На правах рукописи

Ишкина Шаура Хабировна

Комбинаторные оценки переобучения пороговых решающих правил

1.2.1 – Искусственный интеллект и машинное обучение

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2026

Работа выполнена в *Федеральном исследовательском центре «Информатика и управление» Российской академии наук.*

Научный руководитель: **Воронцов Константин Вячеславович**
доктор физико-математических наук, профессор РАН, Московский государственный университет им. М. В. Ломоносова, заведующий кафедрой

Официальные оппоненты: **Двоенко Сергей Данилович**
доктор физико-математических наук, профессор, Федеральное государственное бюджетное образовательное учреждение высшего образования «Тульский государственный университет», профессор

Масляков Глеб Олегович
кандидат физико-математических наук, ООО «Яндекс.Технологии», старший разработчик программного обеспечения

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук

Защита состоится «____» _____ 2026 г. в ____ часов на заседании диссертационного совета *24.1.224.03* при *Федеральном исследовательском центре «Информатика и управление» Российской академии наук* по адресу: *119333, Россия, Москва, ул. Вавилова, д. 42.*

С диссертацией можно ознакомиться в библиотеке *Федерального исследовательского центра «Информатика и управление» Российской академии наук* и на сайте http://www.frccsc.ru.

Автореферат разослан «____» _____ 2026 г.

Ученый секретарь
диссертационного совета
24.1.224.03,
кандидат технических наук *Рейер Иван Александрович*

Общая характеристика работы

Актуальность темы исследования. Диссертация посвящена построению точных (достигаемых) 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 выделяется как одна из основных задач теории статистического обучения[1]. Завышенность полученных оценок может приводить к неоптимальному выбору структурных параметров[2]. 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. Вапник и Червоненкис получили равномерные по семействам классификаторов оценки[3], 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[4] 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, явля-

[1] 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.

[2] 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.

[3] Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974. 416 с.

[4] Vorontsov K. V. Combinatorial probability and the tightness of generalization bounds // Patt. Rec. and Image An. 2008. V. 18, No. 2. P. 243–259.

ется оценкой худшего случая.

Теория статистического обучения продолжает активно развиваться, последователи теории занимаются повышением точности равномерных оценок с учетом особенностей данных и конкретных алгоритмов классификации[5]. Получены более тонкие оценки, которые зависят от свойств отношения частичного порядка на множестве вектор-столбцов матрицы ошибок. Среди плодотворных подходов можно выделить оценки, адаптирующиеся к данным и использующие понятие Радемахеровской сложности, предложенной в 1999 г. В. Колчинским.

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

В комбинаторной теории переобучения[6], предложенной К. В. Воронцовым, вероятностью переобучения называют долю разбиений конечного множества объектов на обучающую и контрольную выборки фиксированной длины, при которых произошло переобучение.

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 модельных 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, хэмминговых шаров и некоторых их разреженных подмножеств. Разработан теоретико-групповой подход, который позволяет получать достигаемые верхние оценки для семейств с произвольными симметриями.

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

В комбинаторной теории для вероятности переобучения получена оценка расслоения–связности[7], учитывающая особенности способа построения классификатора по обучающей выборке, а также локальные свойства семейства – эффекты расслоения и связности. 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.

---

[5] Valle-Pérez G., Louis A.A. Generalization bounds for deep learning. 2020. doi:10.48550/arXiv.2012.04115.

[6] Воронцов К. В. Комбинаторные оценки качества обучения по прецедентам // Доклады РАН. 2004. Т. 394, № 2. С. 175–178.

[7] 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.

В работе[8] получены условия, при которых оценка расслоения–связности является точной. Им удовлетворяют, в частности, монотонные и унимодальные цепи классификаторов. В практических задачах статистического обучения такие цепи могут порождаться элементарными пороговыми правилами, используемых в таких алгоритмах классификации, как решающие деревья, логические закономерности, алгоритмы вычисления оценок, 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 И. С. Гуза для ожидаемой частоты ошибок на контрольной выборке[9] в частном случае, когда признак принимает попарно различные значения на объектах. Различные уточнения оценок расслоения–связности, например, оценка[10] Е. А. Соколова, учитывающая попарную конкуренцию между классификаторами, также остаются завышенными для прямых последовательностей. Однако завышенность верхних оценок остается неизученной.

Цель диссертационной работы. Построение точных (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.

---

[8] Животовский Н. К., Воронцов К. В. Критерии точности комбинаторных оценок обобщающей способности // Интеллектуализация обработки информации (ИОИ-2012). М.: Торус Пресс, 2012. С. 25–28.

[9] 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.

[10] 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.

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.

Методы исследования. Для построения алгоритма использованы методы комбинаторики и динамического программирования. Для оценки вычислительной сложности использованы методы математического анализа. Для проведения вычислительных экспериментов по сравнению полученных достигаемых верхних оценок с существующими оценками алгоритм реализован на языке программирования C++. Для апробации на результатах проведения трассерных исследований алгоритм дерева решений с модифицированным критерием выбора атрибута для разделения узла реализован на языке программирования C++. Для построения суррогатной модели и вычислительных экспериментов по оценке точности и устойчивости модели алгоритм реализован на языке Python. Для выводов о статистической значимости результатов апробации использованы методы математической статистики.

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.

Основные результаты диссертации докладывались на следующих конференциях:

1. Международная школа-конференция «Фундаментальная математика и ее приложения в естествознании», 2023. [10]

2. Международная научно-практическая конференция «Цифровая трансформация в нефтегазовой отрасли», 2023. [11]

3. Межрегиональная школа-конференция «Теоретические и экспериментальные исследования нелинейных процессов в конденсированных средах», 2021. [15]

4. Всероссийская молодежная научно-практическая конференция «Геолого-геофизические исследования нефтегазовых пластов», 2021.[8]

5. Международная конференция «Управление развитием крупномасштабных систем», 2016. [9]

6. Международная конференция «Intelligent Data Processing», 2016. [12]

7. Всероссийская конференция «Математические методы распознавания образов», 2015. [13]

8. Всероссийская конференция «Математические методы распознавания образов», 2013. [14]

Публикации. Результаты диссертации содержатся в 16 публикациях. В изданиях из списка ВАК представлено 7 публикаций [1–7]. Получен 1 патент на изобретение [16]. Все работы индексируются РИНЦ. Работы [1–5] 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 вычислении достигаемой верхней оценки ожидаемой переобученности и верхней оценки частоты ошибок на контрольной выборке для семейства одномерных пороговых решающих правил, проведены вычислительные эксперименты;
- в работах [2, 10, 11] разработан алгоритм интерпретации исследований скважин на неустановившихся режимах с применением методов машинного обучения, проведено тестирование алгоритма;
- в работе [5] разработан алгоритм «виртуального расходомера» на основе стекинга моделей машинного обучения, проведены вычислительные эксперименты;
- в работе [6] 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 [7] разработан алгоритм интерпретации исследований скважин методом эхометрирования с применением методов машинного обучения;
- в работе [8] разработан алгоритм на основе методов машинного обучения для анализа взаимовлияния скважин;
- в работе [9] разработан подход для проверки однородности символьных последовательностей на основе проверки статистических гипотез;
- в работе [14] разработан алгоритм вычисления оценки вероятности переобучения для прямой цепи;

- в работе [15] разработан алгоритм на основе случайного леса для выделения установившихся режимов на динамических данных при гидродинамическом исследовании скважины методом построения индикаторной диаграммы;

- в работе [16] разработан алгоритм построения программы трассерных исследований с использованием методов 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.

Содержание работы

В автореферате нумерация основных утверждений (определений, лемм, теорем) и формул сквозная.

Во введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана практическая значимость полученных результатов, представлены выносимые на защиту научные положения.

В первой главе решается задача вычисления верхних оценок переобучения одномерных пороговых решающих правил при выборе порога.

В **разделе 1.1** формулируется математическая модель задачи классификации как задачи принятия решений в условиях неполноты информации и вводятся основные определения комбинаторной теории переобучения. Предполагается, что дана бинарная матрица I, строки которой соответствуют объектам из генерального множества $\mathbb{X}$, столбцы — классификаторам из семейства $\mathbb{A}$. Мощность генерального множества конечна и равна L. В ячейке матрицы $I(a,x)$ находится единица тогда и только тогда, когда данный классификатор a ошибается на данном объекте x. Из множества $\mathbb{X}$ всех строк матрицы случайно и равновероятно (с вероятностью $P = \frac{1}{C_L^\ell}$) выбирается наблюдаемая обучающая выборка — подмножество $X \subset \mathbb{X}$ фиксированной мощности ℓ. Дополнение

$\bar{X} = \mathbb{X}\backslash X$ называется контрольной выборкой. Затем на основе метода обучения μ из множества $\mathbb{A}$ всех столбцов матрицы по обучающей выборке X выбирается классификатор.

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.

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.

Для оценок обобщающей способности метода обучения в комбинаторной теории переобучения рассматриваются следующие функционалы:

- вероятности переобучения:

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

- полного скользящего контроля, 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 переобученности:

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

В данной работе рассматривается метод обучения, называемый методом минимизации эмпирического риска (МЭР), который выбирает классификатор с минимальной частотой ошибок на X:

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

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 a с наибольшим числом ошибок на множестве $\mathbb{X}$.

Для получения верхних оценок EOF рассматривается 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: при $\ell = \frac{L}{2}$ рademахеровская сложность семейства равна ожидаемой переобученности метода МП.

В **разделе 1.2** вводится понятие прямой последовательности классификаторов. Рассмотрим множества объектов, по которым различаются соседние классификаторы семейства $\mathbb{A} = \{a_0, \ldots, a_P\}$:

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

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

Определение 1.2. *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.*

Доказывается теорема, что между семействами прямых последовательностей и одномерными пороговыми классификаторами имеется биекция, откуда следует, что данные понятия являются синонимами.

В случае, когда числовой признак принимает попарно различные значения на объектах множества $\mathbb{X}$, семейство называется прямой цепью.

Проводятся вычислительные эксперименты, которые показывают, что для семейства пороговых решающих правил актуальна задача определения обобщающей способности. Эффективное вычисление Q_ε, CCV и EOF непосредственно по определению возможно только при малых ℓ. Если ℓ близко к $L/2$, то число слагаемых экспоненциально по L.

Вследствие этого в **разделе 1.3** ставится следующая задача. Для прямой последовательности $\mathbb{A}$ общего вида, методов обучения ПМЭР и МП вычислить точные (достигаемые) верхние оценки вероятности переобучения Q_ε, полного скользящего контроля CCV и ожидаемой переобученности EOF за полиномиальное по L время.

В **разделе 1.4** исследуются свойства МЭР, МП и ПМЭР и доказывается, что они обладают общим свойством, которое определяется как финитность метода обучения. Как показано далее в **разделе 1.5**, данное свойство позволяет получить аналитическое представление достигаемых верхних оценок обобщающей способности произвольного семейства классификаторов.

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 дано произвольное семейство классификаторов $\mathbb{A}$. Пусть на множестве $\mathbb{A} \times \mathbb{A} \times [\mathbb{X}]^\ell$ имеется бинарное отношение $a \succ_X a'$. 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 $\mathbb{X}'$.

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:

$$a = \mu X \quad \Leftrightarrow \quad a \succ_X a', \quad \forall a' \neq a. \tag{1.2}$$

Доказывается теорема о том, что методы МЭР, МП и ПМЭР являются финитными.

Обозначим через $\mathbb{D}$ подмножество объектов, по которым классификаторы семейства $\mathbb{A} = \{a_0, \ldots, a_P\}$ различимы:

$$\mathbb{D} = G_0 \cup \cdots \cup G_{P-1} = \big\{x \in \mathbb{X} \,\big|\, \exists\, a, a' \in \mathbb{A}\colon I(a,x) \neq I(a',x)\big\}, \tag{1.3}$$

где множества G_p определяются согласно (1). Объекты множества $\mathbb{N} = \mathbb{X} \setminus \mathbb{D}$ назовем *нейтральными*. На множестве $\mathbb{N}$ классификаторы семейства неразличимы и допускают одинаковое число ошибок:

$$\begin{aligned} &m = n(a, \mathbb{N}), \quad \forall a \in \mathbb{A}; \\ &m_p = n(a_p, \mathbb{D}). \end{aligned} \tag{1.4}$$

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 1.1. *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*

$$Q_\varepsilon = \frac{1}{C_L^\ell} \sum_{p=0}^{P} \sum_{(t,e)\in\Psi_p} D_p(t,e) N_p(t,e), \tag{1.5}$$

где множество $\mathbb{D}$, параметры m_p и m определяются по (1.3) и (1.4)

$$\Psi_p = \big\{(t,e) \,|\, 0 \leqslant t \leqslant \min\{\ell, |\mathbb{D}|\},\ 0 \leqslant e \leqslant \min\{t, m_p\}\big\}; \tag{1.6}$$

$$N_p(t,e) = C_{L-|\mathbb{D}|}^{\ell-t}\, H_{L-|\mathbb{D}|}^{\ell-t,m}(s_p(e)); \tag{1.7}$$

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

Для функционалов полного скользящего контроля и ожидаемой переобученности имеют место аналогичные теоремы.

Теорема 1.2. *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*

$$\mathsf{CCV} = \frac{1}{(L-\ell)C_L^\ell} \sum_{p=0}^{P} \sum_{(t,e)\in\Psi_p} D_p(t,e) F_p(t,e), \tag{1.8}$$

где

$$F_p(t,e) = \sum_{s=0}^{\min\{\ell-t,m\}} C_m^s C_{L-|\mathbb{D}|-m}^{\ell-t-s}\big(n(a_p,\mathbb{X}) - s - e\big), \tag{1.9}$$

множества $\mathbb{D}$ и Ψ_p определяются по (1.3) и (1.6), параметры m_p и m определяются по (1.4).

Теорема 1.3. *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 функционал ожидаемой переобученности имеет вид*

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

где множества $\mathbb{D}$ *и* Ψ_p *определяются по* (1.3) *и* (1.6), *параметры* m_p *и* m *определяются по* (1.4) *и*

$$K_p(t,e) = \sum_{s=0}^{\min\{\ell-t,m\}} C_m^s C_{L-|\mathbb{D}|-m}^{\ell-t-s} \Big(\tfrac{1}{L-\ell}\big(n(a_p,\mathbb{X}) - (s+e)\big) - \tfrac{1}{\ell}(s+e) \Big). \quad (1.11)$$

Таким образом, задача сводится к вычислению для каждого p количества разбиений $D_p(t,e)$ на всем множестве Ψ_p.

В **разделе 1.6** рассматривается случай прямой последовательности и описывается алгоритм вычисления $D_p(t,e)$. Далее элементы множества $\mathbb{D}$ называются ребрами последовательности. Доказывается теорема, которая для каждого p сводит задачу к расчету количества разбиений множества ребер левой $a_0,\ldots,a_p$ и правой $a_p,\ldots,a_P$ последовательностей относительно классификатора a_p.

Теорема 1.4. *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*

$$D_p(t,e) = \sum_{t'+t''=t} \sum_{e'+e''=e} L_p(t',e') R_p(t'',e''), \quad (1.12)$$

где

$$L_p(t',e') = \#\left\{ (X \cap \mathbb{L}_p, \bar{X} \cap \mathbb{L}_p) \;\middle|\; \begin{array}{l} \forall d = 0,\ldots,p,\;\; a_p \succ_X a_d, \\ t' = |X \cap \mathbb{L}_p|,\;\; e' = n(a_p,\, X \cap \mathbb{L}_p) \end{array} \right\}, \quad (1.13)$$

$$R_p(t'',e'') = \#\left\{ (X \cap \mathbb{R}_p, \bar{X} \cap \mathbb{R}_p) \;\middle|\; \begin{array}{l} \forall d = p+1,\ldots,P,\;\; a_p \succ_X a_d, \\ t'' = |X \cap \mathbb{R}_p|,\;\; e'' = n(a_p,\, X \cap \mathbb{R}_p) \end{array} \right\}, \quad (1.14)$$

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*

$$\Psi'_p = \big\{ (t',e') \;\big|\; 0 \leqslant t' \leqslant \min\{\ell, |\mathbb{L}_p|\}, 0 \leqslant e' \leqslant \min\{t',\, n(a_p,\mathbb{L}_p)\} \big\}, \quad (1.15)$$

$$\Psi''_p = \big\{ (t'',e'') \;\big|\; 0 \leqslant t'' \leqslant \min\{\ell, |\mathbb{R}_p|\}, 0 \leqslant e'' \leqslant \min\{t'',\, n(a_p,\mathbb{R}_p)\} \big\}. \quad (1.16)$$

Для методов ПМЭР и МП предлагается алгоритм для вычисления значений $L_p(t',e')$ и $R_p(t'',e'')$ для всех возможных значений параметров, основанный на рекуррентном подсчете числа траекторий при блуждании по трехмерной сетке между двумя заданными точками с ограничениями специального вида. Доказывается ряд теорем для обоснования корректности алгоритма.

В **разделе 1.7** приводится псевдокод алгоритма вычисления оценок обобщающей способности для прямой последовательности и доказывается его полиномиальная вычислительная сложность.

Результаты первой главы опубликованы в работах [1] и [3].

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 завышенными на несколько порядков. Показано, что точными оказываются оценки Гуза для величины полного скользящего контроля, откуда следует вывод о применимости данных оценок в прикладных задачах, однако отмечается, что они накладывают требования на распределение значений одномерного признака, то есть применимы только в частных случаях.

Результаты второй главы опубликованы в работе [1].

Третья глава посвящена теме трассерных (маркерных) исследований в нефтегазовых месторождениях. В процессе проведения трассерного исследования осуществляется закачка меченой жидкости (трассера) в одну или в несколько нагнетательных скважин. После закачки начинают производить отбор проб жидкости из устья добывающих скважин, которые далее анализируются в лаборатории. В случае обнаружения меченой жидкости в пробах говорят о наличии гидродинамической связи между скважинами и проводят расчеты для оценки гидродинамических свойств пласта в межскважинном пространстве, направления и скорости распространения жидкости в пласте, что важно для решения задач проектирования и мониторинга разработки месторождений.

В **разделе 3.1** описываются принятые критерии выбора скважин при планировании трассерных исследований и отмечаются их недостатки. При построении программы не учитываются 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 трассер, оказывается избыточным и приводит к более значительным затратам на проведение исследования. Для решения проблемы предлагается способ построения программы исследований с применением методов машинного обучения. Согласно способу, пара скважин нагнетательная–добывающая включается в программу на основе ответа классификатора. Признаками для описания пары скважин являются коэффициенты взаимовлияния по методам емкостно-резистивной модели и многопараметрической регрессии, рассчитываемые на основе динамических данных эксплуатации скважин. Апробация подхода на промысловых данных показала, что использование алгоритма машинного обучения позволяет уточнить программу трассерных исследований и повысить долю пар скважин с наличием гидродинамической связи.

Классификатор в предложенном способе основан на алгоритме решающего дерева для получения интерпретируемых результатов, а также в связи с

ограниченным объемом накопленной цифровой базы с отчетами по результатам трассерных исследований, используемой для обучения модели. В **разделе 3.2** обсуждаются недостатки известных алгоритмов генерации дерева решений, связанные с явлением переобучения. Исследуются причины возникновения переобучения, одна из которых заключается в смещенности критериев выбора атрибута при построении разбиения в узле. В **разделе 3.3** ставится задача повышения обобщающей способности дерева решений путем модификации критерия выбора атрибута. Рассматривается бинарное дерево, в котором атрибутом является одномерный пороговый классификатор $[F \leq \theta]$, разделяющий множество примеров, попавших в узел, на два множества. Порог θ определяется по методу МЭР. Выбор атрибута сводится к выбору признака F на основе подвыборки примеров обучающей выборки, попавших в узел.

В **разделах 3.5 и 3.6** описывается предлагаемый критерий для выбора атрибута в узле и приводится псевдокод алгоритма построения дерева решений. Модификация состоит в применении несмещенных достигаемых верхних оценок переобучения пороговых решающих правил: ожидаемой переобученности EOF и ожидаемой частоты ошибок на отложенной выборке CCV для метода ПМЭР. Выбирается признак F, для которого значения критерия оптимальны. Для вычисления критерия применяется алгоритм, разработанный в первой главе. В **разделе 3.7** проводятся вычислительные эксперименты на промысловых данных результатов трассерных исследований на примере двух месторождений Западной Сибири. Результаты показывают, что применение комбинаторных оценок в качестве критериев выбора атрибута в узле приводит к статистически значимому уменьшению переобученности и повышению точности дерева решений. Таким образом, предложенный подход позволяет повысить эффективность проведения трассерных исследований и снизить затраты на промысловые работы.

Результаты третьей главы опубликованы в работах [6] и [16].

В **четвертой главе** решается задача построения суррогатной модели для быстрого вычисления приближенных оценок переобучения семейства пороговых решающих правил. Описывается процесс сбора обучающей выборки для модели, которая состоит из пар «объект, ответ», и каждым объектом является семейство пороговых решающих правил, ответом – 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. По результатам тестирования выбрана модель нейронной сети с MAPE=2.8%. Анализ значимости признаков показывает, что при построении оценок переобучения недостаточно учитывать только количество классификаторов и минимальное число ошибок классификаторов, необходимо использовать геометрическую структуру семейства (расслоение по числу ошибок) и взаимосвязь между классификаторами (связность). Показано, что использование модели позволяет сократить время вычисления оценок обобщающей способности 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 [4].

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 на практике в частных случаях.

5. Проведен анализ завышенности оценки частоты ошибок на контрольной выборке на основе Радемахеровской сложности по сравнению с достигаемыми 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 итогового классификатора.

7. Построена суррогатная модель для вычисления приближенных оценок обобщающей способности семейства пороговых решающих правил с высокой точностью. Показано, что использование суррогатного моделирования позволяет сократить сложность вычисления оценок переобучения с $O(L^5)$ до $O(L^2)$ и может применяться в практических задачах для отбора признаков при построении моделей машинного обучения.

Благодарность. Автор выражает признательность 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.

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. 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
3. 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
4. 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
5. 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
6. 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
7. 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
8. 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.
9. 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.
10. Закирьянов И. И., Ишкина Ш. Х., Сарапулова В. В., Давлетбаев А. Я. Ин-

терпретация 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 технологий», 2023. С. 151.

11. 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.
12. 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.
13. 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.
14. Ишкина Ш. Х., Ивахненко А. А. Комбинаторные оценки переобучения пороговых решающих правил // Математические методы распознавания образов. 2013. Т. 16, № 1. С. 23.
15. 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.
16. 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».

Научное издание

Ишкина Шаура Хабировна

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук на тему:
Комбинаторные оценки переобучения пороговых решающих правил