
    для всех *точек* (t'', e'') *множества* Ψ''_p, *определенного по* (1.18)
        найти $R_p(t'', e'')$ по формулам (1.28), (1.21), (1.24) и (1.26);
    для всех *точек* (t, e) *множества* Ψ_p, *определенного по* (1.7)
        вычислить $N_p(t, e)$ по формуле (1.8);
        вычислить $F_p(t, e)$ по формуле (1.11);
        вычислить $K_p(t, e)$ по формуле (1.13);

$$Q_\varepsilon := \frac{1}{C_L^\ell} \sum_{p=0}^{P} \sum_{(t',e')\in\Psi'_p} \sum_{(t'',e'')\in\Psi''_p} L_p(t', e') R_p(t'', e'') N_p(t' + t'', e' + e'');$$

$$\mathsf{CCV} := \frac{1}{(L-\ell)C_L^\ell} \sum_{p=0}^{P} \sum_{(t',e')\in\Psi'_p} \sum_{(t'',e'')\in\Psi''_p} L_p(t', e') R_p(t'', e'') F_p(t' + t'', e' + e'');$$

$$\mathsf{EOF} := \frac{1}{(L-\ell)C_L^\ell} \sum_{p=0}^{P} \sum_{(t',e')\in\Psi'_p} \sum_{(t'',e'')\in\Psi''_p} L_p(t', e') R_p(t'', e'') K_p(t' + t'', e' + e'');$$

Глава 2

Исследование завышенности существующих оценок обобщающей способности пороговых решающих правил

Недостатком алгоритма, описанного в первой главе, является его большая вычислительная сложность. В данной главе проводится сравнение достигаемых верхних оценок обобщающей способности семейства одномерных пороговых классификаторов, вычисленных с помощью алгоритма, с известными ранее быстро вычислимыми верхними оценками с целью оценить порядки их завышенности и выявить те оценки, которые можно было бы использовать в реальных задачах.

2.1. Обзор известных оценок вероятности переобучения

2.1.1. Оценка Вапника–Червоненкиса

Верхняя оценка вероятности переобучения, полученная Вапником и Червоненкисом, является функцией от мощности множества объектов и сложности семейства алгоритмов. Мерой сложности на заданном множестве объектов является коэффициент разнообразия, определяемый как число попарно различных бинарных векторов ошибок, индуцируемых классификаторами семейства. В комбинаторной теории коэффициент разнообразия равен мощности семейства.

Теорема 2.1 (см. [66]) *Для любого метода обучения, множества* $\mathbb{X}$*, семейства классификаторов* $\mathbb{A}$*, объема* ℓ *обучающей выборки и порога* $\varepsilon \in (0,1)$

$$Q_\varepsilon \leqslant |\mathbb{A}| \max_{n=1,\ldots,L} H_L^{\ell,n}\big(\tfrac{\ell}{L}(n - \varepsilon(L-\ell))\big). \tag{2.1}$$

2.1.2. Оценка расслоения–связности

В комбинаторном подходе учитываются геометрические свойства булевых векторов ошибок классификаторов – расслоение и связность.

Под *расслоением* семейства понимается распределение классификаторов по слоям ошибок. *Слоем* называется множество классификаторов, допускающих на множестве $\mathbb{X}$ равное число ошибок. Чем меньше ошибок допускает классификатор, тем ниже его слой.

Связность предполагает, что для каждого классификатора в семействе найдется множество похожих классификаторов, отличающихся от него только на одном объекте выборки.

Пусть дано семейство классификаторов $\mathbb{A} = \{a_0, \ldots, a_P\}$ с известными векторами ошибок на множестве $\mathbb{X}$. На множестве классификаторов, как векторов ошибок, существует отношение лексикографического порядка $\leqslant$. Говорят, что классификатор a *предшествует* b и записывают $a \prec b$, если $a \leqslant b$ и расстояние Хемминга между ними равно 1.

Для каждого $a \in \mathbb{A}$ через u, q и n будут обозначаться:

$$u = |\{b \in \mathbb{A} \mid a \prec b\}|, \tag{2.2}$$

$$q = |\{x \in \mathbb{X} \mid I(a,x) = 1,\ \exists b < a : I(b,x) = 0\}|, \tag{2.3}$$

$$n = n(a, \mathbb{X}).$$

Теорема 2.2 (Оценки расслоения–связности, см. [59]) *Для произвольного множества $\mathbb{X}$, семейства классификаторов $\mathbb{A}$, метода обучения ПМЭР, объема ℓ обучающей выборки и порога $\varepsilon \in (0,1)$ справедливы оценки*

$$\begin{aligned} Q_\varepsilon &\leqslant \sum_{a\in\mathbb{A}} \frac{C^{l-u}_{L-u-q}}{C^\ell_L} H^{\ell-u,n-q}_{L-u-q}\big(\tfrac{\ell}{L}(n - \varepsilon(L-\ell))\big), \\ \mathsf{CCV} &\leqslant \sum_{a\in\mathbb{A}} \frac{C^{\ell-u}_{L-u-q}}{C^\ell_L}\Big(\frac{n}{L-\ell} - \frac{(n-q)(\ell-u)}{(L-u-q)(L-\ell)}\Big). \end{aligned} \tag{2.4}$$

В [70] оценка была улучшена за счет более тонкого анализа эффекта связности.

Теорема 2.3 (Оценка Соколова, см. [70]) *Пусть S – множество истоков семейства $\mathbb{A}$, т. е. множество классификаторов s таких, что нет классификаторов $a \prec s$. Тогда верна оценка*

$$Q_\varepsilon \leqslant \sum_{p=0}^{P} \min_{s\in S}\left\{ \sum_{i=0}^{\min\{|A_{ps}|,\ |B_{ps}|\}} \frac{C_{|B_{ps}|}^{i} C_{L-u-|B_{ps}|}^{\ell-u-i}}{C_L^\ell} H_{L-u-|B_{ps}|}^{\ell-u-i,\ n-|B_{ps}|}\Big(\tfrac{\ell}{L}(n-\varepsilon(L-\ell))-i\Big)\right\}, \tag{2.5}$$

где

$$A_{ij} = \{x \in \mathbb{X} \mid I(a_i,x) = 0,\ I(a_j,x) = 1\},$$
$$B_{ij} = \{x \in \mathbb{X} \mid I(a_i,x) = 1,\ I(a_j,x) = 0\}.$$

2.2. Обзор известных оценок полного скользящего контроля

В комбинаторной теории наряду с оценкой расслоения–связности (2.4) для частного случая семейства имеются оценки Гуза.

2.2.1. Оценки Гуза

Рассмотрим семейство одномерных пороговых решающих правил над числовым признаком, принимающим попарно различные значения на объектах множества $\mathbb{X}$. Пусть порог пробегает все возможные значения. Для данного семейства в [72] был предложен полиномиальный алгоритм вычисления верхней и нижней оценок полного скользящего контроля.

Теорема 2.4 (см. [72]) *Для произвольного множества $\mathbb{X}$, семейства классификаторов $\mathbb{A}$, метода обучения ПМЭР, объема ℓ обучающей выборки справедливы верхняя и нижняя оценки полного скользящего контроля:*

$$\frac{1}{L-\ell}\frac{1}{C_L^\ell}\sum_{i=1}^{L}|E^1(i)| \leqslant \mathsf{CCV} \leqslant 1 - \frac{1}{L-\ell}\frac{1}{C_L^\ell}\sum_{i=1}^{L}|E^0(i)|, \tag{2.6}$$

где для каждого i множества $E^0(i)$ безошибочных выборок и $E^1(i)$ ошибочных выборок определены как

$$E^0(i) = \{X \mid x_i \in \bar{X},\ \forall \mu X \in M(X)\ I(\mu X, x_i) = 0\} \subseteq [\mathbb{X}]^\ell,$$
$$E^1(i) = \{X \mid x_i \in \bar{X},\ \forall \mu X \in M(X)\ I(\mu X, x_i) = 1\} \subseteq [\mathbb{X}]^\ell.$$

2.3. Оценка частоты ошибок на контрольной выборке

В данном разделе доказывается теорема для вычисления оценок частоты ошибок классификатора, выбранного методом ПМЭР, на контрольной выборке.

Пусть задана вероятностная мера P_σ. Для семейства $\mathbb{A}$ и множества $\mathbb{X}$ определим *Радемахеровскую сложность*

$$\mathcal{R}_L(\mathbb{A}, \mathbb{X}) = 2\mathsf{E}_\sigma \sup_{a \in \mathbb{A}} \frac{1}{L} \sum_{i=1}^{L} \sigma_i I(a, x_i),$$

где $I(a, x_i)$ обозначает ошибку классификатора, E_σ – математическое ожидание по мере P_σ, а радемахеровские случайные величины $\sigma_1, \ldots, \sigma_L$, независимые в совокупности, определяются как

$$\sigma_i = \begin{cases} +1, \mathsf{P}_\sigma = \frac{1}{2}, \\ -1, \mathsf{P}_\sigma = \frac{1}{2}. \end{cases}$$

Радемахеровская сложность описывает сложность семейства классификаторов. Чем больше Радемахеровская сложность, тем лучше ошибки классификаторов семейства могут коррелировать со случайным шумом σ_i.

В [58] было продемонстрировано, что функционал ожидаемой переобученности возникает в выражении Радемахеровской сложности, таким образом связывая комбинаторную теорию с теорией эмпирических процессов и неравенств концентрации вероятностной меры:

Лемма 2.1 (см. [58]) *Для метода обучения μ_δ МП, конечного семейства классификаторов $\mathbb{A}$, множества $\mathbb{X}$ мощности L, объема обучающей выборки $\ell = \frac{L}{2}$ верно*

$$\mathsf{EOF}(\mu_\delta, \mathbb{A}, \mathbb{X}, \ell) = \mathcal{R}_L(\mathbb{A}, \mathbb{X}).$$

Из определения ожидаемой переобученности и метода МП следует лемма 2.2.

Лемма 2.2. *Для методов обучения μ_δ МП и μ ПМЭР, конечного семейства классификаторов $\mathbb{A}$, множества $\mathbb{X}$ мощности L, объема обучающей выборки ℓ верно*

$$\mathsf{EOF}(\mu, \mathbb{A}, \mathbb{X}, \ell) \leq \mathsf{EOF}(\mu_\delta, \mathbb{A}, \mathbb{X}, \ell).$$

Из лемм 2.1 и 2.2 следует теорема 2.5.

Теорема 2.5. *Для метода обучения μ ПМЭР, конечного семейства классификаторов $\mathbb{A}$, множества $\mathbb{X}$ мощности L с вероятностью ε верно*

$$\nu(\mu X, \bar{X}) \leq \nu(\mu X, X) + \mathsf{EOF}(\mu, \mathbb{A}, \mathbb{X}, \ell) + \eta(\varepsilon), \tag{2.7}$$

$$\nu(\mu X, \bar{X}) \leq \nu(\mu X, X) + \mathcal{R}_L(\mathbb{A}, \mathbb{X}) + \eta(\varepsilon), \tag{2.8}$$

где поправка $\eta(\varepsilon) = \sqrt{-\frac{1}{2}\ln\varepsilon}$ и объем обучающей выборки $\ell = \frac{L}{2}$.

Доказательство. С помощью неравенства Хёфдинга [32] с вероятностью ε можно оценить отклонение $\eta = \eta(\varepsilon)$ переобученности от ее математического ожидания:

$$\delta(\mu X, \bar{X}) \leq \mathsf{EOF}(\mu) + \eta(\varepsilon),$$

где отклонение $\eta = \sqrt{-\frac{1}{2}\ln\varepsilon}$.

Тогда частоту ошибок классификатора, выбранного методом ПМЭР, на контрольной выборке, можно оценить непосредственно через частоту ошибок

на обучающей выборке и математическое ожидание переобученности:

$$\nu(\mu X, \bar{X}) \;\; = \nu(\mu X, \bar{X}) + \delta(\mu X, \bar{X}) \;\; \leq \nu(\mu X, X) + \mathsf{EOF}(\mu) + \eta(\varepsilon),$$

что обосновывает неравенство (2.7). Отсюда и из лемм 2.1 и 2.2 следует неравенство (2.8). Теорема 2.5 доказана.

Оценки, предложенные в теореме 2.5, различаются в следующем. Из леммы 2.2 следует, что оценка (2.7) является более точной, чем оценка (2.8). Однако недостатком оценки (2.7), вычислямой по теореме 1.3, является большая вычислительная сложность $O(L^5)$, тогда как для Рademахеровской сложности возможно построить быстро вычислимую верхнюю оценку, основанную на лемме Массара [51].

В данной работе мы сравним достигаемые верхние оценки ожидаемой переобученности методов обучения ПМЭР μ и МП μ_δ на примере семейства прямых последовательностей. Если зазор между значениями $\mathsf{EOF}(\mu_\delta)$ и $\mathsf{EOF}(\mu)$ окажется достаточно малым, то в правой части неравенства (2.8) можно заменить величину $\mathcal{R}_L(\mathbb{A}, \mathbb{X})$ на ее быстро вычислимую оценку и использовать полученную оценку в практических задачах.

2.4. Вычислительные эксперименты

Напомним обозначения. Дана прямая последовательность $\mathbb{A} = \{a_0, \ldots, a_P\}$, множество $\mathbb{X}$ мощности L, множество $\mathbb{D}$ ребер последовательности. Объем обучающей выборки ℓ. Точность ε. Параметр m равен числу ошибок на множестве $\mathbb{X} \setminus \mathbb{D}$.

2.4.1. Модельные данные

В экспериментах используются случайные прямые последовательности. Для порождения таких семейств генерируются класс нулей $X_0 \sim \mathcal{N}(0, 1)$ и класс единиц $X_1 \sim \mathcal{N}(\Delta, 1)$ как выборки равной мощности $\frac{L}{2}$ из нормальных

распределений. Параметр Δ влияет на минимальное количество ошибок классификаторов семейства. Объединение выборок является множеством $\mathbb{X}$. Множество $\mathbb{D}$ соответствует значениям, которые пробегает порог.

В экспериментах по сравнению оценок полного скользящего контроля в силу ограниченности оценок Гуза рассматриваются прямые цепи. Для порождения таких цепей из выборок X_0 и X_1 удаляются совпадающие элементы. Также оценки Гуза справедливы только для случая, когда порог пробегает все возможные значения, т. е. множество $\mathbb{D}$ совпадает с $\mathbb{X}$.

2.4.2. Сравнение с существующими оценками вероятности переобучения

На рис. 2.1 в логарифмической шкале отложены значения оценки Вапника–Червоненкиса (точки ■), оценки расслоения–связности (точки ◆) и оценки Соколова (точки ●) в сравнении с достигаемой верхней оценкой вероятности переобучения прямой последовательности (точки ▶). Горизонтальной линией указано значение $Q_\varepsilon = 1$.

Оценки расслоения–связности и Соколова являются точными только в одном случае, когда минимальное количество ошибок совпадает с параметром m. В этом случае семейство является унимодальной цепью (второй график в верхнем ряду на рис. 1.2), т. е. на границе классов отсутствует шум и граница определяется четко. С увеличением минимального количества ошибок оценка (2.5) начинает превосходить реальное значение вероятности переобучения. Оценка Вапника–Червоненкиса для рассматриваемой последовательности оказывается завышенной при любом значении минимального количества ошибок.

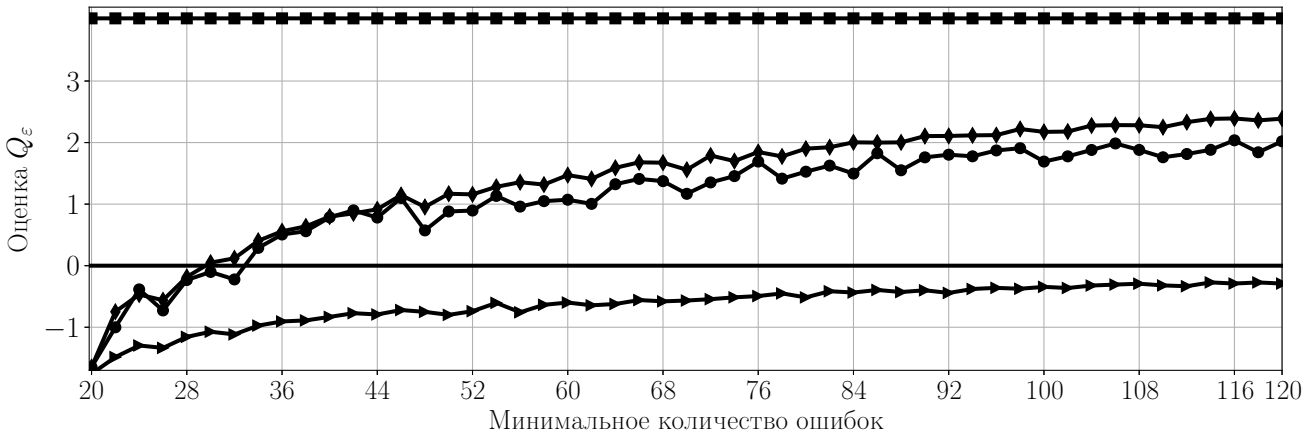


Рис. 2.1. Сравнение верхних оценок вероятности переобучения в логарифмической шкале. Условия эксперимента: $L = 240$, $\ell = 160$, $m = 20$, $\varepsilon = 0,05$. По горизонтали отложено минимальное количество ошибок классификаторов.

2.4.3. Сравнение с существующими оценками полного скользящего контроля

На рис. 2.2 по вертикали в логарифмической шкале отложены значения нижней (точки •) и верхней оценок Гуза (точки ▪) и оценки расслоения–связности (точки ◆) в сравнении с достигаемой верхней оценкой (точки ▶).

Оценка расслоения–связности оказывается точной только в случае, когда минимальное количество ошибок равно нулю, т. е. когда классы линейно разделимы, для остальных значений параметра она является завышенной. Верхняя оценка Гуза практически совпадает с достигаемой, из чего можно сделать вывод о высокой точности оценок.

2.4.4. Сравнение с Радемахеровской сложностью

На рис. 2.3 по вертикали в логарифмической шкале отложены значения ожидаемой переобученности классификатора, выбираемого методами обучения ПМЭР (точки •) и МП (точки ▶). Можно отметить, что с увеличением минимального числа ошибок классификаторов в семействе, т. е. с увеличением шума, два метода обучения начинают давать близкие значения ожидаемой переобученности. В этом случае для получения верхней оценки частоты ошибок

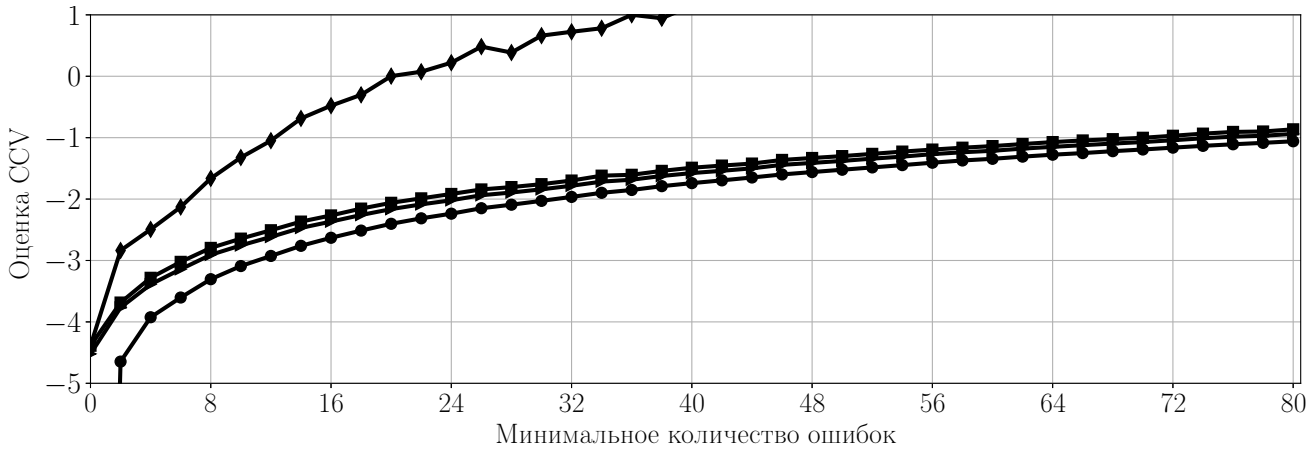


Рис. 2.2. Сравнение верхних оценок полного скользящего контроля в логарифмической шкале. Условия эксперимента: $L = 240$, $\ell = 160$, $m = 0$. По горизонтали отложено минимальное количество ошибок классификаторов.

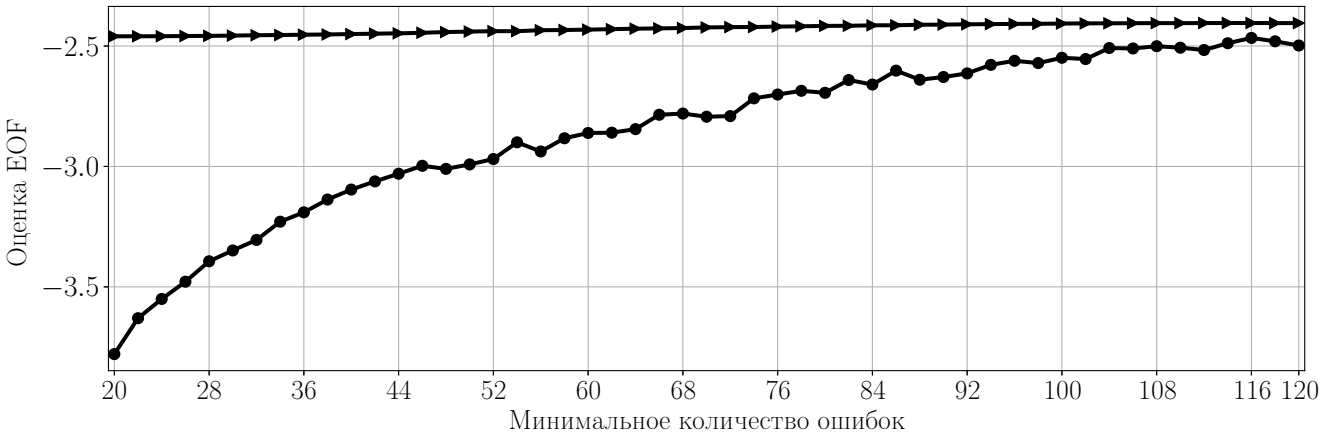


Рис. 2.3. Сравнение оценок ожидаемой переобученности для методов ПМЭР и МП в логарифмической шкале. Условия эксперимента: $L = 240$, $\ell = 120$, $m = 0$. По горизонтали отложено минимальное количество ошибок классификаторов.

классификатора, выбранного методом ПМЭР, на контрольной выборке, можно использовать оценку (2.8). Но при малом уровне шума ожидаемая переобученность метода ПМЭР оказывается ниже на два порядка. В этом случае для повышения точности оценок необходимо пользоваться оценкой (2.7).

2.5. Выводы ко второй главе

Проведено исследование обобщающей способности семейства одномерных пороговых решающих правил, определяемой с помощью функционалов вероят-

ности переобучения, полного скользящего контроля и ожидаемой переобученности.

Показано, что имеющиеся верхние оценки вероятности переобучения являются завышенными на 1–2 порядка. Оценки Вапника–Червоненкиса не учитывают геометрическую структуру классов, от которой, как показано в экспериментах, зависит обобщающая способность семейства. Комбинаторные оценки, несмотря на то, что принимают во внимание такие геометрические свойства, как расслоение и связность, все равно являются завышенными для данного семейства.

Оценки Гуза для полного скользящего контроля, полученные в рамках комбинаторной теории переобучения и вычислимые за полиномиальное от общего количества объектов время, демонстрируют высокую точность, что обосновывает возможность применения данных оценок в реальных задачах.

Оценки Радемахеровской сложности оказываются достаточно точными только для задач с высоким уровнем шума на границе классов. В противном случае, когда граница между классами определяется четко, оценки Радемахеровского типа являются завышенными на несколько порядков и неприменимыми на практике.

Поскольку имеющиеся верхние оценки обобщающей способности, за исключением оценок Гуза, не продемонстрировали высокую точность, возникает задача улучшения алгоритма вычисления достигаемых верхних оценок и уменьшения его вычислительной сложности. Также следующей задачей является применение достигаемых верхних оценок для повышения качества алгоритмов классификации, в частности для модификации критериев отбора признаков в жадных алгоритмах индукции конъюнктивных логических закономерностей и других логических алгоритмах классификации.

Глава 3

Применение комбинаторных оценок при планировании трассерных исследований в нефтегазовых месторождениях

Трассерное исследование (ТИ) скважин или исследования методом закачки меченой жидкости является одним из наиболее распространенных методов для оценки наличия гидродинамической связи между нагнетательными и добывающими скважинами [44, 79]. Промысловые данные по этим исследованиям позволяют определить гидродинамические свойства пласта в межскважинном пространстве, направление и скорость распространения жидкости в пласте и проверить наличие гидродинамической связи между скважинами, что важно для решения задач проектирования и мониторинга разработки месторождений [24, 49]. Выбор трассера, план исследования зависят от решаемых задач, особенностей продуктивного пласта и насыщающих его флюидов, требований к экологической безопасности [34, 60].

В последнее десятилетие инженеры-нефтяники при решении практических задач активно начинают применять подходы к интерпретации данных, основанные на интеллектуальном анализе данных. Например, анализ петрофизических данных с использованием передовых методов статистики и машинного обучения получил широкое распространение благодаря уменьшению неопределенностей и прогнозированию более точных трендов в данных по сравнению с классическими методами [36, 37, 53]. В [4] описан алгоритм автоматической интерпретации результатов гидродинамических исследований нефтяных и газовых добывающих скважин на неустановившихся режимах фильтрации с применением методов машинного обучения, который позволяет снизить неопределенность при выборе релевантных моделей системы «скважина-пласт» и повысить достовер-

ность определяемых параметров при анализе промысловых данных с высокой зашумленностью и наличии помех и выбросов. В [12] показано, как сверточные нейронные сети применяются для решения задачи интерпретации исследований методом эхометрирования и определения уровня жидкости в затрубном пространстве скважины. Известны примеры решения задач по созданию алгоритма «виртуального расходомера» на основе методов машинного обучения для повышения дискретности замеров дебитов жидкости в скважине по динамическим данным забойного давления с телеметрических систем и параметров со станции управления установки электроцентробежных насосов [15, 85].

Данная глава посвящена применению методов машинного обучения для проектирования трассерных исследований с целью повышения достоверности результатов по выявлению гидродинамической связи в пласте между нагнетательными и добывающими скважинами. Предложен новый алгоритм построения дерева решений с критерием выбора атрибута для разделения узла на основе оценок обобщающей способности пороговых решающих правил.

3.1. Планирование трассерных исследований с применением методов машинного обучения

В процессе проведения трассерного исследования для оценки наличия гидродинамической связи между скважинами осуществляется закачка меченой жидкости в одну или в несколько нагнетательных скважин. Закачиваемая меченая жидкость оттесняется к реагирующим окружающим добывающим скважинам. После закачки трассера начинают производить отбор проб жидкости из устья добывающих скважин, которые далее анализируются в лаборатории. Масса вынесенного трассера, кривые изменения концентрации трассера в устьевых пробах от времени позволяют определить фильтрационную неоднородность продуктивного пласта, каналы фильтрации между скважинами, а также выполнить количественные оценки фильтрационных параметров пласта [79].

В том случае, если меченую жидкость закачали в нагнетательную скважину и зафиксировали в реагирующей добывающей скважине, делают вывод, что скважины имеют гидродинамическую связь. Если по результатам исследований скважины не имеют гидродинамическую связь, то этот факт учитывается при дальнейшем планировании мероприятий по заводнению изученного участка месторождения. Если обнаружена ярко выраженная гидродинамическая связь между скважинами с приходом меченой жидкости в короткое время и значительном объемом агента закачки, то это свидетельствует о риске преждевременного обводнения таких реагирующих добывающих скважин [77]. В последнем случае планируются мероприятия по ограничению нагнетательных скважин или планируется закачка потокоотклоняющих составов.

При планировании ТИ, согласно [79], для отбора скважин и участка проведения трассерного исследования должен быть выполнен ряд условий:

- ТИ ранее на скважинах не проводились или с момента ранее проведенного ТИ прошло более 1 года;
- для добывающей скважины: скважина не является кандидатом на проведение геолого-технических мероприятий и на ней отсутствует заколонная циркуляция (ЗКЦ), обводненность больше 20%;
- для нагнетательной скважины: скважина отсутствует в списке на ограничение закачки, работает в установившемся режиме, техническое состояние глубинного и устьевого оборудования не препятствует закачке, отсутствует ЗКЦ, объект разработки совпадает с объектом разработки добывающей скважины;
- для участка исследования: есть уверенная оценка проницаемости по результатам гидродинамических исследований или по результатам анализа добычи/давления.

Недостатком описанного способа построения программы ТИ является то,

что перечисленные критерии не учитывают фактические динамические промысловые данные по эксплуатации скважин, вследствие чего среди выбранных скважин могут оказаться те, между которыми нет гидродинамической связи, и меченая жидкость в таком случае в добывающей скважине обнаружена не будет. Таким образом, список скважин, в которые закачивается индикатор, оказывается избыточным и приводит к более значительным затратам на проведение исследования.

Согласно [81], программа ТИ может быть расширена парами добывающих и нагнетательных скважин, между которыми установлено наличие гидродинамической связи по итогам других гидродинамических исследований, например, исследований методом гидропрослушивания (ГП). Массовое проведение исследований методом гидропрослушивания невозможно из-за множества ограничений и недостатков:

- высокая стоимость исследования ввиду необходимости затрат на составление программы и контроль ГП, проведение измерений;
- потери в добыче/закачке при остановке возмущающих (нагнетательных) и реагирующих (добывающих) скважин;
- с увеличением количества анализируемых пар скважин увеличивается длительность и количество измерений и срок получения результатов анализа измерений;
- технические ограничения на проведение ГП, например, неисправность задвижек и фонтанной арматуры; ввиду высокой стоимости исследования, ГП требует заблаговременного планирования и согласования.

С целью устранения указанных недостатков и повышения информативности трассерных исследований в нефтегазовых месторождениях предлагается способ планирования ТИ с применением методов машинного обучения.

Согласно способу, проводится построение классификатора для определения гидродинамической связи между парой скважин. Для этого в качестве обучающих примеров для настройки модели используются пары скважин нагнетательная-добывающая, на которых ранее проводились трассерные исследования. Для каждой пары известно целевое значение – заключение исследования о приходе трассера: «да», если трассер пришел в добывающую скважину, и «нет», если трассер не пришел. Далее на основе ответов классификатора для каждой пары скважин нагнетательная–добывающая формируется программа исследования. Если модель машинного обучения выдает ответ «да», то пара скважин включается в программу исследования, в нагнетательную скважину производится закачка трассера и в добывающей скважине проводится отбор проб. Если алгоритм выдает ответ «нет», то пара скважин из программы исследования исключается.

В качестве признаков для описания объектов обучающей выборки используются коэффициенты взаимовлияния по методам емкостно-резистивной модели (CRMIP) и многопараметрической регрессии (MLR), рассчитываемые на основе динамических данных эксплуатации скважин [64].

В рамках метода CRMIP анализируются приемистость нагнетательной скважины, дебит и забойное давление добывающей скважины, в методе MLR дополнительно анализируются забойное давление нагнетательной скважины и другие данные нормальной эксплуатации скважин.

Обозначения динамических параметров, используемых при расчете коэффициентов по методам MLR и CRMIP, на j-й добывающей скважине и i-й нагнетательной скважине, даны в таблице 3.1.

При наличии взаимовлияния между скважинами изменение динамического параметра на возмущающей скважине приводит к изменению на реагирующей скважине.

Коэффициенты, рассчитанные по методу MLR, являются мерой корреляции между динамическими параметрами на добывающей и нагнетательной

Обозначение	Динамический параметр	Единицы измерения
q_j^k	дебит жидкости в k-й момент времени	м3/сут
p_j^k	забойное давление на добывающей скважине в k-й момент времени	МПа
w_i^k	приемистость в k-й момент времени	м3/сут
p_i^k	забойное давление на нагнетательной скважине в k-й момент времени	МПа

Таблица 3.1. Обозначения динамических параметров

скважинах. Метод основан на решении задачи многопараметрической линейной регрессии:

$$\hat{\psi}_j^k = \beta_{0j} + \sum_{i=1}^{I} \beta_{ij}\psi_i^k, \qquad k = [0, \ldots, T], \tag{3.1}$$

где β_{0j} – свободный член, β_{ij} – вероятностный коэффициент влияния i-й скважины на j-ю скважину, $\hat{\psi}_j^k$ – значение динамического параметра на наблюдательной j-й скважине в k-й момент времени, ψ_i^k – значение динамического параметра в возмущающей i-й скважине в k-й момент времени, I – число возмущающих скважин, T – количество точек по времени.

При рассмотрении пар скважин i-я нагнетательная и j-я добывающая параметр $I = 1$ и уравнение (3.1) переписывается в виде:

$$\hat{\psi}_j^k = \beta_{0j} + \beta_{ij}\psi_i^k, \qquad k = [0, \ldots, T]. \tag{3.2}$$

Далее система линейных уравнений (3.2) решается методом наименьших квадратов, и вычисляются искомые коэффициенты по формулам (3.3):

$$\beta_{ij} = \frac{(T+1)\sum_{k=0}^{T}\psi_i^k\hat{\psi}_j^k - (\sum_{k=0}^{T}\psi_i^k)(\sum_{k=0}^{T}\hat{\psi}_j^k)}{(T+1)\sum_{k=0}^{T}(\psi_i^k)^2 - (\sum_{k=0}^{T}\psi_i^k)^2}. \tag{3.3}$$

С помощью уравнения (3.1) анализируется одновременное изменение динамических параметров скважин. В зависимости от анализируемых динамических параметров коэффициенты β_{ij} обозначаются согласно таблице 3.2.

Коэффициент MLR β_{ij}	Параметр на добывающей скважине $\hat{\psi}_j^k$	Параметр на нагнетательной скважине ψ_i^k
$p(p)_{ij}$	p_j^k	p_i^k
$p(w)_{ij}$	p_j^k	w_i^k
$q(p)_{ij}$	q_j^k	p_i^k
$q(w)_{ij}$	q_j^k	w_i^k

Таблица 3.2. Обозначения коэффициентов в методе MLR

Метод CRMIP на основе уравнения математического баланса позволяет оценить взаимовлияние между скважинами и временной параметр, который отражает реакцию добывающей скважины на изменения в работе нагнетательной скважины:

$$\tau_{ij}\frac{dq_{ij}(t)}{dt} + q_{ij}(t) = f_{ij}(t)w_{ij}(t) - \tau_{ij}J_{ij}\frac{dp_j(t)}{dt}, \tag{3.4}$$

где $w_i(t)$ – приемистость i-й нагнетательной скважины, м3/сут; $p_j(t)$ – забойное давление j-й добывающей скважины, МПа; $q_{ij}(t)$ – дебит j-й добывающей скважины (нефть и вода), обусловленный влиянием i-й нагнетательной скважины, м3/сут; f_{ij} – параметр взаимосвязи, определяющий объемную долю закачанной в нагнетательную скважину жидкости, которая фильтруется к добывающей скважине ($0 \leq f_{ij} \leq 1$); τ_{ij} – временной параметр ($\tau_{ij} \geq 0$), сут; J_{ij} - параметр продуктивности ($J_{ij} \geq 0$), м3/(сут $\cdot$ МПа); t – время от начала исследования, сут.

Аналитическое решение уравнения (3.4) представляется в виде

$$q_{ij}^k = q_{ij}^{k-1}e^{-\frac{\Delta t}{\tau_{ij}}} + \left(1 - e^{-\frac{\Delta t}{\tau_{ij}}}\right)\left(f_{ij}w_i^k - J_{ij}\tau_{ij}\frac{p_j^k - p_j^{k-1}}{\Delta t}\right). \tag{3.5}$$

Для определения коэффициентов f_{ij}, τ_{ij}, J_{ij} отдельно для каждой i-ой нагнетательной скважины решается задача минимизации невязки методом Нелдера-Мида [47]:

$$\sum_{k=1}^{T}\sum_{j=1}^{N}\left(\frac{\tilde{q}_j^k - q_{ij}^k}{\tilde{q}_{ij}^k}\right)^2 \to \min_{f_{ij},\tau_{ij},J_{ij}}, \tag{3.6}$$

где N – число наблюдательных скважин; $\tilde{q}_j^k$ – фактический дебит j-й добывающей скважины в k-й момент времени, м3/сут; q_{ij}^k – рассчитанный по формуле (3.5) дебит j-й добывающей скважины в k-й момент времени, м3/сут. Для поиска точки минимума методом Нелдера-Мида необходимо задать $\tau_{0,ij}$ – начальное приближение для параметра τ_{ij}.

Для численного описания влияния взаимовлияния между i-ой нагнетательной и j-й добывающей скважинами используются коэффициенты MLR $p(p)_{ij}$, $p(w)_{ij}$, $q(p)_{ij}$ и $q(w)_{ij}$ и коэффициенты CRMIP f_{ij}, J_{ij}, τ_{ij} и $\tau_{0,ij}$.

Практическая реализация предложенного способа осуществлялась на примере двух месторождений Западной Сибири. Выборка для обучения и тестирования алгоритма состояла из 289 пар скважин нагнетательная–добывающая, на которых ранее проводились ТИ. На тестовой выборке полнота алгоритма составила 78%, то есть алгоритм практически полностью находит пары скважин, в которых обнаружен трассер. Также показано, что алгоритм позволяет уточнить множество скважин для проведения ТИ, повысив долю пар скважин с обнаруженным трассером с 40% до 60%. Таким образом, использование алгоритма машинного обучения позволяет оптимизировать проведение ТИ, повысить эффективность и снизить затраты на промысловые работы.

Классификатор для выявления гидродинамической связи между парами скважин в предложенном способе основан на алгоритме решающего дерева. Несмотря на то, что в последнее время наиболее часто для решения задач анализа данных используются нейронные сети или алгоритм градиентного бустинга [23], в случае ТИ накопленная цифровая база с отчетами по результатам исследований является небольшой, и при таких условиях более приемлемым явля-

ется выбор алгоритма дерева решений, поскольку требует настройки меньшего числа параметров. Выбор алгоритма также обосновывается тем, что для специалистов по анализу промысловых данных необходима интерпретируемость результатов: полученные пороговые значения в решающих правилах проверяются с точки зрения согласованности с результатами математического моделирования в гидродинамических симуляторах [77] и аналитическими решениями задач по распространению меченой жидкости в пласте [78]. На основе заключения экспертов по инженерному сопровождению исследований делается вывод о применимости алгоритма в текущем бизнес-процессе.

3.2. Явление переобучения в деревьях решений

Деревья решений классифицируют примеры, сортируя их вниз по дереву от корня до некоторого листового/конечного узла, при этом листовой/конечный узел обеспечивает классификацию примера. В каждом узле принятия решения в дереве проводится проверка некоторого дискретного атрибута, который разбивает значения непрерывного признака на конечный набор интервалов. Каждое ребро, спускающееся с узла, соответствует возможным значениям атрибута. Этот процесс носит рекурсивный характер и повторяется для каждого поддерева с корнем в новом подузле. В [61] приведено описание известных алгоритмов построения дерева решений с указанием их отличий, достоинств и недостатков.

Алгоритм генерации дерева решений генерирует дерево решений до тех пор, пока оно не может продолжиться. Недостатком данного подхода является то, что дерево часто является очень точным в классификации на обучающей выборке, но классификация на неизвестной тестовой выборке не настолько точна, то есть происходит переобучение. Причина переобучения состоит в том, что выбор атрибутов во внутренних узлах, а также выбор решений в листовых узлах, происходит по малым подвыборкам обучающей выборки – только по тем примерам, которые попадают в данный узел. Даже если исходная обучающая

выборка была репрезентативной и имела достаточно большой объём, решения в листовых узлах и внутренних узлах, далёких от корня дерева, оказываются статистически ненадёжными (смещёнными), так как производятся по малым подвыборкам. Чем больше в дереве узлов, то есть чем дерево сложнее, тем больше в нём оказывается статистически ненадёжных узлов.

Имеются различные подходы для контроля сложности дерева. Стратегии раннего усечения дерева (pre-pruning) используют ограничения на минимальное количество примеров, которое должно быть в узле перед разделением; максимальное количество листовых узлов; максимальная глубина дерева [45]. В данном подходе с использованием критерия останова в каждом узле предварительно проверяется значение критерия и затем принимается решение о целесообразности разделения узла на два или более подузла. Стратегии позднего усечения (post-pruning), применяемые в методах CART [21], C4.5 [41], C5.0 [21] сначала строят максимально сложное дерево, затем обрезают те его ветви, которые не удовлетворяют критерию качества на независимой контрольной выборке.

При выборе атрибута для разделения узла на подузлы дерево решений разбивает узел по всем доступным атрибутам, а затем выбирает разбиение, в результате которого получаются наиболее однородные подузлы по отношению к целевой переменной. Для принятия решения о выборе атрибута разработано множество числовых критериев, например, индекс Джини (Gini Index), коэффициент прироста информации (Gain Ratio) или площадь под ROC-кривой (AUC) [43, Глава 9]. Недостатком известных критериев является их смещенность, поскольку расчет производится только на фиксированном наборе примеров, попавших в узел, и никак не анализируется устойчивость результата расчета при изменении набора примеров, что приводит к переобучению.

Отсюда возникает задача модификации критерия выбора атрибута для минимизации переобученности алгоритма дерева решений, описанная в следующем разделе.

3.3. Постановка задачи

Дано: множество примеров S, бинарная целевая переменная T, принимающая значения 0 и 1; для каждого примера известны значения признаков (признаковое описание) и значения целевой переменной. Все признаки — непрерывные переменные.

Задача: модифицировать критерий выбора атрибута в алгоритме дерева решений для минимизации переобученности прогноза целевой переменной.

3.4. Предлагаемый критерий

Согласно алгоритму построения дерева решений, в узле с множеством примеров S и для данного непрерывного признака F рассматривается набор бинарных атрибутов $[F \leq \theta]$, равных 1, если выполнено $F \leq \theta$, и 0 – иначе. Пороги θ всеми возможными способами разбивают множество примеров S на два подмножества S_0, где значение атрибута равно 0, и S_1, где значение атрибута равно 1. Далее на множестве S_0 ответы дерева решений полагаются равными 0, на множестве S_1 ответы полагаются равными 1. Ошибкой называется пример, на котором ответ модели отличается от известного значения целевой переменной. В качестве атрибута выбирается классификатор $[F \leq \theta]$ с минимальным числом ошибок на множестве S, то есть реализуется метод обучения МЭР.

В разделе 1.2 было показано, что пороговые решающие правила переобучаются при выборе порога описанным способом, причем величина переобученности зависит от геометрической структуры классов. Например, при линейной разделимости классов переобученность минимальна, тогда как при наличии зашумленных данных на границе классов переобученность возрастает. Исходя из предположения, что минимизация переобученности итогового классификатора невозможна при наличии переобученных атрибутов в узлах, предлагается реализовать жадную стратегию и в каждом узле решать локально задачу мини-

мизации переобученности. Для этого предлагается использовать достигаемые верхние оценки переобучения пороговых решающих правил: критерий ожидаемой переобученности EOF и критерий полного скользящего контроля CCV. Такое решение мотивировано результатами работы [59], где было показано, что применение комбинаторных оценок в качестве критерия отбора признаков при построении элементарных конъюнкций в логических алгоритмах классификации уменьшает переобучение. Правило принятия решений для каждого класса на основе дерева решений является дизъюнкцией элементарных конъюнкций, то есть аналогично по строению изученному в [59] семейству.

Критерий EOF оценивает ожидаемую переобученность, возникающую при выборе порога θ для данного признака F. Критерий CCV оценивает ожидаемую долю ошибок классификатора $[F \leq \theta]$ на отложенном множестве примеров. Таким образом, данные критерии дают несмещенные оценки для переобученности и доли ошибок, в отличие от критериев, перечисленных в разделе 3.2. Также отметим, что в [59] была использована оценка расслоения-связности. В разделе 2.4 было показано, что данная оценка для пороговых решающих правил сильно завышена и превосходит достигаемые верхние оценки на 1–2 порядка. Преимущество предлагаемого подхода в том, что критерии EOF и CCV позволяют вычислять величину переобученности непосредственно, тогда как оценки расслоения-связности являются верхними, приводя к смещённости критериев ветвления.

Для расчета достигаемых верхних оценок EOF и CCV предлагается использовать полиномиальный алгоритм 1 для метода ПМЭР.

3.5. Псевдокод алгоритма

Проведено тестирование предлагаемых критериев путем сравнения с известными критериями Gini Index (CART) и Gain Ratio (C4.5). Для этого модифицирован алгоритм ID3: изменен критерий выбора атрибута для разделения

на два подузла и определен критерий останова – минимальный размер узла. ID3 – базовый алгоритм для построения дерева решений, при работе которого можно отследить влияние предлагаемого критерия на переобучение. Для критериев EOF и CCV выбирается признак, при котором критерий минимизируется, для критериев Gini Index и Gain Ratio – максимизируется.

Псевдокод алгоритма аналогичен приведенному в [43, Глава 9] и описан в алгоритмах 2 и 3. Критерием останова в алгоритме 2 является минимальный размер узла η. В Алгоритме 2 на шаге 8 используется понятие оптимальности. Для критериев EOF, CCV и Gini Index это значит, что необходимо искать признак, при котором критерий минимизируется, для критерия Gain Ratio – максимизируется.

Алгоритм реализован на языке C++ и доступен в репозитории GitHub [29].

3.6. Тестирование подхода

Модифицированный алгоритм был протестирован на данных с результатами проведения трассерных исследований на двух месторождениях Западной Сибири. Распределение по классам 0 «нет взаимовлияния по ТИ» и 1 «есть взаимовлияние по ТИ» указано в таблице 3.3.

	Месторождение П.	Месторождение М.	Всего
Класс 1	86	40	126
Класс 0	130	33	163
Всего	216	73	289

Таблица 3.3. Распределение по классам в данных с результатами проведения трассерных исследований на двух месторождениях Западной Сибири

Оценка алгоритма проводилась методом 15-кратной кросс-валидации. Данные случайно разделялись на обучающую и тестовую выборки в отношении 3:1. Параметр алгоритма выбран по кросс-валидации.

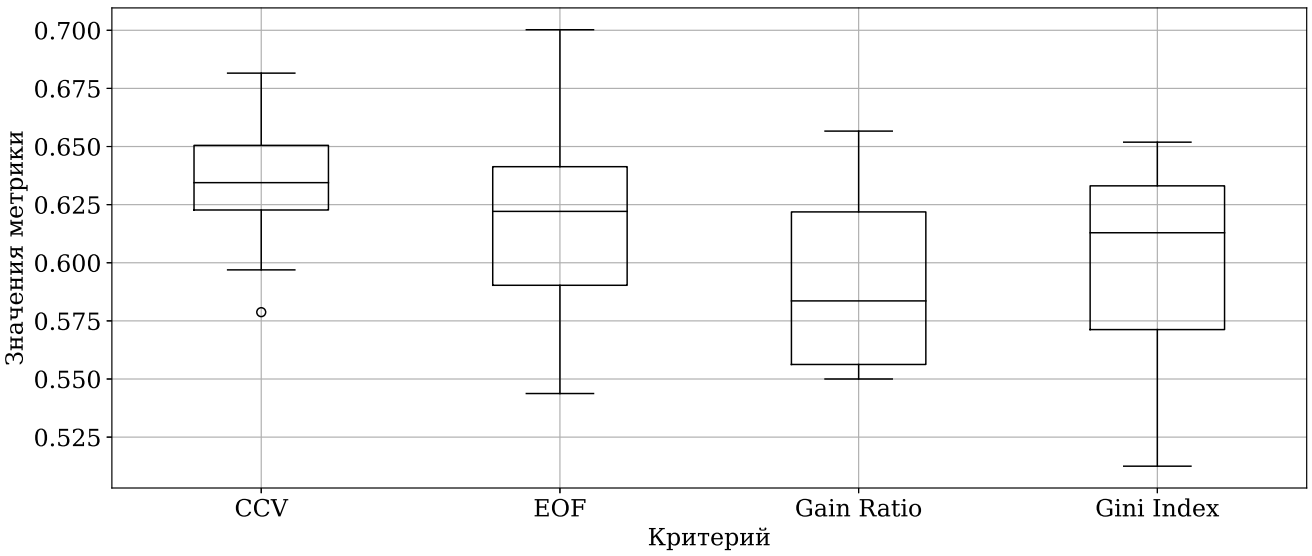


Рис. 3.1. Значения метрики AUC для различных критериев

Для оценки качества алгоритма была выбрана метрика AUC. Дерево решений для каждого примера рассчитывает вероятность принадлежности к классу 1, затем на основе порогового значения выдает ответ – метку класса. Данный функционал показывает, насколько хорошо согласуется ранжирование по рассчитываемой вероятности с ранжированием по истинным меткам классов 0 и 1. Достоинством метрики AUC является то, что она не зависит от выбора порога, недостатком является плохая интерпретируемость. Для более качественного анализа результата классификации были использованы метрики Precision и Recall. Метрика Precision равна доле пар скважин с наличием взаимовлияния (истинная метка класса – 1) среди всех пар, выбранных алгоритмом в программу исследования (ответ алгоритма – 1). Метрика Recall равна доле пар скважин с наличием взаимовлияния, которые попали в программу исследования, среди всех пар скважин с наличием взаимовлияния. Переобученность алгоритма оценивалась как разность метрики на обучающей и контрольной выборках.

Проведен статистический анализ полученных значений метрики AUC. Рассмотрены подходы для связанных выборок, поскольку каждое значение метрики получено путем фиксации пары обучающая-контрольная выборка и варьированием моделей и критериев отбора. Уровень значимости 0.05.

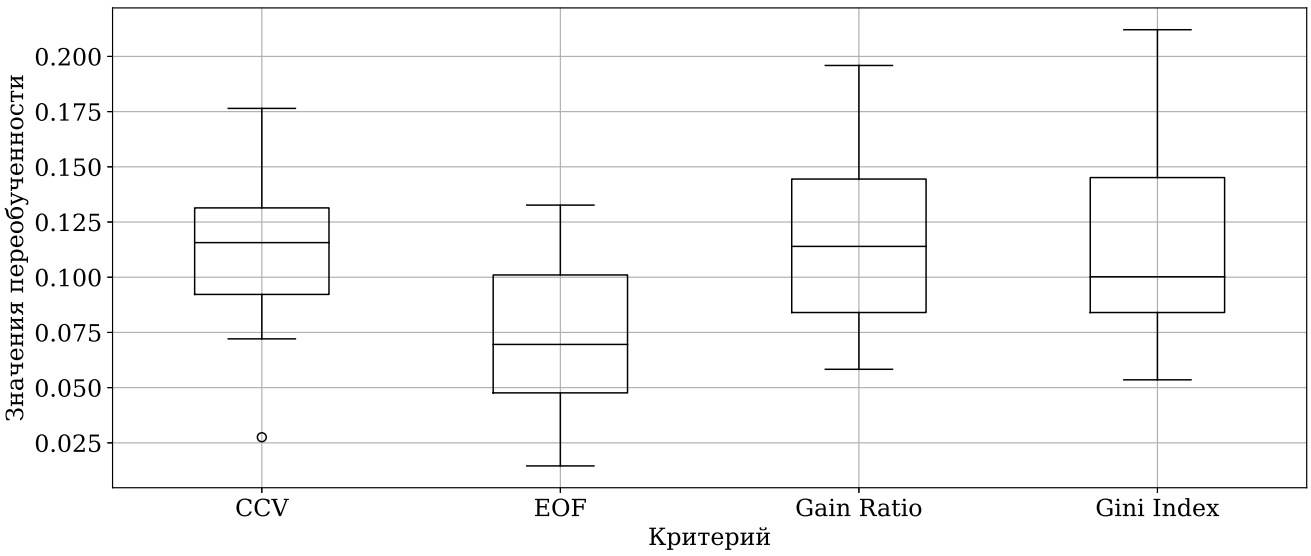


Рис. 3.2. Значения переобученности для метрики AUC для различных критериев

На рисунке 3.1 приведены боксплоты для метрики AUC для всех критериев. Однофакторный дисперсионный анализ ANOVA [50] для фактора Критерий подтверждает наличие значимого отличия метрики для различных критериев ($p\text{-}value = 0.0007$).

Средние значения метрики для критериев CCV и EOF превосходят значения для критериев Gain Ratio и Gini Index. С помощью критерия Уилкоксона [76] проверим гипотезу о равенстве средних против односторонней альтернативы. Результаты проверок гипотез приведены в таблице 3.4. После поправки на множественность методом Холма [33] отвергаются обе гипотезы для критерия CCV, откуда следует вывод о статистически значимом улучшении качества классификации при использовании критерия CCV.

Критерий 1	Критерий 2	p-value
Gain Ratio	CCV	0.0006
Gain Ratio	EOF	0.0026
Gini Index	CCV	0.0008
Gini Index	EOF	0.0962

Таблица 3.4. Результаты проверки гипотезы о равенстве средних

На рисунке 3.2 приведены боксплоты со значениями переобученности для метрики AUC. Видно, что средние значения переобученности для критерия EOF ниже, чем для остальных критериев. Поэтому проведем только две проверки гипотез о равенстве средних критерием Уилкоксона против односторонней альтернативы. Проверка для критериев EOF и Gain Ratio дает $p\text{-}value$=0.005. Проверка для критериев EOF и Gini Index дает $p\text{-}value$=0.004. Обе проверки позволяют отклонить нулевую гипотезу и подтверждают, что EOF действительно приводит к статистически значимому уменьшению переобученности.

В таблице 3.5 приведены результаты вычислений метрик качества: средние значения и 95% доверительный интервал. Жирным выделены максимальные значения по строкам, то есть указан критерий для выбора атрибута, приводящий к максимальному значению метрики. Можно увидеть, что критерии CCV и EOF дают более высокие значения метрик качества, причем чаще лучшим оказывается именно критерий CCV. Для метрики Recall использование критериев CCV или EOF дает прирост на 6%. Для метрики Precision использование критериев дает прирост на 18% относительно исходного распределения по классам, приведенного в таблице 3.3.

Метрика	Gini Index	Gain Ratio	CCV	EOF
AUC	0.60(± 0.08)	0.59(± 0.07)	**0.64(± 0.05)**	0.62(± 0.09)
Precision	0.56(± 0.12)	0.56(± 0.17)	**0.59(± 0.11)**	0.57(± 0.12)
Recall	0.55(± 0.24)	0.60(± 0.34)	0.60(± 0.20)	**0.61(± 0.26)**

Таблица 3.5. Значения метрик качества, вычисленные по кросс-валидации, и 95% доверительный интервал

В таблице 3.6 приведены значения переобученности для каждой метрики. Жирным выделены минимальные значения по строкам. Во всех строках выделены значения в столбце EOF, то есть, как и предполагалось, использование данного критерия приводит к уменьшению переобученности.

Проведенное тестирование позволяет сделать вывод о статистически значи-

Метрика	Gini Index	Gain Ratio	CCV	EOF
AUC	0.11	0.11	0.11	**0.07**
Precision	0.17	0.15	0.17	**0.14**
Recall	0.16	0.16	0.18	**0.13**

Таблица 3.6. Значения переобученности

мом улучшении качества модели Дерева решений при использовании критериев отбора признаков CCV и EOF.

3.7. Выводы к третьей главе

Показано, что имеющиеся критерии выбора атрибута для разделения узла дерева решений являются смещенными и приводят к переобучению. Для устранения этой проблемы разработаны критерии выбора атрибута, основанные на комбинаторных оценках переобучения. Данные критерии оценивают математическое ожидание доли ошибок и переобученности классификатора. Предложенные критерии апробированы на результатах проведения трассерных исследований на двух месторождениях Западной Сибири. Показано, что модификация алгоритма дерева решений приводит к уменьшению переобученности алгоритма и повышению его точности.

В дальнейшем предполагается развитие в трех направлениях: улучшение алгоритма построения плана трассерных исследований с помощью дерева решений, анализ предложенного подхода и продолжение исследований по комбинаторной теории переобучения.

В рамках первого направления предполагается использовать выводы, полученные в работе [82], где показано, что анализ корреляции накопленной концентрации трассерной жидкости в нагнетательной скважине и вынесенного трассера в добывающей скважине с помощью метода MLR позволяют уточнить наличие взаимовлияния между скважинами. Задачей будущего исследования является расширение набора используемых признаков в модели значениями данной

корреляции, что может усилить вклад подхода MLR в модель.

В рамках второго направления необходимо провести теоретический анализ предложенной модификации алгоритма построения дерева решений для исследования и обоснования масштабируемости подхода на другие предметные области и выявления отличительных особенностей применения алгоритма в задаче планирования ТИ.

В рамках третьего направления ставится задача улучшения разработанного алгоритма расчета комбинаторных оценок переобучения для устранения ограничения, связанного с большой вычислительной сложностью, которое не дает использовать его на выборках большого объема. Также необходимо исследовать другие алгоритмы, при построении которых решается задача выбора порога, например, нейронные сети. Решение поставленных задач даст универсальный критерий отбора признаков, с помощью которого удастся контролировать переобучение модели.

Алгоритм 2: Построение разбиения в узле дерева решений

Вход: множество примеров X в узле, бинарная целевая переменная T, полное множество примеров S, критерий выбора признаков I, множество признаков $\mathcal{F}$, параметр η для критерия останова.

Выход: разбиение в узле X.

проверить выполнение критерия останова;

если *критерий выполнен* **то**

  вернуть *описание листового узла: ответ дерева решений для примеров, попадающих в этот узел, заданный как класс с максимальной частотой в X;*

иначе

  для всех $F \in \mathcal{F}$

    построить атрибут $[F \leq \theta]$ для разделения на два подузла, определив порог θ по критерию минимизации ошибок классификатора $[F \leq \theta]$ на множестве X;

    вычислить значение критерия I для выбранного атрибута, обозначить его как I_F;

  выбрать признак F^*, для которого значение критерия I_F оптимально;

  определить разбиение на основе атрибута $A^* = [F^* \leq \theta]$:

$$X_{left} = \{x \in X \mid A^*(x) = 1\}$$
$$X_{right} = \{x \in X \mid A^*(x) = 0\}$$

вернуть *описание разбиения внутреннего узла – атрибут A^*, множества X_{left} и X_{right};*

Алгоритм 3: Построение дерева решений

Вход: полное множество примеров S, бинарная целевая переменная T, критерий выбора атрибута для разделения I, критерий останова.

Выход: построенное дерево решений.

```
1  определить множество 𝓕 непрерывных признаков;
2  определить корневой узел со множеством примеров S;
3  определить разбиение узла по алгоритму 2;
4  если это листовой узел то
5      вернуть описание узла — ответ алгоритма 2
6  иначе
7      определить левый подузел через S_left, правый – через S_right;
8      исключить выбранный в алгоритме 2 признак из множества 𝓕;
9      рекурсивно выполнить алгоритм 2 для правого S_right и левого S_left
         подузлов и множества 𝓕;
10 вернуть построенное дерево – набор атрибутов во внутренних
     узлах и ответы алгоритма в листовых узлах
```

Глава 4

Суррогатное моделирование для вычисления оценок обобщающей способности пороговых решающих правил

Большое количество научных и технических областей сталкивается с необходимостью компьютерного моделирования для изучения сложных явлений реального мира или решения сложных проблем проектирования. Например, чтобы найти оптимальную форму аэродинамического профиля для крыла самолета, инженер симулирует воздушный поток вокруг крыла для разных переменных формы (длина, кривизна, материал и т.д.) [17]. Несмотря на неуклонный рост вычислительных мощностей, затраты на проведение этих сложных, высокоточных симуляций и подсчет факторов прочности элементов обшивки по-прежнему огромны. Моделирование может занять много часов, дней или даже недель [25, 52]. Это затрудняет решение таких рутинных задач, как задача оптимального проектирования (design optimization), исследование пространства проектных параметров (design space exploration, DSE), анализ чувствительности (sensitivity analysis) и анализ сценариев (what-if analysis). Одним из способов решения проблемы является процесс суррогатного моделирования, состоящий в замене одних моделей другими, аппроксимационными, близкими к исходным, но более простыми в вычислительном смысле.

Построение суррогатной модели производится на основе ответов исходной модели на конечном множестве специально подобранных точек с минимальным привлечением знаний из предметной области. Данные могут быть получены из различных источников, например, из эксперимента или в результате численного моделирования. Точное внутреннее устройство исходной модели считается неизвестным, важны исключительно входные и выходные данные. Областями

применения суррогатного моделирования являются военное дело [46], строительство [71] и аэродинамика [63].

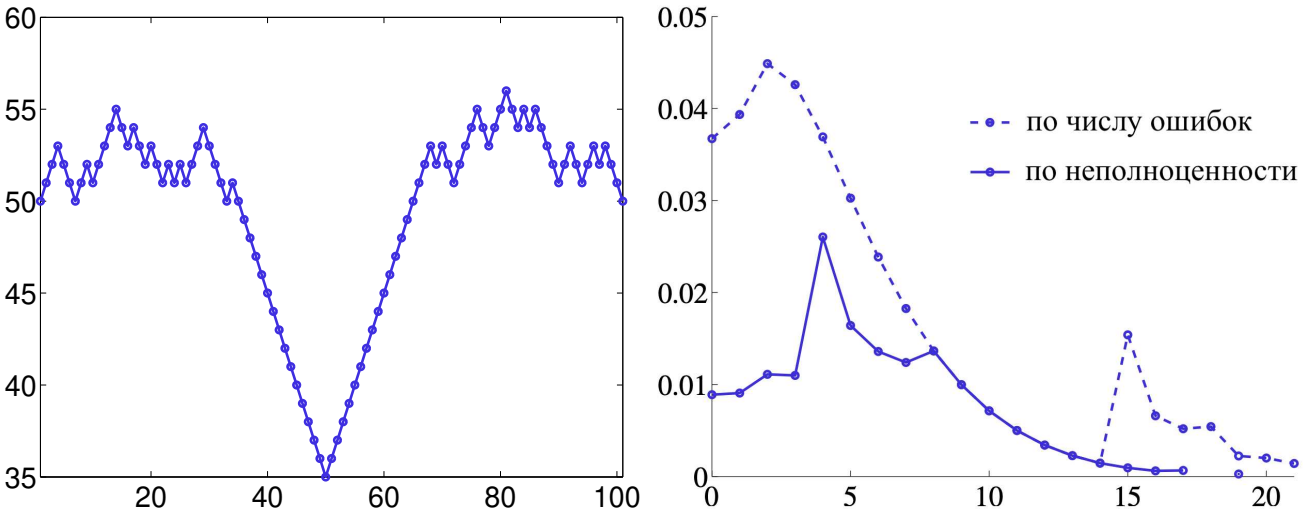
Ограничением алгоритма, описанного в главе 1, является его вычислительная сложность $O(L^5)$, что делает невозможным расчет оценок для значений L, больших 300. В данной главе суррогатное моделирование применяется в задаче получения быстро вычислимых оценок обобщающей способности семейства пороговых классификаторов над одномерными признаками. Рассматривается функционал CCV, поскольку в главе 3 была показана его практическая значимость в качестве критерия отбора признаков при построении модели машинного обучения.

4.1. Вклад порогового классификатора в переобучение семейства

Функционалы обобщающей способности, рассматриваемые в данной работе, представимы в виде суммы *вкладов* классификаторов. Это переобучение, которое возникает при выборе порогового классификатора методом обучения.

Оценка расслоения–связности (2.4) представляет собой сумму величин, которые также назовем вкладами классификаторов. Первый множитель во вкладах экспоненциально убывает по $u(a)$ – величине, определенной в (2.2) и характеризующей связность семейства, и $q(a)$ – величине, определенной в (2.3) и характеризующей расслоение. Величина $u(a)$ называется *верхней связностью*, $q(a)$ – *неполноценностью* [59].

В семействе пороговых классификаторов верхняя связность $u(a)$ принимает три значения: 0 – у локальных минимумов графика частоты ошибок (классификаторы a_2 и a_6 на рис. 1.1), 2 – у локальных максимумов (классификатор a_5 на рис. 1.1) и 1 у всех остальных.



(*а*) График частоты ошибок классификаторов. По горизонтали отложены номера классификаторов семейства.

(*б*) Зависимость вклада классификаторов от номера слоя (горизонтальная ось) и неполноценности (вертикальная ось).

Рис. 4.1. Зависимость значений вкладов классификаторов слоя $\mathbb{A}_n$ в значение функционала полного скользящего контроля от номера слоя n и от неполноценности q. Значения параметров $L = 100$, $\ell = 25$, $n_0 = 36$.

Обозначим слой с номером n через

$$\mathbb{A}_n = \{a \in \mathbb{A} \mid n = n(a, \mathbb{X}) - n_0\}.$$

Семейство $\mathbb{A}$ представимо в виде $\mathbb{A} = \mathbb{A}_0 \sqcup \cdots \sqcup \mathbb{A}_{L-n_0}$.

В [57] проводились эксперименты по исследованию зависимости вкладов классификаторов в значение функционала полного скользящего контроля от номера слоя. Было показано, что вклады убывают с возрастанием номера слоя. Следующий пример обнаруживает более сложную зависимость.

На рис. 4.1, *а* изображено семейство пороговых классификаторов при $L = 100$, $\ell = 25$, $n_0 = 36$. Данное семейство соответствует задаче классификации со сбалансированными классами.

На рис. 4.1, *б* изображена зависимость среднего значения вклада классификатора a от номера слоя n и неполноценности $q(a)$. Усреднение производится по всем классификаторам из одного слоя с равным q.

Графики демонстрируют, что вклады велики только у классификаторов с небольшими значениями неполноценности. Но в то же время вклады у клас-

сификаторов с равной неполноценностью убывают с увеличением номера слоя. Действительно, на кривой зависимости вклада от номера слоя имеется два пика на расстоянии d, приблизительно равном 15. Правый пик соответствует локальным минимумам, выделенным на рис. 4.1, *а*, при наличии которых на слоях с большими номерами появляются классификаторы с малой неполноценностью. Примеры пар классификаторов с равной неполноценностью отмечены на рис. 4.1, *а*, они расположены на слоях n и $n + d$. Вклады у классификаторов со слоя n выше, чем у классификаторов со слоя $n + d$.

Таким образом, можно сделать следующие выводы:

- значимый вклад вносят классификаторы с небольшими значениями неполноценности $q(a)$;
- вклады убывают с возрастанием номера слоя, но необходимо учитывать их неполноценность;
- на вклад влияет значение верхней связности $u(a)$, то есть то, является ли классификатор локальным минимумом или максимумом.

Данные наблюдения будут использоваться далее при построении признаков для суррогатных моделей.

4.2. Суррогатное моделирование

Суррогатную модель можно понимать как «модель» модели. Она описывает взаимосвязь между входами (то есть регулируемыми параметрами модели) и выходами истинной, имитационной, модели.

Суррогатное моделирование состоит из трех шагов. Первый шаг – это построение обучающей выборки $\mathbb{Y}$ – множества пар $(\mathbb{A}, \mathcal{F})$ объектов и ответов на них, которые получаются путем запуска имитационной модели с различными наборами параметров, выбранных в допустимом пространстве параметров.

В данной работе каждым объектом $\mathbb{A}$ является семейство одномерных пороговых классификаторов, ответом $\mathcal{F}$ – достигаемая верхняя оценка полного скользящего контроля семейства. Второй шаг – это построение некоторым образом подобранных признаков для описания объектов. Третий шаг – это построение регрессионной модели для последующего быстрого вычисления ответов на новых объектах.

4.2.1. Построение обучающей выборки

При построении обучающей выборки пар $(\mathbb{A}, \mathcal{F})$ каждый объект $\mathbb{A}$ устроен следующим образом. Генерируем класс нулей X_0 и класс единиц X_1 как выборки из нормальных распределений и строим для них семейство пороговых классификаторов $\mathbb{A}$ варьированием порога. В определенных выше обозначениях объединение выборок X_0 и X_1 является множеством $\mathbb{X}$.

Зафиксируем значение математического ожидания распределения класса нулей, задав его равным 0. Репрезентативность обучающей выборки обеспечим путем варьирования степени погруженности одного класса в другой, который определяется значением Δ математического ожидания распределения класса единиц. Стандартное отклонение обоих распределений зададим равным 1. На рисунке 4.2 изображен пример таких классов.

Параметр Δ принимает значения, равные удвоенным квантилям u_α стандартного нормального распределения $\xi \sim \mathcal{N}(0,1)$, где под квантилем понимается число u_α, такое, что

$$P[\xi \leqslant u_\alpha] = \alpha.$$

Отметим, что плотности распределений $x_0 \sim \mathcal{N}(0,1)$ и $x_1 \sim \mathcal{N}(2u_\alpha, 1)$ симметричны относительно значения u_α, поскольку справедливо равенство $P(x_0 \leq u_\alpha) = P(x_1 \geq u_\alpha)$. Таким образом, варьируя значения α, мы получим всевозможные взаимные расположения двух классов. Ограничимся теми, где класс единиц находится правее класса нулей, поскольку в обратном

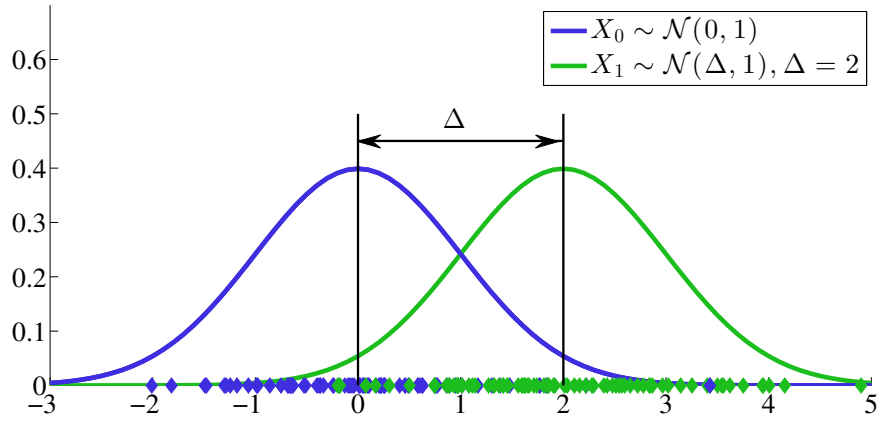


Рис. 4.2. Выборка классов из двух нормальных распределений $X_0^{n_0} \sim \mathcal{N}(0,1)$ и $X_1^{n_1} \sim \mathcal{N}(\Delta,1)$, используемая для построения семейства пороговых классификаторов. Варьируемые параметры: Δ – расстояние между центрами распределений, n_0, n_1 – размеры классов. Стандартное отклонение равно 1. Значения параметров $\Delta = 2$, $n_0 = n_1 = 80$

случае в задаче классификации необходимо переопределить пороговое правило принятия решений, изменив знак $\geq$ на обратный. Отсюда следует, что α пробегает значения $0.6, \ldots, 1$.

Мощность множества $\mathbb{X}$ равна $L = n_0 + n_1$ и пробегает значения $50, \ldots, 320$. Для каждого L параметр ℓ принимает значения 0.15, 0.3, 0.5, 0.9. Отношение $n_0 : n_1$ размеров классов пробегает значения $0.25 \ldots 4$. Значения параметра L ограничиваются вычислительными мощностями рабочего компьютера, поскольку для каждого семейства требуется вычислить достигаемую верхнюю оценку ыполного скользящего контроля $\mathcal{F} = \mathsf{CCV}(\mu, \mathbb{A}, \mathbb{X}, \ell)$ по алгоритму 1 со сложностью $O(L^5)$.

Полученное множество $\mathbb{Y}$ пар $(\mathbb{A}, \mathcal{F})$ используется для построения суррогатной модели. Мощность множества $\mathbb{Y}$ составляет 14 000 объектов.

4.2.2. Построение признакового пространства

Для построения признаков воспользуемся свойствами оценки расслоения–связности (2.4). Необходимо описывать некоторым образом слои семейства с учетом неполноценностей классификаторов и того, является ли классификатор

локальным минимумом или максимумом графика частоты ошибок на множестве $\mathbb{X}$.

В работе [83] обобщающая способность слоя $\mathbb{A}_n$ классификаторов оценивалась сверху через подмножество хеммингова шара булевого куба, подможеством которого представлялось исходное семейство $\mathbb{A}$. Завышенность оценки зависела от того, насколько близкими в смысле расстояния Хемминга являлись классификаторы в слое $\mathbb{A}_n$. Поэтому будем использовать следующие признаки:

- объем обучающей выборки ℓ;
- минимальное число ошибок классификаторов n_0;
- мощность слоя $\{|\mathbb{A}_n|\}_{n=0}^{L-n_0}$;
- среднее значение неполноценности $q(a)$ в слое;
- профиль расслоения-связности [57]

$$\Delta(n,u) = |\{a \in \mathbb{A}_n \mid u(a) = u\}|,$$

  равный количеству локальных минимумов (при $u = 2$) и максимумов (при $u = 0$) в слое;
- среднее расстояние Хемминга между классификаторами в слое;
- количество локальных минимумов и максимумов с равной неполноценностью;
- среднее значение числа ошибок у классификаторов с равной неполноценностью;

Разделим все признаки на L – мощность множества $\mathbb{X}$ – и добавим признак, равный L.

Кроме того, для уменьшения количества признаков будем считать, что классификаторы из соседних слоев вносят одинаковый вклад. Поэтому будем

группировать слои на B блоков путем усреднения значений признаков по $\frac{L}{B}$ соседних слоев, где B – параметр.

Поскольку для каждого n сумма $\Delta(n,0) + \Delta(n,1) + \Delta(n,2) = |\mathbb{A}_n|$, то для устранения мультиколлинеарности профиль расслоения-связности будем вычислять только для $u = 0$ и $u = 2$. Таким образом, всего имеется $8 \cdot B + 3$ признаков.

4.2.3. Построение модели

Будем использовать методы построения моделей, позволяющие получать интерпретируемые результаты и выявлять признаки и характеристики семейства, которые дают значимый вклад в его обобщающую способность. Рассмотрим следующие модели:

- линейные на основе модели elastic net EN с возможностью задания l_2- или l_1- регуляризации и ограничения на неотрицательность коэффициентов;
- на основе дерева решений: случайного леса RF и градиентного бустинга GB;
- нейронную сеть NN.

Преимуществом линейных моделей помимо интерпретируемости является их малая вычислительная сложность. Линейные модели с l_1-регуляризацией или ограничением на неотрицательность коэффициентов проводят отбор признаков, что приводит к сокращению временных затрат на вычисление ответа. Модель с ограничением на неотрицательность коэффициентов позволяет гарантировать неотрицательность ответов, вычисляемых моделью.

Преимуществом моделей на основе дерева решений является возможность обнаружения нелинейных взаимосвязей между признаками и целевой переменной. Для того чтобы снять указанное ограничение для линейных моделей и нейронных сетей в реализации библиотеки sklearn [48], при построении EN и

NN добавим полиномиальные признаки в набор данных (обозначим модели как EN+Poly и NN+Poly соответственно).

4.3. Анализ суррогатных моделей

Качество модели на выборке Y будем оценивать по метрике среднего относительного отклонения MAPE, учитывающего абсолютное значение целевого и приближенного значений:

$$MAPE(Y) = \frac{1}{k} \sum_{(\mathbb{A},\mathcal{F})\in Y} \frac{\left|\hat{\mathcal{F}} - \mathcal{F}\right|}{\mathcal{F}},$$

где $\hat{\mathcal{F}}$ – ответ, вычисленный с помощью модели, на объекте выборки $(\mathbb{A}, \mathcal{F}) \in Y$, k – объем выборки Y. Меньшие значения метрики соответствуют более высокому качеству модели.

Пусть множество $\mathbb{Y}$ разделено на обучающую выборку Y_t и контрольную Y_v в отношении 5:1, то есть $\mathbb{Y} = Y_t \sqcup Y_v$. Гиперпараметры для каждой модели будем подбирать по 5-блочной кросс-валидации на обучающей выборке Y_t.

Проведем анализ рассматриваемых суррогатных моделей. Для начала сравним точность моделей на выборке Y_v при различных значениях параметра B из списка $5, 10, 15, 20$. Также исследуем, насколько модель устойчива при увеличении L – мощности множества $\mathbb{X}$ или, другими словами, размера семейства классификаторов. Цель построения суррогатной модели – вычислять приближенные оценки обобщающей способности в тех случаях, когда достигаемые верхние оценки с помощью алгоритма вычислить становится невозможно ввиду больших временных затрат, обусловленных параметром L. Поэтому модель должна быть масштабируема по параметру L, то есть давать достаточно точные приближенные оценки CCV при больших L, тогда как обучение модели проводится на объектах при малых значениях данного параметра.

На множестве $\mathbb{Y}$ и выборке Y_v значения параметра L не превосходят 300 в силу больших временных затрат на вычисление точных значений ответов.

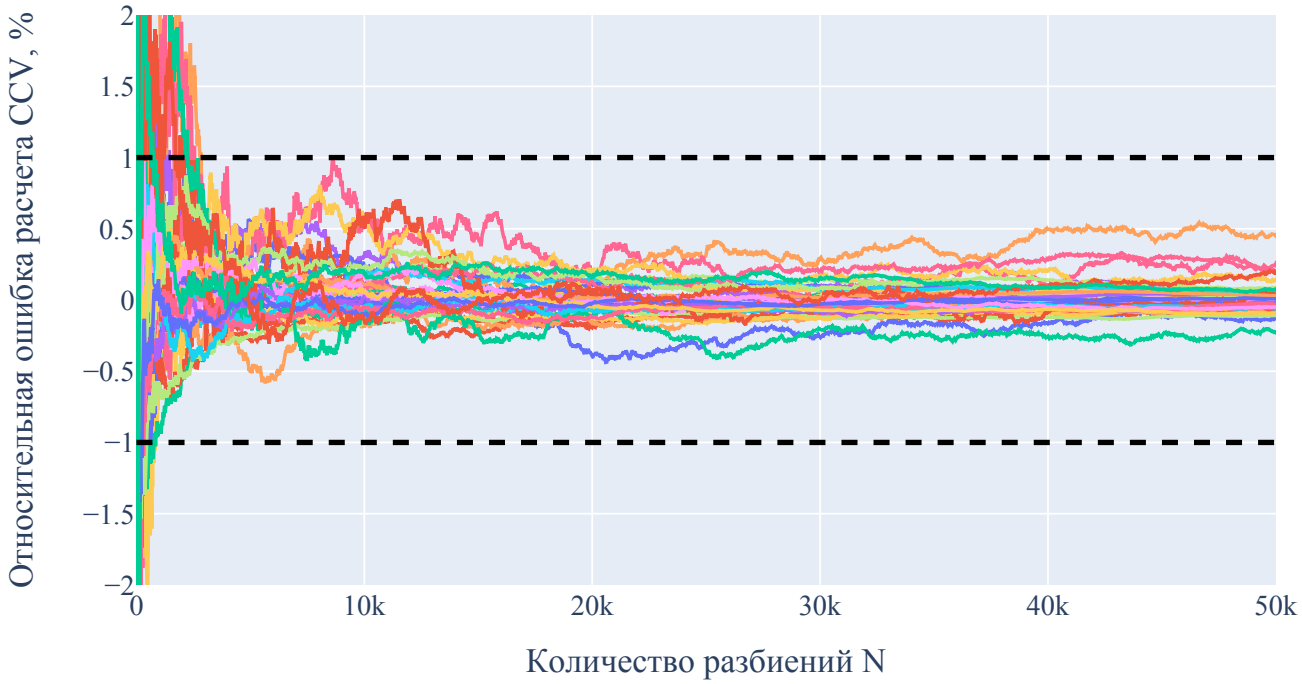
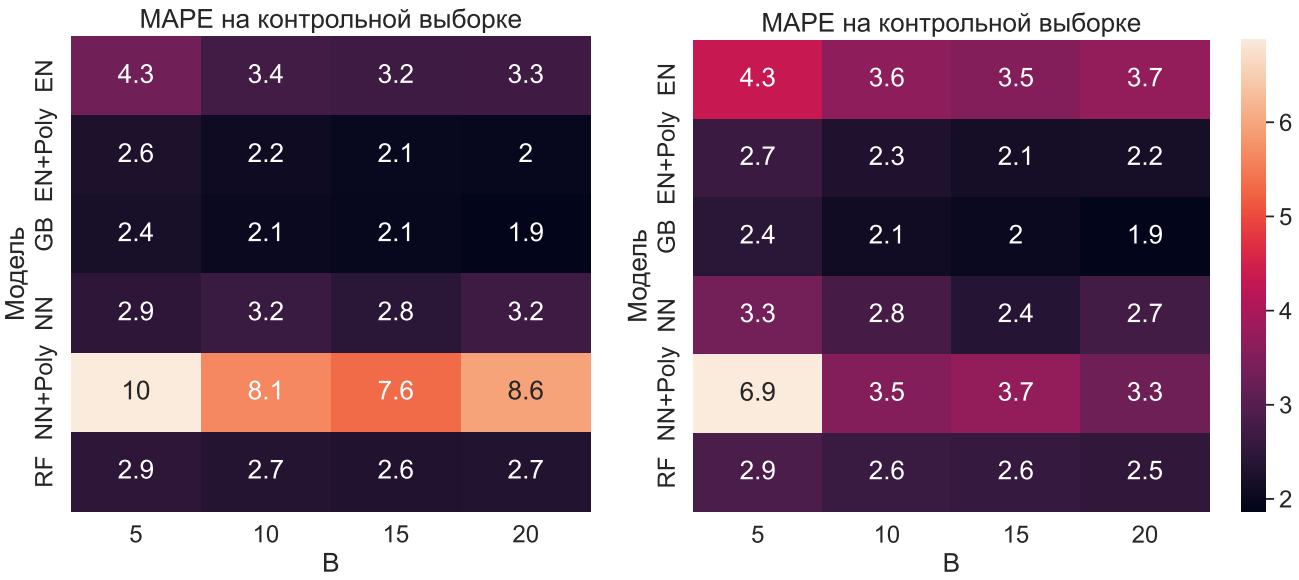


Рис. 4.3. Зависимость относительной ошибки расчета CCV методом Монте–Карло от количества N разбиений. Горизонтальные линии соответствуют ошибке в 1%

Построим множество Y_{mc} объектов с большими значениями параметра L в диапазоне $400, \ldots, 800$.

Вычислим для этих объектов приближенные ответы методом Монте–Карло: усредним значения $\nu(\mu X, \bar{X})$ по N случайным разбиениям $(X, \bar{X})$ множества $\mathbb{X}$. На рисунке 4.3 изображена относительная ошибка расчета CCV методом Монте–Карло для 30 случайных объектов в зависимости от N, параметры $L = 800, \ell = 400$.

По горизонтали отложены значения параметра N. Видно, что начиная с $N = 10\ 000$ относительная ошибка стабилизируется и не превышает значения 1%. Для исключения внесения погрешности в результаты тестирования проведем расчеты с $N = 50\ 000$. Отметим, что для 10 000 разбиений время вычисления CCV для одного объекта может достигать 10 секунд, что на практике неприменимо.



(*а*) Без отбора признаков

(*б*) С отбором признаков

Рис. 4.4. Сравнение моделей на контрольной выборке

4.3.1. Точность моделей на контрольной выборке

На рисунке 4.4, *а* показано сравнение моделей на контрольной выборке Y_v по метрике $MAPE(Y_v)$ при разных значениях параметра B.

Для уменьшения сложности моделей также были рассмотрены различные методы сокращения размерности. Метод на основе взаимной информации [22] позволил предварительно убрать нерелевантные признаки. Наилучший результат по точности модели удалось достичь при дальнейшем применении метода на основе оценивания вектора Шепли [28]. Метод Шепли основан на концепции кооперативной игры и теории распределения выигрышей между игроками. В машинном обучении выигрышем считается верный ответ модели, игроками являются признаки. Метод позволяет оценить вклад признака в прогнозы модели, для чего перебираются все возможные комбинации признаков и рассчитывается разница в ответах модели при добавлении или удалении конкретного признака.

На рисунке 4.4, *б* приведены значения метрики после отбора признаков. Можно видеть, что уменьшение сложности не снижает точность модели, поэтому далее проводится анализ для моделей с отбором признаков.

Видно, что наилучшие значения метрики достигаются при использовании

моделей GB, EN+Poly, NN и RF. При этом с увеличением значений параметра B качество улучшается. Но с увеличением параметра B растет и количество вычисляемых признаков, то есть возникает необходимость найти компромисс (trade-off) между сложностью модели и ее точностью. Отметим, что у модели GB более низкие значения MAPE достигаются при малых значениях параметра $B = 10$, что позволяет сократить время на расчет признаков. Проанализируем устойчивость выделенных моделей.

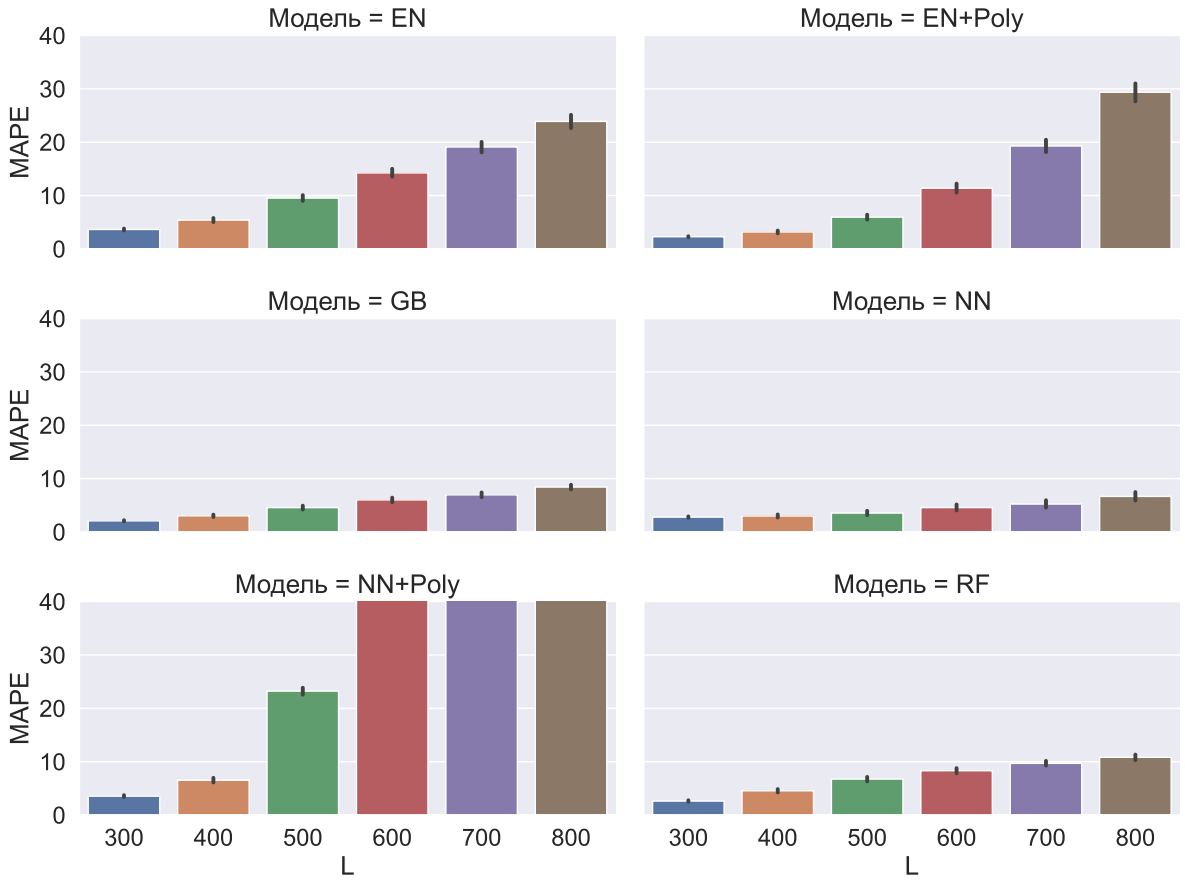
4.3.2. Устойчивость моделей к увеличению размера семейства классификаторов

На рисунке 4.5, *а* приведены значения метрики качества на контрольной выборке Y_v (в столбце $L = 300$) и на множестве Y_{mc} (в столбцах $L > 300$) при $B = 10$. Для других значений параметра B поведение моделей аналогично.

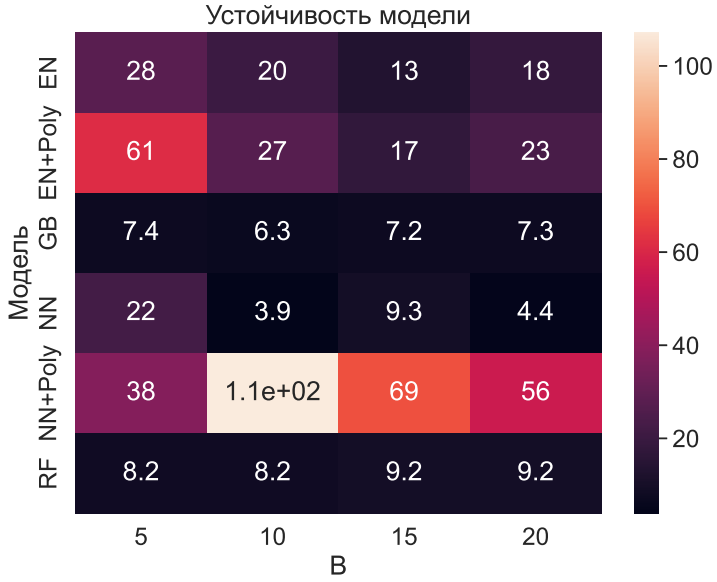
Высота столбцов увеличивается с увеличением L, поэтому как оценку устойчивости рассмотрим максимальную разницу в значениях метрики качества: при $L = 800$ на множестве Y_{mc} и при $L = 300$ на выборке Y_v. На рисунке 4.5, *б* показано сравнение моделей по данной оценке при разных значениях параметра B. Более низкие значения соответствуют более устойчивой модели.

Видно, что наилучшими по рассматриваемому критерию являются модели GB (при $B = 10$) и NN (при $B = 10$ и $B = 20$). Модель EN+Poly показывает низкое качество на множестве Y_{mc} при $L = 800$ по сравнению с выборкой Y_v, откуда делаем вывод о ее неустойчивости при увеличении семейства классификаторов и непригодности для решения поставленной задачи. Модель RF оказывается менее устойчивой по сравнению с GB.

На основе проведенного анализа можно утверждать, что наилучшими суррогатными моделями являются модели градиентного бустинга GB и нейронной сети NN. Поскольку нейронная сеть более устойчива при $B = 10$, тогда как для градиентного бустинга выбор однозначен, то для обеих моделей выберем значение параметра $B = 10$.



(*а*) Значения метрики на Y_{MC} при увеличении параметра L



(*б*) Разность метрики на Y_{mc} при $L = 800$ и на Y_v при $L = 300$

Рис. 4.5. Анализ устойчивости моделей

Гиперпараметры для модели GB: глубина деревьев 25, количество деревьев 300; для модели NN: количество слоев 2, количество узлов в одном слое 10,

функция активации $ReLU$.

4.3.3. Вычислительная сложность моделей

Сложность вычисления ответа для одного объекта в модели градиентного бустинга составляет $O(T \cdot H)$, где T – количество деревьев, H – высота деревьев, в однослойной нейронной сети – $O(K)$, где K – количество узлов в одном слое. С учетом значений гиперпараметров T, H, K, приведенных выше, можно утверждать, что расчеты на основе нейронной сети выполняются за меньшее время. Затраты по памяти на хранение модели градиентного бустинга имеют аналогичную асимптотику и оказываются больше, чем у нейронной сети.

Вследствие этого на практике будем использовать модель NN при $B = 10$.

Признаки вычисляются за квадратичное по L время, значит, сложность вычисления ответа линейной модели составляет $O(L^2)$. На рассмотренной выборке для $L = 300$ вычисление признаков и ответа модели не превосходило 0.4 секунды, тогда как время расчета по алгоритму 1 достигало 400 секунд. Таким образом, использование суррогатного моделирования уже на малых выборках позволяет сократить время вычисления оценки обобщающей способности в 1000 раз.

4.4. Обсуждение результатов

Проанализируем значимость признаков в модели NN при $B = 10$.

Как уже было сказано выше, при построении модели предварительно был проведен отбор признаков на основе метода Шепли. Для $B = 10$ алгоритм на основе описанного метода Шепли выделяет в качестве значимых признаки, описывающие 20% нижних слоев. Таким образом, суррогатное моделирование подтверждает эмпирическое наблюдение о том, что вклад в значение переобучения вносят только нижние слои.

Для интерпретируемости результатов признаки были разделены на груп-

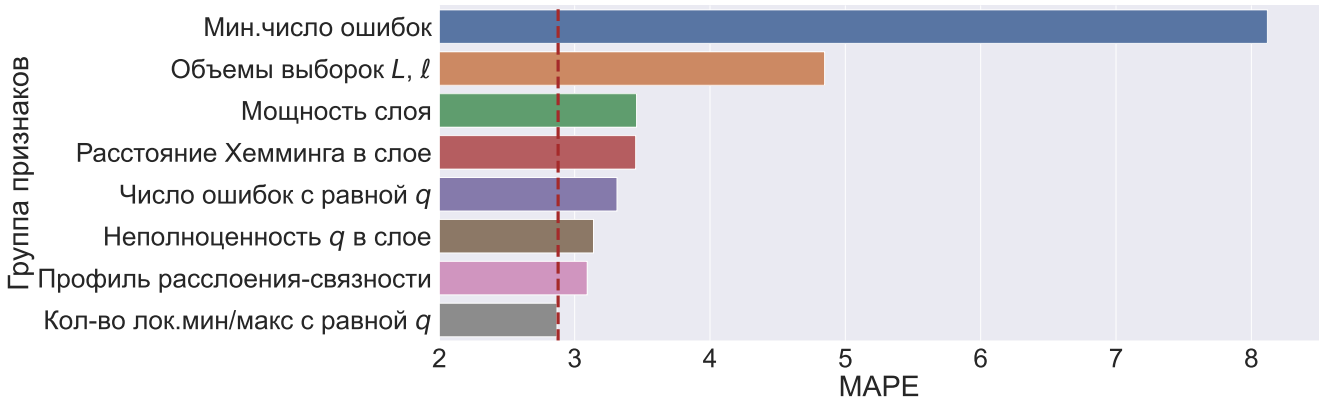


Рис. 4.6. Значение метрики при удалении группы признаков из модели NN. Вертикальной линией обозначено значение метрики при добавлении всех признаков в модель

пы, описанные в разделе 4.2.2, и анализ проводился для каждой группы. Значимость оценивалась следующим образом: из модели, построенной на полном наборе признаков, удалялась группа и вычислялась метрика MAPE относительной ошибки расчета CCV. Рассматривалась выборка при $L = 300$, то есть ответы модели сравнивались с достигаемыми верхними оценками переобучения.

Результаты расчета приведены на рисунке 4.6. Вертикальной линией обозначено значение метрики для модели на полном наборе признаков. Можно видеть, что наибольшую значимость имеют признаки, отвечающие за минимальное число ошибок и размерные характеристики задачи – объем обучающей выборки ℓ и параметр L. Признаки, описывающие мощности слоев ошибок, вносят равный вклад с признаками, отвечающие за расстояние Хемминга в слое. Также надо учитывать взаимосвязь между числом ошибок классификатора и его неполноценностью и профиль расслоения–связности. Связь между профилем расслоения–связности и неполноценностью, выражаемую через количество локальных минимумов/максимумов с равной неполноценностью, не вносит вклада в модель, этот признак можно исключить.

На рисунке 4.7 показано сопоставление истинных и рассчитанных значений CCV при последовательном добавлении наиболее значимых признаков в модель: минимального числа ошибок, объемов выборок и мощности слоев. Можно видеть, что рассеянность точек относительно диагональной линии при добавлении

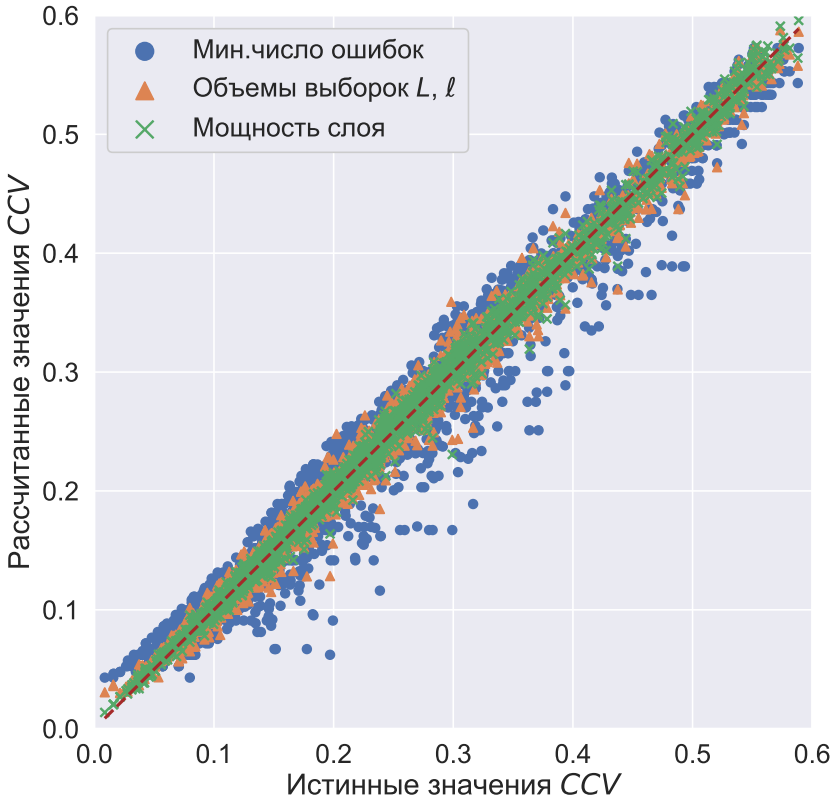


Рис. 4.7. Графики рассеяния при добавлении группы признаков в модель NN

очередной группы признаков уменьшается, то есть рассчитанные значения приближаются к истинным, при этом не наблюдается стабильного завышения или занижения ответов модели. Также интересно отметить, что именно добавление мощности слоев в модель уточняет ее ответы на отрезке $[0, 0.1]$.

Таким образом, можно сделать следующие выводы. Значений минимального числа ошибок и размерных характеристик семейства недостаточно, учет геометрической структуры семейства необходим при построении модели. Полученные результаты согласуются с результатами работы [59]: при вычислении оценок обобщающей способности значимый вклад вносят эффекты связности семейства и расслоения по числу ошибок.

4.5. Выводы к четвертой главе

Проведено построение суррогатной модели для быстрого вычисления оценок переобучения семейства пороговых решающих правил. Рассмотрены модели различной структуры, наилучшей по результатам тестирования выбрана модель нейронной сети с MAPE=2.8% и устойчивостью при увеличении размера выборки, на которой решается задача классификации с использованием пороговых решающих правил.

Интерпретация результатов и анализ значимости признаков показали, что учет только объема обучающей выборки и минимального числа ошибок классификаторов недостаточен, необходимо использовать в модели внутреннюю структуру семейства и взаимосвязь между классификаторами. Построение модели и отбор признаков подвердили результаты ранних работ и эмпирические наблюдения о том, что вклад в переобучение вносят только классификаторы из нижних слоев семейства, то есть те, которые допускают наименьшее число ошибок.

На основе полученных результатов следующей задачей становится исследование возможности аппроксимации семейства пороговых рещающих правил меньшим семейством, состоящим из классификаторов из нижних слоев. Другим направлением является применение построенной суррогатной модели в практических задачах при отборе признаков.

Заключение

Основные результаты данной работы заключаются в следующем:

1. Доказаны теоремы о представлении достигаемых верхних оценок обобщающей способности произвольного семейства классификаторов в виде произведения числа разбиений двух непересекающихся множеств объектов генеральной совокупности.

2. Доказаны теоремы и разработан алгоритм полиномиальной сложности для вычисления достигаемых верхних оценок обобщающей способности семейства пороговых решающих правил над одномерным признаком при варьировании параметра порога. В качестве характеристик обобщающей способности используются функционалы вероятности переобучения, полного скользящего контроля и ожидаемой переобученности. Алгоритм основан на рекуррентном подсчете числа допустимых траекторий при блуждании по трехмерной сетке между двумя заданными точками с ограничениями специального вида.

3. Проведен анализ завышенности известных оценок вероятности переобучения: Вапника-Червоненкиса, расслоения–связности и Соколова. Показано, что данные оценки завышены по сравнению с достигаемыми верхними оценками, рассчитанными с помощью полученного алгоритма.

4. Полученный алгоритм применен для анализа завышенности известной оценки Гуза для полного скользящего контроля. Показано, что оценки Гуза являются достаточно точными для применения в реальных задачах. Однако данные оценки накладывают требования на распределение значений одномерного признака, то есть применимы только в частных случаях.

5. Проведен анализ завышенности оценки частоты ошибок на контрольной выборке на основе Радемахеровской сложности по сравнению с достигае-

мыми верхними оценками, рассчитанными с помощью полученного алгоритма. Показано, что данные оценки оказываются точными только для задач с высоким уровнем шума на границе классов. В противном случае, когда граница между классами определяется четко, оценки Радемахеровского типа являются завышенными на несколько порядков и неприменимыми на практике.

6. Полученные достигаемые верхние оценки полного скользящего контроля и ожидаемой переобученности применены в качестве критерия отбора признаков при построении дерева решений. Проведены эксперименты на промысловых данных трассерных исследований и показано статистически значимое повышение обобщающей способности при использовании комбинаторных оценок в деревьях решений.

7. Построена суррогатная модель для быстрого вычисления приближенных оценок обобщающей способности семейства пороговых решающих правил с высокой точностью. Показано, что использование суррогатного моделирования позволяет сократить на несколько порядков время вычисления оценок переобучения даже на выборках малого объема и может применяться в практических задачах для отбора признаков при построении моделей машинного обучения.

Список иллюстраций

1.1 Пример прямой цепи 21

1.2 Сравнение переобучения прямых цепей различной формы. По горизонтали отложены номера p классификаторов цепи. Условия эксперимента: $L = 200$, $\ell = 150$, $\varepsilon = 0,05$. Минимальная частота ошибок равна $0,245$. 22

1.3 Соответствие разбиения цепи (нижний график) проекции траектории (верхний график). Двойными линиями выделены ребра цепи, попавшие в обучающую выборку 35

2.1 Сравнение верхних оценок вероятности переобучения в логарифмической шкале. Условия эксперимента: $L = 240$, $\ell = 160$, $m = 20$, $\varepsilon = 0,05$. По горизонтали отложено минимальное количество ошибок классификаторов. 51

2.2 Сравнение верхних оценок полного скользящего контроля в логарифмической шкале. Условия эксперимента: $L = 240$, $\ell = 160$, $m = 0$. По горизонтали отложено минимальное количество ошибок классификаторов. 52

2.3 Сравнение оценок ожидаемой переобученности для методов ПМ-ЭР и МП в логарифмической шкале. Условия эксперимента: $L = 240$, $\ell = 120$, $m = 0$. По горизонтали отложено минимальное количество ошибок классификаторов. 52

3.1 Значения метрики AUC для различных критериев 67

3.2 Значения переобученности для метрики AUC для различных критериев 68

4.1 Зависимость значений вкладов классификаторов слоя $\mathbb{A}_n$ в значение функционала полного скользящего контроля от номера слоя n и от неполноценности q. Значения параметров $L = 100$, $\ell = 25$, $n_0 = 36$. 76

4.2 Выборка классов из двух нормальных распределений $X_0^{n_0} \sim \mathcal{N}(0,1)$ и $X_1^{n_1} \sim \mathcal{N}(\Delta,1)$, используемая для построения семейства пороговых классификаторов. Варьируемые параметры: Δ – расстояние между центрами распределений, n_0, n_1 – размеры классов. Стандартное отклонение равно 1. Значения параметров $\Delta = 2$, $n_0 = n_1 = 80$ 79

4.3 Зависимость относительной ошибки расчета CCV методом Монте–Карло от количества N разбиений. Горизонтальные линии соответствуют ошибке в 1% 83

4.4 Сравнение моделей на контрольной выборке 84

4.5 Анализ устойчивости моделей 86

4.6 Значение метрики при удалении группы признаков из модели NN. Вертикальной линией обозначено значение метрики при добавлении всех признаков в модель 88

4.7 Графики рассеяния при добавлении группы признаков в модель NN 89

Список таблиц

3.1 Обозначения динамических параметров 59
3.2 Обозначения коэффициентов в методе MLR 60
3.3 Распределение по классам в данных с результатами проведения трассерных исследований на двух месторождениях Западной Сибири 66
3.4 Результаты проверки гипотезы о равенстве средних 68
3.5 Значения метрик качества, вычисленные по кросс-валидации, и 95% доверительный интервал 69
3.6 Значения переобученности 70

Список обозначений

$\mathbb{A}$	семейство классификаторов (Стр.16)
a, a_p	классификатор/классификатор с номером p (Стр.16)
L	мощность генеральной совокупности объектов (Стр.16)
ℓ	объем обучающей выборки (Стр. 16)
$\mathbb{X}$	генеральная совокупность объектов (Стр.16)
X	обучающая выборка объектов (Стр.16)
$\bar{X}$	контрольная выборка объектов (Стр.16)
$[\mathbb{X}]^\ell$	множество всех подмножеств $\mathbb{X}$ мощности ℓ (Стр.16)
$n(a, X)$	число ошибок классификатора a на выборке X (Стр.16)
$\nu(a, X)$	частота ошибок классификатора a на выборке X (Стр.16)
$\delta_p(a, X)$	переобученность классификатора a на выборке X (Стр.17)
$\mathsf{P}\{\}$	вероятность события (Стр.16)
$\mathsf{E}\{\}$	математическое ожидание события (Стр.17)
ε	вещественное число от 0 до 1 (Стр.17)
μ	метод обучения (Стр.17)
CCV	полный скользящий контроль (Стр.17)
EOF	ожидаемая переобученность (Стр.17)
Q_ε	вероятность переобучения (Стр.17)
P	мощность семейства классификаторов (Стр.18)
p	номер классификатора (Стр.18)
x	объект множества $\mathbb{X}$/значение одномерного признака (Стр.19)
$\Delta(a, X)$	запас ошибок классификатора a на выборке X (Стр.24)
H	гипергеометрическая функция распределения (Стр.26)

$\mathbb{D}$	множество ребер семейства классификаторов (Стр.26)
$\mathbb{N}$	нейтральное множество объектов (Стр.26)
m	число ошибок семейства на множестве $\mathbb{N}$ (Стр.26)
t	множество объектов из $X \cap \mathbb{D}$ (Стр.26)
e	число ошибок классификатора a_p на $X \cap \mathbb{D}$ (Стр.26)
D_p	количество разбиений множества $\mathbb{D}$ (Стр.26)
$\mathbb{L}_p$	множество ребер левой последовательности (Стр.31)
$\mathbb{R}_p$	множество ребер правой последовательности (Стр.31)
L_p	количество разбиений множества $\mathbb{L}_p$ (Стр.31)
R_p	количество разбиений множества $\mathbb{R}_p$ (Стр.31)
Ω_p	трехмерная сетка (Стр.33)
$\mathbb{T}_p$	множество траекторий на Ω_p (Стр.33)
T_p	количество траекторий из $\mathbb{T}_p$ с ограничениями (Стр.35)
$q(a)$	неполноценность классификатора a (Стр.45)
$u(a)$	верхняя связность классификатора a (Стр.45)
$\mathcal{R}_L(\mathbb{A}, \mathbb{X})$	Радемахеровская сложность семейства классификаторов (Стр.47)
МП	метод максимизации переобученности (Стр.18)
МЭР	метод минимизации эмпирического риска (Стр.18)
ПМЭР	метод пессимистичной минимизации эмпирического риска (Стр.18)
ТИ	трассерное исследование (Стр.54)
ГП	гидропрослушивание (Стр.57)
CRMIP	метод емкостно-резистивной модели (Стр.58)
MLR	метод многопараметрической регрессии (Стр.58)
MAPE	среднее относительное отклонение (Стр.82)

Публикации автора по теме диссертации

1. Ishkina Sh. Kh., Vorontsov K. V. Sharpness estimation of combinatorial generalization ability bounds for threshold decision rules // Autom. Remote Control. 2021. V. 82. No. 5 P. 863–876. doi:10.31857/S000523102105010X
2. Бикметова А. Р., Фахреева Р. Р., Ишкина Ш. Х., Питюк Ю. А. Комплексный подход к анализу взаимовлияния скважин по динамическим данным эксплуатации скважин // Геолого-геофизические исследования нефтегазовых пластов: сборник научных статей по материалам VI Всероссийской молодежной научно-практической конференции (Уфа, 27 мая 2021 года). Уфа: Башкирский государственный университет, 2021. С. 66–69.
3. Жариков И. Н., Ишкина Ш. Х., Воронцов К. В. Статистические тесты однородности символьных последовательностей для информационного анализа электрокардиосигналов // Управление развитием крупномасштабных систем (MLSD'2016): Материалы IX Международной конференции (Москва, 3–5 октября 2016). М.: ИПУ РАН, 2016. Т. 2. С. 375–377.
4. Закирьянов И. И., Ишкина Ш. Х., Кунафин А. Ф. и др. Интерпретация результатов гидродинамических исследований скважин на неустановившихся режимах с применением методов машинного обучения // Нефтяное хозяйство. 2024. № 4. С. 54–59. doi:10.24887/0028-2448-2024-4-54-59
5. Закирьянов И. И., Ишкина Ш. Х., Сарапулова В. В., Давлетбаев А. Я. Интерпретация гидродинамических исследований скважин с применением методов машинного обучения в ПК «РН-ВЕГА» // Фундаментальная математика и ее приложения в естествознании: спутник Международной научной конференции «Уфимская осенняя математическая школа-2023»: Тезисы докладов XIV Международной школы-конференции студентов, аспирантов и молодых ученых, посвящённой 75-летнему юбилею профессоров Я. Т. Султанаева и М. Х. Харрасова (Уфа, 08–11 октября 2023 года). Уфа: ФГБОУ ВО «Уфимский университет науки и техноло-

гий», 2023. С. 151.

6. Закирьянов И. И., Ишкина Ш. Х.,Сарапулова В. В., Давлетбаев А. Я. Применение методов машинного обучения при интерпретации гидродинамических исследований скважин // Международная научно-практическая конференция «Цифровая трансформация в нефтегазовой отрасли» (Москва, 8–10 ноября 2023 года). Москва, 2023.
7. Ишкина Ш. Х. Аппроксимация комбинаторных оценок переобучения пороговых классификаторов // Интеллектуализация обработки информации ИОИ-2016: тезисы докладов 11-й международной конференции (Москва-Барселона, 10–14 октября 2016 года). Москва-Барселона: Общество с ограниченной ответственностью «ТОРУС ПРЕСС», 2016. С. 30–31.
8. Ишкина Ш. Х. Комбинаторные оценки переобучения одномерных пороговых классификаторов // Математические методы распознавания образов : тезисы докладов 17-й Всероссийской конференции с международным участием (Светлогорск, 19–25 сентября 2015 года). Москва: Общество с ограниченной ответственностью «ТОРУС ПРЕСС», 2015. С. 76–77.
9. Ишкина Ш. Х. Комбинаторные оценки переобучения пороговых решающих правил // Уфимск. матем. журн. 2018. Т. 10, №1. С. 50–65. doi:10.13108/2018-10-1-49
10. Ишкина Ш. Х. Суррогатное моделирование для вычисления оценок обобщающей способности пороговых решающих правил // Челябинский физико-математический журнал. 2025. Т. 10, №1. С. 53–69. doi:10.47475/2500-0101-2025-10-1-53-69
11. Ишкина Ш. Х., Воронцов К. В., Давлетбаев А. Я., Мирошниченко В. П. Применение комбинаторных оценок переобучения при планировании трассерных исследований в нефтегазовых месторождениях // Искусственный интеллект и принятие решений. 2024. № 1. С. 68–78. doi:10.14357/20718594240106

12. Ишкина Ш. Х., Закирьянов И. И., Сагдеев Э. И. и др. Апробация подхода по автоматической интерпретации эхограмм методами машинного обучения // Экспозиция Нефть Газ. 2024. № 5. С. 51–56. doi:10.24412/2076-6785-2024-5-51-56
13. Ишкина Ш. Х., Ивахненко А. А. Комбинаторные оценки переобучения пороговых решающих правил // Математические методы распознавания образов. 2013. Т. 16, № 1. С. 23.
14. Ишкина Ш. Х., Питюк Ю. А., Асалхузина Г. Ф. и др. Способ повышения информативности трассерных исследований в нефтегазовых месторождениях // Патент РФ № 2776786 C1, МПК E21B 47/11, опубл. 26.07.2022. Бюл.№ 21. Заявитель ООО «РН-Юганскнефтегаз».
15. Сагдеев Э. И., Ишкина Ш. Х., Давлетбаев А. Я. и др. Апробация подхода к восстановлению замеров дебита жидкости механизированных скважин с применением методов машинного обучения в программном комплексе «РН-ВЕГА» // Нефтяное хозяйство. 2024. № 4. С. 42–48. doi:10.24887/0028-2448-2024-4-42-48
16. Сахибгареев Э. Э., Ишкина Ш. Х. Автоматическая интерпретация гидродинамических исследований скважин на установившихся режимах добычи/закчки методами машинного обучения // Теоретические и экспериментальные исследования нелинейных процессов в конденсированных средах: материалы VII Межрегиональной школы-конференции студентов, аспирантов и молодых ученых, посвященной 60-летию первого полёта человека в космос (Уфа, 20–21 мая 2021 года). Уфа: Башкирский государственный университет, 2021. С.213–214.

Список литературы

17. Besnard E., Schmitz A., Boscher E., et al. Two-dimensional Aircraft High Lift System Design and Optimization // Proceedings of the 36th AIAA aerospace sciences meeting and exhibit, Reno, NV, USA. Reston: AIAA, 1998. doi:10.2514/6.1998-123.
18. Botov P. V. Exact estimates of the probability of overfitting for multidimensional modeling families of algorithms // Pattern Recogn. and Image Anal. 2010. V. 20, No. 4. P. 52–65.
19. Boucheron S., Bousquet O., Lugosi G. Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. 2005. V. 9. P. 323–375.
20. Bousquet O., Klochkov Y., Zhivotovskiy N. Sharper Bounds for Uniformly Stable Algorithms // Proceedings of Machine Learning Research. 2020. V. 125. P. 610–626.
21. Breiman L., Friedman J. H., Olshen R. A., Stone C. J. Classification And Regression Trees (1st ed.). Taylor & Francis, 1984. 368 p. doi:10.1201/9781315139470
22. Cover T. M., Thomas J. A. Elements of information theory. John Wiley & Sons, 2012. 784 p.
23. Dorogush A. V., Ershov V., Gulin A. CatBoost: gradient boosting with categorical features support. Workshop on ML Systems at NIPS 2017. doi:10.48550/arXiv.1810.11363
24. Dugstad Ø., Viig S., Krognes B., Kleven R., Huseby O. Tracer monitoring of enhanced oil recovery projects // EPJ Web of Conferences. 2013. V. 50, No. 02002. doi:10.1051/epjconf/20135002002
25. Forrester A., Sobester A., and Keane A. Engineering Design Via Surrogate Modelling: A Practical Guide. John Wiley & Sons, 2008. 240 p. doi:10.1002/9780470770801.
26. Frei A. I. Accurate estimates of the generalization ability for symmetric set of

predictors and randomized learning algorithms // Pattern Recogn. and Image Anal. 2010. V. 20, No. 3. P. 241–250.

27. Freund Y., Schapire R. E. A decision-theoretic generalization of on-line learning and an application to boosting // Journal of Computer and System Sciences. 1997. V. 55, No. 1. P. 119–139.
28. Fryer D., Strümke I., Nguyen H. Shapley Values for Feature Selection: The Good, the Bad, and the Axioms // IEEE Access, 2021. V. 9. P. 144352–144360. doi:10.1109/ACCESS.2021.3119110
29. GitHub Project https://github.com/shaurushka/decision-tree-with-ccv-and-eof (дата обращения 01.06.2023)
30. GiHub Project https://github.com/shaurushka/theshold-clfs-gen-bound (дата обращения 01.06.2023)
31. Haussler D., Littlestone N., Warmuth M. K. Predicting $\{0,1\}$-functions on randomly drawn points // Information and Computation. 1994. V. 115, No. 2. P. 248–292.
32. Hoeffding W. Probability inequalities for sums of bounded random variables // Journal of the American Statistical Association. 1963. No. 58. P. 13–30.
33. Holm S. A simple sequentially rejective multiple test procedure // Scandinavian Journal of Statistics. 1979. V. 6, No. 2. P. 65–70.
34. Joshi D., Patidar A. K., Mishra A., et al. Prediction of sonic log and correlation of lithology by comparing geophysical well log data using machine learning principles // GeoJournal. 2021. doi:10.1007/S10708-021-10502-6
35. Kearns M. J., Mansour Y., Ng A. Y., Ron D. An experimental and theoretical comparison of model selection methods // Computational Learning Theory. 1995. P. 21–30.
36. Khilrani N., Prajapati P., Patidar A. K. Contrasting machine learning regression algorithms used for the estimation of permeability from well log data // Arab. J. Geosci. 2021. V. 14, No. 20. P. 1–14. doi:10.1007/S12517-021-08390-8
37. Knackstedt M. A, Latham S., Madadi M., et al. Digital rock physics: 3D imaging

of core material and correlations to acoustic and flow properties // The Lead Edge. 2009. V. 28, No. 1. P. 28–33. doi:10.1190/1.3064143

38. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // Proc. Int. Joint Conf. on Artificial Intelligence. 1995. P. 1137–1143.
39. Koltchinskii V. Rademacher Penalties and Structural Risk Minimization // IEEE Trans. Inf. Theory. 2001. Vol. 47, No. 5. P. 1902–1914.
40. Koltchinskii V. Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics. Springer, 2011.
41. Kuhn M., Johnson K. Classification Trees and Rule-Based Models // Applied Predictive Modeling. NY:Springer, 2013. doi:10.1007/978-1-4614-6849-3_14
42. Langford J. Quantitatively Tight Sample Complexity Bounds: Ph.D. thesis // Carnegie Mellon Thesis, 2002.130 p.
43. Maimon O., Rokach L. Data Mining and Knowledge Discovery Handbook, 2nd ed. Springer, 2010. 1285 p. doi:10.1007/978-0-387-09823-4
44. Shook G. M., Ansley Sh. L., Wylie A. Tracers and tracer testing: design, implementation, tracer selection, and interpretation methods, report, January 1, 2004. Idaho Falls, Idaho: INL, 2004. 36 p. doi:10.2172/910642
45. Mitchell T. Machine Learning. McGraw Hill, 1997. 414 p.
46. Mogilicharla A., Mittal P., Majumbar S., Mitra K. Kriging surrogate based multi-objective optimization of bulk vinyl acetate polymerization with branching // Materials and Manufacturing Processes. 2015. No. 30. P. 394–402.
47. Nelder J.A., Mead R. A simplex method for function minimization // Computer Journal. 1965. V. 7. P. 308–313.
48. Pedregosa F., et al. Scikit-learn: Machine Learning in Python // JMLR. 2011. V. 12, No. 85. P. 2825–2830.
49. Patidar A. K., Joshi D., Dristant U. et al. A review of tracer testing techniques in porous media specially attributed to the oil and gas industry // J. Petrol. Explor.

Prod. Technol. 2022. V. 12. P. 3339–3356. doi:10.1007/s13202-022-01526-w

50. Rutherford A. Anova and ANCOVA: a GLM approach. John Wiley & Sons, 2011. 360 p.
51. Shalev-Shwartz S., Ben-David S. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014. 449 p.
52. Simpson T., Toropov V., Balabanov V., Viana F. Design and analysis of computer experiments in multidisciplinary design optimization: a review of how far we have come or not // Proceedings of the 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference (Victoria, British Columbia, Canada, 10–12 September 2008). doi:10.2514/6.2008-5802
53. Sprunger C., Muther T., Syed F. I., et al. State of the art progress in hydraulic fracture modeling using AI/ML techniques // Model Earth Syst. Environ. 2022. V. 8. P. 1–13. doi:10.1007/S40808-021-01111-W
54. Valle-Pérez G., Louis A.A. Generalization bounds for deep learning. 2020. doi:10.48550/arXiv.2012.04115.
55. Vorontsov K. V. Combinatorial probability and the tightness of generalization bounds // Patt. Rec. and Image An. 2008. V. 18, No. 2. P. 243–259.
56. Vorontsov K. V. Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // Pattern Recogn. and Image Anal. 2009. V. 19, No. 3. P. 412–420.
57. Vorontsov K. V. Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recogn. and Image Anal. 2010. V. 20, No. 3. P. 269–285. doi:10.1134/S105466181003003X
58. Vorontsov K. V. Combinatorial Theory of Overfitting: How Connectivity and Splitting Reduces the Local Complexity // 9th IFIP WG 12.5 Int. Conf., AIAI (Paphos, Cyprus, September 30–October 2, 2013). Springer Berlin, Heidelberg, 2013.
59. Vorontsov K. V., Ivahnenko A. A. Tight combinatorial generalization bounds for threshold conjunction rules // 4th Int. Conf. on Pattern Recognition

and Machine Intelligence, 2011. Lecture Notes in Computer Science. Springer–Verlag, 2011. P. 66–73.

60. åberg G. The use of natural strontium isotopes as tracers in environmental studies // Water Air Soil Pollut. 1995. V. 79, No. 1. P. 309–322. doi:10.1007/BF01100444
61. Çetinkaya Z., Horasan F. Decision Trees in Large Data Sets // International Journal of Engineering Research and Development, 2021. V. 13, No. 1. P. 140–151. doi:10.29137/umagd.763490
62. Ботов П. В. Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Математические методы распознавания образов-14. М.: МАКС Пресс, 2009. С. 7–10.
63. Бурнаев Е., Ерофеев П., Зайцев А. и др. Суррогатное моделирование и оптимизация профиля крыла самолета на основе гауссовских процессов. URL: http://itas2012.iitp.ru/pdf/1569602325.pdf (дата обращения 14.07.2024)
64. Бухмастова С. В., Фахреева Р. Р., Питюк Ю. А., Давлетбаев А. Я., и др. Апробация методов MLR и CRMIP при исследовании взаимовлияния скважин // Нефтяное хозяйство, 2020. № 8. С. 58–62. doi:10.24887/0028-2448-2020-8-58-62
65. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979. 448 с.
66. Вапник В. Н., Червоненкис А. Я. О равномерной сходимости частот появления событий к их вероятностям // Теория вероятности и ее применения, 1971. Т. 16, № 2. С. 264–280.
67. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974. 416 с.
68. Воронцов К. В. Комбинаторные оценки качества обучения по прецедентам // Доклады РАН. 2004. Т. 394, № 2. С. 175–178.
69. Воронцов К. В. Точные оценки вероятности переобучения // До-

кл. РАН. 2009. Т. 429, № 1. С. 15–18.

70. Воронцов К. В., Фрей А. И., Соколов Е. А. Вычислимые комбинаторные оценки вероятности переобучения // Машинное обучение и анализ данных. 2013. Т. 1, № 6. С. 734–743.
71. Гарифуллин М., Барабаш А., Наумова Е., и др. Суррогатное моделирование для определения начальной жесткости вращения сварных трубчатых соединений // Инженерно-строительный журнал. 2016. Т. 3, № 63. С. 53–76. doi:10.5862/MCE.63.4.
72. Гуз И. С. Конструктивные оценки полного скользящего контроля для пороговой классификации // Математическая биология и биоинформатика, 2011. Т. 6, № 2. С. 173–189. doi:10.17537/2011.6.173.
73. Животовский Н. К., Воронцов К. В. Критерии точности комбинаторных оценок обобщающей способности // Интеллектуализация обработки информации (ИОИ-2012). М.: Торус Пресс, 2012. С. 25–28.
74. Журавлёв Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. М.: Фазис, 2006. 176 с.
75. Журавлёв Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики: Вып. 33. 1978. С. 5–68.
76. Лагутин М. Б. Наглядная математическая статистика: учебное пособие. 2-е изд., испр. М.: БИНОМ, Лаборатория знаний, 2009. 472 с.
77. Мирзаянов А. А., Асалхузина Г. Ф., Питюк Ю. А. и др. Матрицы применимости трассерных исследований на примере элемента девятиточечной системы разработки с трещинами гидроразрыва // Нефтегазовое дело. 2021. Т. 19, № 4 С. 41–49. doi: 10.17122/ngdelo-2021-4-41-49
78. Соколовский Э. В., Соловьев Г. Б., Тренчиков Ю. И. Индикаторные методы изучения нефтегазоносных пластов. М.: Недра, 1986. 157 с.
79. Соколовский Э. В., Чижов С. И., Тренчиков Ю. И. и др. Методическое руководство по технологии проведения индикаторных исследований и интер-

претации их результатов для регулирования и контроля процесса заводнения нефтяных залежей. РД 39-014-7428-235-89. Грозный: СевКавНИПИнефть, 1989. 79 с.

80. Толстихин И. О. Вероятность переобучения некоторых разреженных семейств алгоритмов // Междунар. конф. ИОИ-8. М.: МАКС Пресс, 2010. С. 83–86.
81. Трофимов А. С., Леонов В. А., Алпатов А. А. Способ исследования и разработки многопластового месторождения углеводородов // Патент РФ № 2315863 C2, МПК E21B 47/10, E21B 43/00, опубл. 27.01.2008. Бюл. № 3. Заявитель ООО Научно-исследовательский Институт «СибГеоТех».
82. Фахреева Р. Р., Питюк Ю. А., Асалхузина Г. Ф. Давлетбаев А. Я. и др. Развитие метода многопараметрической линейной регрессии для анализа трассерных исследований // Вестник Башкирского университета. 2021. С. 554–558. doi:10.33184/bulletin-bsu-2021.3.2.
83. Фрей А. И., Толстихин И. О. Комбинаторные оценки вероятности переобучения на основе кластеризации и покрытий множества алгоритмов // Машинное обучение и анализ данных. 2013. Т. 1, № 6. С. 761–778.
84. Фрей А. И., Толстихин И. О. Комбинаторные оценки вероятности переобучения на основе покрытий множества алгоритмов // Доклады РАН, 2014. Т. 455, № 3. С. 265–268.
85. Юдин Е. В., Андрианова А. М., Ганеев Т. А. и др. Контроль дебита жидкости нестабильно работающего фонда скважин при помощи виртуального расходомера // Нефтяное хозяйство. 2023. № 8. С. 82–87. doi:10.24887/0028-2448-2023-8-82-87

Приложение А

Акт внедрения

УТВЕРЖДАЮ
Начальник управления по
моделированию и анализу
исследований скважин и пластов
ООО «РН-БашНИПИнефть»

_____ Давлетбаев А.Я.
«11» декабря 2025 г.

АКТ
внедрения основных результатов
диссертационной работы Ишкиной Ш.Х.

Мы, представители ООО «РН-БашНИПИнефть», настоящим актом подтверждаем, что ряд результатов, полученных в диссертационной работе Ишкиной Шауры Хабировны на тему «Комбинаторные оценки переобучения пороговых решающих правил» на соискание ученой степени кандидата физико-математических наук, а именно:

1) алгоритм построения программы трассерных (маркерных) исследований с применением деревьев решений;

2) алгоритм построения дерева решений с использованием комбинаторных оценок обобщающей способности (полного скользящего контроля и ожидаемой переобученности) одномерных пороговых решающих правил в качестве критерия выбора атрибута для разделения узла

используются при выполнении научно-исследовательских работ и инженерно-аналитических работ, а также учтены при разработке методических указаний по планированию, проведению и интерпретации трассерных (маркерных) исследований на месторождениях Западной Сибири.

Начальник отдела гидродинамических
исследований скважин
ООО «РН-БашНИПИнефть» Абдуллин Р.И.

Эксперт отдела
гидродинамического моделирования
ООО «РН-БашНИПИнефть» Штинов В.А.