На правах рукописи

ВОРОНЦОВ КОНСТАНТИН ВЯЧЕСЛАВОВИЧ

КОМБИНАТОРНАЯ ТЕОРИЯ НАДЁЖНОСТИ ОБУЧЕНИЯ ПО ПРЕЦЕДЕНТАМ

05.13.17 — теоретические основы информатики

Автореферат диссертации на соискание ученой степени
доктора физико-математических наук

Москва, 2010

Работа выполнена в Учреждении Российской академии наук
Вычислительный центр им. А. А. Дородницына РАН

Научный консультант:	доктор физико-математических наук, чл.-корр. РАН **Константин Владимирович Рудаков**
Официальные оппоненты:	доктор физико-математических наук, академик РАН, **Виктор Леонидович Матросов**
	доктор физико-математических наук, профессор, **Алексей Иванович Чуличков**
	доктор физико-математических наук, профессор, **Владислав Викторович Сергеев**
Ведущая организация:	Московский физико-технический институт (государственный университет)

Защита диссертации состоится « ____ » _____ 2010 г. в _____ на заседании диссертационного совета Д 002.017.02 в Учреждении Российской академии наук Вычислительный центр им. А. А. Дородницына РАН по адресу: 119333, г. Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан « ____ » _____ 2010 г.

Учёный секретарь диссертационного совета
Д 002.017.02, д.ф.-м.н., профессор В. В. Рязанов

Общая характеристика работы

Диссертационная работа посвящена проблемам обобщающей способности в задачах обучения по прецедентам. Предлагается комбинаторный подход, позволяющий получать точные оценки вероятности переобучения, учитывающие эффекты расслоения и связности в семействах алгоритмов.

Актуальность темы. Вопрос о качестве восстановления зависимостей по эмпирическим данным является фундаментальной проблемой *теории статистического обучения* (statistical learning theory, SLT).

Основным объектом исследования в SLT является задача обучения по прецедентам: задана *обучающая выборка* пар «объект–ответ»; требуется восстановить функциональную зависимость ответов от объектов, т. е. построить алгоритм, способный выдавать адекватный ответ для произвольного объекта. К этому классу задач относятся задачи распознавания образов, классификации, восстановления регрессии, прогнозирования.

Основной задачей SLT является получение оценок вероятности ошибки построенного алгоритма на объектах, не входивших в обучающую выборку. Эта задача нетривиальна, поскольку частота ошибок на обучающей выборке, как правило, является смещённой (сильно заниженной) оценкой вероятности ошибки. Это явление называют *переобучением*. Способность алгоритмов восстанавливать неизвестную зависимость по конечной выборке данных называют *обобщающей способностью*.

Возникновение SLT связывают с появлением статистической теории Вапника–Червоненкиса (далее VC-теории) в начале 70-х годов. Основным результатом VC-теории являются верхние оценки вероятности ошибки, зависящие от длины обучающей

выборки и сложности семейства функций, из которого выбирается искомый алгоритм. Согласно VC-теории, для получения надёжных алгоритмов необходимо ограничивать сложность семейства. Естественной мерой сложности конечного семейства является его мощность. Однако на практике гораздо чаще используются бесконечные семейства. Чтобы свести этот случай к конечному, вводится бинарная функция потерь. Тогда лишь конечное число алгоритмов оказываются попарно различимыми на выборке конечной длины. Зависимость этого числа от длины выборки называется *функцией роста семейства*. В худшем случае она растёт экспоненциально, но если её рост ограничен сверху полиномом фиксированной степени, то оценки являются состоятельными — частота ошибок на обучающей выборке стремится к вероятности ошибки с ростом длины выборки.

Основной проблемой VC-теории является сильная завышенность оценок вероятности ошибки. Попытка их практического применения приводит либо к требованию явно избыточного наращивания обучающей выборки, либо к переупрощению семейства алгоритмов. Наиболее интересные случаи — малых выборок и сложных семейств — находятся за границами применимости VC-теории. В частности, сложные алгоритмические композиции на практике могут обеспечивать высокое качество классификации, даже когда VC-оценка вероятности ошибки равна единице. Примерами таких конструкций являются корректные линейные и алгебраические композиции алгоритмов вычисления оценок (Ю. И. Журавлёв, 1977). Нетривиальные оценки вероятности ошибки для таких композиций были получены В. Л. Матросовым в серии работ (1980–1985). Тем самым было показано, что применение сложных композиций не противоречит VC-теории. Однако эти оценки также были сильно завы-

шены, поскольку опирались на VC-теорию. Намного позже широкое распространение получили методы обучения линейных композиций — бустинг (И. Фройнд, Р. Шапир, 1995) и бэггинг (Л. Брейман, 1995). Их статистические обоснования удалось получить П. Бартлетту и др. (1998). Было показано, что верхние оценки вероятности ошибки не зависят от числа базовых алгоритмов в композиции, а только от сложности семейства базовых алгоритмов. Эти оценки опираются на усовершенствованный вариант VC-теории, но также не являются численно точными.

Основной причиной завышенности VC-оценок является их чрезмерная общность. Они справедливы для любой выборки, любой восстанавливаемой зависимости и любого метода обучения, включая худшие случаи, никогда не встречающиеся на практике. Очевидно также, что одна скалярная характеристика сложности семейства, не зависящая от решаемой задачи, не несёт достаточно информации о таком сложном оптимизационном процессе, как обучение по прецедентам. Дальнейшее развитие SLT шло по пути повышения точности оценок путём учёта индивидуальных особенностей задач и методов обучения. Большое разнообразие исследований в SLT за последние 40 лет связано с неоднозначностью ответов на вопросы: какие именно характеристики задачи, семейства алгоритмов и метода обучения наиболее существенны, и в то же время достаточно удобны для практического оценивания и управления качеством алгоритма в процессе его обучения. В идеале хотелось бы предсказывать вероятность ошибки примерно с той же точностью, с которой закон больших чисел предсказывает частоту выпадения орла при подбрасывании монеты. Однако проблема получения точных оценок обобщающей способности оказалась гораздо более сложной, и до сих пор не имеет окончательного решения.

Современные оценки основаны, главным образом, на теории эмпирических процессов и неравенствах концентрации вероятностной меры. Несмотря на развитость этих математических техник, они обладают рядом существенных недостатков. В процессе вывода верхних оценок трудно оценить количественно, на каких именно шагах происходит основная потеря точности. Не менее трудно устранить одновременно все причины завышенности (автору такие работы неизвестны). Наиболее точные оценки основаны на байесовском подходе, требующем задания субъективной априорной информации, что не всегда оправдано.

Для устранения этих недостатков в данной работе предлагается слабая вероятностная аксиоматика и комбинаторный подход, позволяющий получать точные (не завышенные, не асимптотические) оценки вероятности переобучения.

Цель работы — создание нового математического аппарата для получения точных оценок вероятности переобучения.

Научная новизна. До сих пор вопрос о получении *точных* оценок в SLT даже не ставился. Задача считалась безнадёжной, и обычно речь шла лишь о «не сильно завышенных» оценках (tight bounds). Для получения точных оценок приходится отказываться от стандартного инструментария SLT — завышенных неравенств Маркова, Хёфдинга, Чернова, МакДиармида, Буля, и др. Комбинаторный подход потребовал радикального пересмотра всей теории, начиная с аксиоматики. Впервые предложена методика эмпирического измерения факторов завышенности VC-оценок и локального эффективного коэффициента разнообразия. Предложены новые оценки обобщающей способности на основе порождающих и запрещающих множеств объектов, профилей расслоения, связности, компактности, монотонности.

Методы исследования. Вместо завышенных функционалов равномерного отклонения, введённых в VC-теории и применяемых в SLT до сих пор, вводится более точный функционал *вероятности переобучения*, зависящий от задачи и метода обучения, и основанный на принципе полного скользящего контроля.

Обычно под *скользящим контролем* понимают среднюю частоту ошибок на контрольных данных, вычисленную по небольшому (например, случайному) подмножеству разбиений выборки на обучение и контроль. При *полном* скользящем контроле берётся множество *всех* разбиений, что практически исключает возможность непосредственного точного вычисления таких функционалов. С другой стороны, для функционала вероятности переобучения справедливы те же верхние VC-оценки, что и для функционала равномерного отклонения, а предлагаемые в работе комбинаторные методы позволяют получать также и точные оценки.

Теория надёжности эмпирических предсказаний опирается не на колмогоровскую теоретико-мерную аксиоматику, а на *слабую вероятностную аксиоматику*, основанную на единственном вероятностном допущении, что все разбиения конечной генеральной выборки на обучающую и контрольную части равновероятны. Этого допущения оказывается достаточно, чтобы получить аналог закона больших чисел, установить сходимость эмпирических распределений и воспроизвести основные результаты VC-теории. Кроме того, в слабой аксиоматике естественным образом строятся непараметрические статистические критерии и доверительные интервалы.

Применяемые в данной работе методы относятся скорее к области дискретной математики, в первую очередь комбинаторики, чем к математической статистике и теории вероятностей.

В то же время, все комбинаторные результаты имеют прозрачный вероятностный смысл.

Хотя работа является теоретической, ход исследования в значительной степени определялся по результатам экспериментов на реальных и модельных задачах классификации.

Результаты, выносимые на защиту.

1. Слабая вероятностная аксиоматика, основанная на предположении о равной вероятности всех разбиений выборки.
2. VC-оценки вероятности переобучения, учитывающие степень некорректности метода обучения.
3. Методика эмпирического измерения факторов завышенности VC-оценок вероятности переобучения.
4. Блочный метод вычисления вероятности переобучения.
5. Метод получения точных оценок вероятности переобучения, основанный на порождающих и запрещающих множествах.
6. Рекуррентный алгоритм вычисления точных, верхних и нижних оценок вероятности переобучения.
7. Точные оценки вероятности переобучения для модельных семейств алгоритмов: слоя и интервала булева куба, монотонных и унимодальных цепочек, единичной окрестности.
8. Верхние оценки вероятности переобучения через профиль расслоения и связности семейства алгоритмов.
9. Точные оценки полного скользящего контроля для метода ближайшего соседа через профиль компактности выборки.
10. Верхние оценки полного скользящего контроля для монотонных алгоритмов через профиль монотонности выборки.

Теоретическая значимость. В настоящее время в теории обобщающей способности наметилась стагнация. Ценой сильного усложнения математического аппарата удаётся добиться

лишь незначительного повышения точности оценок. Интерес научного сообщества к проблематике оценок обобщающей способности заметно снизился в последние годы. Тем временем остаются открытыми фундаментальные проблемы — как преодолеть завышенность оценок, и как их использовать на практике для управления процессом обучения. Сложившаяся ситуация не раз повторялась в истории науки: очевидно, что для дальнейшего развития теории требуются радикально новые идеи и подходы. Данная работа является попыткой выхода из тупика.

Практическая значимость. Большинство оценок, полученных в данной работе, пока не нашли непосредственного практического применения, за исключением результатов главы 5. Точные оценки в большинстве случаев требуют адаптации к прикладной задаче. Ожидается, что одним из первых применений станет разработка новых методов поиска логических закономерностей и построения логических алгоритмов классификации.

Апробация работы. Результаты работы неоднократно докладывались на научных семинарах ВЦ РАН и на конференциях:

— всероссийская конференция «Математические методы распознавания образов» ММРО-7, 1995 г. [1];
— международная конференция «Интеллектуализация обработки информации» ИОИ-4, 2002 г. [4];
— всероссийская конференция «Математические методы распознавания образов» ММРО-11, 2003 г. [5];
— международная конференция «Интеллектуализация обработки информации» ИОИ-5, 2004 г. [10];
— всероссийская конференция «Математические методы распознавания образов» ММРО-12, 2005 г. [11];

— международная конференция «Интеллектуализация обработки информации» ИОИ-6, 2006 г. [12, 13];
— всероссийская конференция «Математические методы распознавания образов» ММРО-13, 2007 г. [14–19];
— 7-й открытый немецко-российский семинар «Распознавание образов и понимание изображений», Эттлинген, Германия, 20–25 августа 2007 г. [22];
— ломоносовские чтения, МГУ, 17 апреля, 2008 г.;
— международная конференция «Интеллектуализация обработки информации» ИОИ-7, 2008 г. [20];
— международная конференция «Распознавание образов и анализ изображений: новые информационные технологии» РОАИ-9, Нижний Новгород, 2008 г. [21];
— международная конференция «Современные проблемы математики, механики и их приложений», Москва, 30 марта–2 апреля 2009 г.;
— семинар «Знания и онтологии ELSEWHERE 2009», ассоциированный с 17-й международной конференцией по понятийным структурам ICCS-17, Москва, Высшая школа экономики, 21–26 июля 2009 г. [25];
— всероссийская конференция «Математические методы распознавания образов» ММРО-14, 2009 г. [26–28].

Материалы диссертационной работы легли в основу спецкурса «Теория надёжности обучения по прецедентам», читаемого студентам старших курсов на факультете ВМК Московского государственного университета им. М. В. Ломоносова.

Публикации по теме диссертации в изданиях из списка ВАК: [2, 3, 6–8, 22–24]. Другие публикации по теме диссертации: [1, 4, 5, 9–21, 26–28]. Полный текст диссертации доступен на странице автора `http://www.ccas.ru/voron`.

Структура и объём работы. Работа состоит из оглавления, введения, пяти глав, заключения, списка обозначений, списка иллюстраций (34 пункта), списка таблиц (6 пунктов), списка литературы (224 пункта) и предметного указателя.

Общий объём работы — 271 стр.

Краткое содержание работы по главам

В автореферате сохранена нумерация основных утверждений (аксиом, гипотез, определений, лемм, теорем и их следствий), принятая в тексте работы. Нумерация формул сквозная.

Глава 1. Слабая вероятностная аксиоматика

Рассмотрим фундаментальную задачу теории вероятностей, тесно связанную с законом больших чисел: оценить вероятность большого отклонения частоты $\nu(S, X)$ события S на конечной выборке X от вероятности $P(S)$ данного события:

$$P_\varepsilon = \mathsf{P}\big\{|\nu(S, X) - P(S)| > \varepsilon\big\}. \tag{1}$$

Если вероятностная мера P неизвестна, то для вычисления вероятности события $P(S)$ необходимо провести бесконечное число наблюдений, что на практике невозможно. В результате оказывается, что вероятность P_ε не может быть измерена в эксперименте как частота события $\big\{X\colon |\nu(S, X) - P(S)| > \varepsilon\big\}$, поскольку само наступление этого события не может быть точно идентифицировано. Данная проблема не возникает, если вместо вероятности $P(S)$ оценивать частоту $\nu(S, X')$ события S на произвольной случайной выборке X':

$$Q_\varepsilon = \mathsf{P}\big\{|\nu(S, X) - \nu(S, X')| > \varepsilon\big\}. \tag{2}$$

Если предполагать, что выборки X и X' независимы, то для определения вероятности Q_ε не нужно ни бесконечного числа испытаний, ни введения вероятностной меры на исходном пространстве событий. Вероятность Q_ε может быть легко измерена в эксперименте или вычислена комбинаторными методами как доля разбиений объединённой выборки $X \cup X'$ на две подвыборки, при которых имеет место большое отклонение частот.

Слабая вероятностная аксиоматика запрещает использование инфинитарных вероятностей и событий, которые не могут быть идентифицированы в эксперименте. Понятие вероятности вводится без использования теории меры и без предельного перехода к выборкам бесконечной длины.

§1.1. Пусть $\mathbb{X} = \{x_1, \ldots, x_L\}$ — фиксированное множество попарно различных объектов, называемое *генеральной выборкой*. Обозначим через S_L группу перестановок L элементов. Всевозможные перестановки элементов генеральной выборки будем обозначать через $\tau\mathbb{X}$, $\tau \in S_L$.

Аксиома 1.1. Все $L!$ перестановок генеральной выборки $\tau\mathbb{X}$, $\tau \in S_L$, имеют одинаковые шансы реализоваться.

Пусть задан предикат $\psi\colon \mathbb{X}^L \to \{0,1\}$. Если $\psi(\tau\mathbb{X}) = 1$, то будем говорить, что событие ψ произошло на перестановке $\tau\mathbb{X}$. *Вероятность события* ψ определяется как доля перестановок выборки, на которых оно произошло:

$$\mathsf{P}_\tau\, \psi(\tau\mathbb{X}) = \frac{1}{L!} \sum_{\tau \in S_L} \psi(\tau\mathbb{X}). \tag{3}$$

Пусть предикат ψ является функцией двух выборок: $X \subset \mathbb{X}$ длины ℓ и её дополнения $\bar{X} = \mathbb{X}\backslash X$ длины $k = L - \ell$, причём значение предиката $\psi(\mathbb{X}) = \varphi(X, \bar{X})$ не зависит от порядка элементов в подвыборках X и $\bar{X}$. Тогда вероятность определяется

как доля разбиений выборки $\mathbb{X}$:

$$\mathsf{P}\,\varphi(X,\bar{X}) = \frac{1}{C_L^\ell} \sum_{X\in[\mathbb{X}]^\ell} \varphi(X,\bar{X}),$$

где $[\mathbb{X}]^\ell = \big\{X \subset \mathbb{X}\colon |X| = \ell\big\}$.

Слабая аксиоматика основана на единственном вероятностном предположении, что объекты выборки появляются в случайном порядке. Оно соответствует стандартному предположению о независимости наблюдений в выборке. Столь слабого предположения оказывается достаточно, чтобы установить сходимость частот (аналог закона больших чисел), сходимость эмпирических распределений (критерий Колмогорова-Смирнова), получить многие ранговые и перестановочные критерии. Далее с позиций слабой аксиоматики рассматриваются задачи эмпирического предсказания и статистического обучения.

§1.1.1. *Задача эмпирического предсказания* состоит в том, чтобы, получив выборку X, предсказать некоторые свойства пока ещё неизвестных новых данных $\bar{X}$ и заранее оценить точность предсказания. Рассмотрим эксперимент, в котором реализуется одно из C_L^ℓ равновероятных разбиений генеральной выборки $\mathbb{X} = X \sqcup \bar{X}$. После реализации разбиения наблюдателю сообщается подвыборка X. Не зная скрытой подвыборки $\bar{X}$, требуется предсказать значение $t = T(\bar{X}, X)$ заданной функции $T\colon \mathbb{X}^k \times \mathbb{X}^\ell \to R$, существенно зависящее от скрытой подвыборки $\bar{X}$. Для этого строится *предсказывающая функция* $\hat{T}\colon \mathbb{X}^\ell \to R$, значение которой на наблюдаемой подвыборке $\hat{t} = \hat{T}(X)$ приближало бы неизвестное значение $t = T(\bar{X}, X)$. Требуется оценить надёжность предсказаний, указав невозрастающую *оценочную функцию* $\eta(\varepsilon)$ такую, что

$$\mathsf{P}\big[d\big(\hat{T}(X), T(\bar{X}, X)\big) > \varepsilon\big] \leqslant \eta(\varepsilon), \tag{4}$$

где $d\colon R \times R \to \mathbb{R}$ — заданная функция, характеризующая величину отклонения $d(\hat{t}, t)$ предсказанного значения $\hat{t}$ от неизвестного истинного значения t. Параметр ε называется *точностью*, а величина $(1 - \eta(\varepsilon))$ — *надёжностью* предсказания. Если в (4) достигается равенство, то $\eta(\varepsilon)$ называется *точной оценкой*.

Выбирая множество R, функции T, $\hat{T}$ и d, можно получать многие классические задачи теории вероятностей, математической статистики, статистического обучения.

§1.1.2. §1.1.3. §1.1.4. Рассматриваются некоторые базовые приёмы обращения с оценками надёжности эмпирических предсказаний в слабой аксиоматике, а также связь задач эмпирического предсказания и статистической проверки гипотез.

§1.1.5. Рассматривается связь слабой аксиоматики с сильной (колмогоровской) аксиоматикой. Обсуждаются преимущества, недостатки и границы применимости слабой аксиоматики.

§1.2. Пусть $S \subseteq \mathbb{X}$ — некоторое множество объектов; будем называть его «событием». Введём функции числа элементов и частоты события S на произвольной конечной выборке $U \subseteq \mathbb{X}$:

$$n(U) = |S \cap U|, \quad \nu(U) = n(U)/|U|.$$

Задача оценивания частоты события состоит в том, чтобы предсказать частоту на скрытой выборке $\bar{X}$ по частоте на наблюдаемой выборке X и оценить надёжность предсказания:

$$\mathsf{P}\big[\nu(\bar{X}) - \nu(X) \geqslant \varepsilon\big] \leqslant \eta(\varepsilon); \tag{5}$$

Лемма 1.2. Если $n(\mathbb{X}) = m$, то число элементов события S в наблюдаемой подвыборке $n(X)$ подчиняются гипергеометрическому распределению:

$$\mathsf{P}\big[n(X) = s\big] = h_L^{\ell,m}(s) = \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}, \tag{6}$$

где s принимает значения от $s_0 = \max\{0, m-k\}$ до $s_1 = \min\{\ell, m\}$.

Теорема 1.3. Если $n(\mathbb{X}) = m$, то для любого $\varepsilon \in [0, 1)$

$$\mathsf{P}\big[\nu(\bar{X}) \geqslant \varepsilon\big] = H_L^{\ell,m}\left(\lfloor m - \varepsilon k \rfloor\right); \tag{7}$$

$$\mathsf{P}\big[\nu(\bar{X}) - \nu(X) \geqslant \varepsilon\big] = H_L^{\ell,m}\left(\left\lfloor \tfrac{\ell}{L}(m - \varepsilon k)\right\rfloor\right), \tag{8}$$

где $H_L^{\ell,m}(z) = \sum\limits_{s=s_0}^{\lfloor z \rfloor} h_L^{\ell,m}(s)$ — функция гипергеометрического распределения.

При пропорциональном увеличении L, ℓ и m относительная ширина гипергеометрического пика уменьшается. Поэтому в пределе при $L, \ell, m \to \infty$ возможно сколь угодно точно предсказывать частоту события S в скрытой выборке $\nu(\bar{X})$ по его частоте на наблюдаемой выборке $\nu(X)$. Равенство (8) даёт точную оценку скорости сходимости частот (справедлива также аналогичная двусторонняя оценка). Классический закон больших чисел утверждает сходимость частоты события к её вероятности. В слабой аксиоматике понятие «вероятности события S» не определено, поэтому (8) можно интерпретировать как *аналог закона больших чисел* в слабой аксиоматике.

§1.3. Рассматривается задача оценивания функции распределения, тесно связанная с двухвыборочным критерием Колмогорова-Смирнова. Выводятся точные формулы для распределения величины равномерного отклонения эмпирических функций распределения на двух выборках. Это распределение описывается усечённым треугольником Паскаля. В рамках слабой аксиоматики критерий Смирнова, обычно формулируемый для случайных величин с непрерывным распределением, легко обобщается и на случай дискретных величин.

§1.4. Демонстрируется возможность получения стандартных статистических критериев и доверительных оценок в рам-

ках слабой аксиоматики. При этом сначала выводятся точные комбинаторные формулы и результат формулируется в терминах конечных выборок. Затем, при необходимости, применяется предельный переход $L \to \infty$. Его можно понимать как способ получения приближёной формулы, не подразумевающий существования выборок сколь угодно большой длины.

§1.5. Задача оценивания вероятности переобучения является основной в данной работе. Задано множество A, элементы которого называются *алгоритмами*. Существует бинарная функция $I\colon A \times \mathbb{X} \to \{0,1\}$, называемая *индикатором ошибки*. Если $I(a,x) = 1$, то алгоритм a допускает ошибку на объекте x, если же $I(a,x) = 0$, то a даёт верный ответ на x. Определяются функции числа ошибок $n(a,U) = \sum\limits_{x \in U} I(a,x)$ и частоты ошибок $\nu(a,U) = n(a,U)/|U|$ алгоритма a на выборке $U \subseteq \mathbb{X}$.

Бинарный вектор-столбец $\vec{a} = \big(I(a,x_i)\big)_{i=1}^{L}$ называется *вектором ошибок* алгоритма a. Совокупность всех попарно различных векторов ошибок, порождаемых алгоритмами $a \in A$, образует *матрицу ошибок* размера $L \times D$. Строки этой матрицы соответствуют объектам, столбцы — алгоритмам.

Далее предполагается, что A — это конечное множество алгоритмов с попарно различными векторами ошибок.

Методом обучения называется отображение $\mu\colon 2^{\mathbb{X}} \to A$, которое произвольной *обучающей выборке* $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $a = \mu X$ из A.

Метод μ называется методом *минимизации эмпирического риска*, если для любого $X \subset \mathbb{X}$

$$\begin{aligned} \mu X \in A(X) &= \operatorname{Arg}\min_{a \in A} n(a,X) = \\ &= \big\{a \in A\colon n(a,X) \leqslant n(a',X),\ \forall a' \in A\big\}. \end{aligned} \tag{9}$$

Переобученностью метода μ относительно пары выборок X и $\bar{X}$ называется отклонение частоты ошибок алгоритма $a = \mu X$ на скрытой *контрольной* выборке $\bar{X}$ от частоты его ошибок на наблюдаемой *обучающей* выборке X:

$$\delta_\mu(X, \bar{X}) = \nu(\mu X, \bar{X}) - \nu(\mu X, X).$$

Если $\delta_\mu(X, \bar{X}) \geqslant \varepsilon$ при некотором достаточно малом $\varepsilon \in (0, 1)$, то говорят, что метод μ *переобучен* относительно $X, \bar{X}$.

Итак, основная задача заключает в том, чтобы оценить *вероятность переобучения*:

$$Q_\varepsilon \equiv Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\delta_\mu(X, \bar{X}) \geqslant \varepsilon\big]. \tag{10}$$

§1.5.3. Вводятся некоторые важные в VC-теории понятия.

$\vec{A} = \big\{\vec{a}\colon a \in A\big\}$ — множество векторов ошибок, порождаемых заданным множеством алгоритмов A.

$\Delta(A, \mathbb{X}) = |\vec{A}|$ — *коэффициент разнообразия* множества алгоритмов A на выборке $\mathbb{X}$. В задачах классификации на два класса коэффициент разнообразия равен числу различных *дихотомий* (способов разделить выборку $\mathbb{X}$ на два класса), реализуемых всевозможными алгоритмами из A.

$A_L^\ell \equiv A_L^\ell(\mu, \mathbb{X}) = \big\{\mu X\colon X \in [\mathbb{X}]^\ell\big\}$ — множество алгоритмов, индуцируемых методом обучения μ на всевозможных обучающих подвыборках X.

$\Delta_L^\ell \equiv \Delta_L^\ell(\mu, \mathbb{X}) = \Delta\big(A_L^\ell(\mu, \mathbb{X}), \mathbb{X}\big)$ — *локальный коэффициент разнообразия* метода μ на выборке $\mathbb{X}$.

$\Delta^A(L) = \max\limits_{\mathbb{X}} \Delta(A, \mathbb{X})$ — *глобальный коэффициент разнообразия* или *функция роста* множества алгоритмов A. Максимум берётся по всевозможным выборкам $\mathbb{X} \subset \mathscr{X}$ длины L из некоторого (как правило, бесконечного) множества допустимых объектов $\mathscr{X}$. Функция роста является мерой сложности множества

алгоритмов A. В отличие от локального коэффициента разнообразия, она не зависит ни от задачи, ни от метода обучения μ.

$A_m = \{a \in A\colon n(a, \mathbb{X}) = m\}$ — подмножество алгоритмов, называемое m-м *слоем* множества A. Будем также говорить, что множество A *расслаивается по уровням ошибок*.

$\Delta_m \equiv \Delta_m(\mu, \mathbb{X}) = \Delta\big((A_L^\ell)_m, \mathbb{X}\big)$ — локальный коэффициент разнообразия m-го слоя множества $A_L^\ell(\mu, \mathbb{X})$. Совокупность величин $(\Delta_m)_{m=0}^L$ будем называть *профилем расслоения*.

§1.5.4. VC-теория переносится в слабую аксиоматику.

Теорема 1.12. Для любых μ, $\mathbb{X}$ и $\varepsilon \in [0, 1]$

$$Q_\varepsilon \leqslant \sum_{m=\lceil \varepsilon k \rceil}^{L} \Delta_m H_L^{\ell,m}\left(\left\lfloor \tfrac{\ell}{L}(m - \varepsilon k)\right\rfloor\right) \leqslant$$
$$\leqslant \Delta_L^\ell \max_{m=1,\ldots,L} H_L^{\ell,m}\left(\left\lfloor \tfrac{\ell}{L}(m - \varepsilon k)\right\rfloor\right). \qquad (11)$$

Оценка (11) имеет следующую интерпретацию. Вероятность переобучения не превышает вероятности большого отклонения частот для наихудшего алгоритма (максимум $H_L^{\ell,m}$ достигается при $m \approx L/2$), умноженной на число алгоритмов.

Оценивая локальный коэффициент разнообразия глобальным $\Delta_L^\ell(\mu, \mathbb{X}) \leqslant \Delta^A(L)$ и заменяя функцию гипергеометрического распределения экспоненциальной верхней оценкой, получаем ещё одну известную классическую VC-оценку:

Следствие 1.12.2. Для любых μ, $\mathbb{X}$, $\varepsilon \in [0, 1)$ при $\ell = k$

$$Q_\varepsilon \leqslant \Delta^A(2\ell) \cdot \tfrac{3}{2} e^{-\varepsilon^2 \ell}. \qquad (12)$$

§1.5.5. В VC-теории отдельно рассматривается *детерминистская постановка задачи*, когда $n(\mu X, X) = 0$ на любой обучающей выборке X. Однако в общей постановке задачи не делается никаких предположений о величине $n(\mu X, X)$. В данной работе предлагается исследовать и промежуточные ситуации.

Степенью некорректности метода обучения μ на выборке $\mathbb{X}$ называется максимальная частота ошибок на обучении

$$\sigma(\mu, \mathbb{X}) = \max_{X \in [\mathbb{X}]^\ell} \nu(\mu X, X).$$

Теорема 1.13. Для любых μ, $\mathbb{X}$ с ограниченной некорректностью $\sigma(\mu, \mathbb{X}) \leqslant \sigma$ и любого $\varepsilon \in [0, 1]$ справедлива оценка

$$Q_\varepsilon \leqslant \Delta_L^\ell \max_{m:\ \varepsilon k \leqslant m \leqslant k + \sigma\ell} H_L^{\ell,m} \left(\min\left\{ \left\lfloor \tfrac{\ell}{L}(m - \varepsilon k) \right\rfloor, \sigma\ell \right\} \right). \qquad (13)$$

Следствие 1.13.1. Если метод μ *корректный на генеральной выборке* $\mathbb{X}$, то есть $\sigma(\mu, \mathbb{X}) = 0$, то для любых μ, $\mathbb{X}$ и $\varepsilon \in [0, 1]$

$$Q_\varepsilon \leqslant \sum_{m=\lceil \varepsilon k \rceil}^{\lfloor k + \sigma\ell \rfloor} \Delta_m \frac{C_{L-m}^\ell}{C_L^\ell} \leqslant \Delta^A(L) \frac{C_{L-\lceil \varepsilon k \rceil}^\ell}{C_L^\ell} \leqslant \Delta^A(L) \left(\frac{k}{L} \right)^{\varepsilon k}. \qquad (14)$$

Следствие 1.13.2. Если $\ell = k$ и метод μ *корректный на генеральной выборке* $\mathbb{X}$, то для любых μ, $\mathbb{X}$ и $\varepsilon \in [0, 1]$

$$Q_\varepsilon \leqslant \Delta^A(2\ell)\, 2^{-\varepsilon\ell}. \qquad (15)$$

§1.5.6. Основной недостаток VC-оценок в том, что они чрезвычайно завышены — настолько, что их применение практически теряет смысл. Для наглядности в работе приводятся результаты численного расчёта требуемой длины обучающей выборки ℓ как функции от ёмкости h, точности ε и уровня значимости (надёжности) Q_ε. Расчётные значения ℓ на порядки превосходят характерные длины выборок в практических задачах.

В детерминистском случае требуемая длина обучения ℓ заметно меньше, но также сильно завышена. Учёт априорной информации о степени некорректности уточняет VC-оценки, но не решает проблему завышенности.

§1.5.7. Причины завышенности видны из доказательства теоремы 1.12, в котором сделаны три оценки сверху, а при выводе следствия 1.13.1 — ещё две. Те же причины остаются и в теореме 1.13. Всего имеется 5 факторов завышенности.

1. Принцип равномерной сходимости приводит к завышенности, когда множество A_L^ℓ расслаивается по уровням ошибок.

2. Неравенство Буля приводит к сильной завышенности, когда среди векторов ошибок имеется много похожих.

3. Пренебрежение профилем расслоения является малозначимым фактором.

4. Пренебрежение эффектом локализации приводит к сильной завышенности, когда локальное подмножество $A_L^\ell(\mu, \mathbb{X})$ много меньше всего семейства A.

5. Экспоненциальная оценка гипергеометрического распределения представляет результат в удобной форме, но в компьютерных вычислениях её использование едва ли оправдано.

Глава 2. Теория статистического обучения

В данной главе представлен обзор оценок обобщающей способности, начиная с VC-теории, и заканчивая работами последних лет. Особое внимание уделяется оценкам, показывающим, что обобщающая способность может не ухудшаться с ростом сложности семейства, а также оценкам, учитывающим эффекты расслоения и сходства в семействах алгоритмов. Современные подходы существенно улучшают VC-оценки, но всё ещё не дают точных оценок. Более того, применение сложных математических методов затрудняет понимание причин завышенности и постановку экспериментов по измерению факторов завышенности. Поэтому в следующей главе за основу берутся клас-

сические VC-оценки, для которых методика эмпирического измерения завышенности разрабатывается сравнительно легко.

Глава 3. Эмпирический анализ факторов завышенности VC-оценок

Описывается методика экспериментального количественного измерения факторов завышенности VC-оценок. В рамках классической VC-теории такие измерения практически невозможны, поскольку в функционале равномерной сходимости

$$P_\varepsilon = \mathsf{P}\big\{ \sup_{a\in A}(P(a) - \nu(a,X)) \geqslant \varepsilon \big\}$$

вероятности $P(a)$ неизвестны, а супремум берётся по чрезвычайно широкому множеству A. Хотя теоретически вероятность любого события можно оценить по конечному числу наблюдений, в данном случае вероятность P_ε не удаётся оценить как частоту события, поскольку само наступление этого события трудно идентифицировать.

В слабой аксиоматике оценки *вероятности переобучения*

$$Q_\varepsilon = \mathsf{P}\big\{\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon\big\},$$

получаются аналогично классическим VC-оценкам. При этом Q_ε легко измеряется по небольшому подмножеству разбиений $(X, \bar{X})$, в частности, методом Монте-Карло.

С методологической точки зрения интересно отметить, что именно стремление выделить и измерить факторы завышенности VC-оценок и привело к отказу от принципа равномерной сходимости и введению слабой вероятностной аксиоматики.

§3.1. Зная эмпирическую оценку $\hat{Q}_\varepsilon$, можно сказать, какое значение должна была бы иметь функция роста Δ, чтобы VC-оценка $Q_\varepsilon \leqslant \Delta\eta(\varepsilon)$ не была завышенной и обращалась в точное

равенство. Здесь $\eta(\varepsilon)$ — вероятность большого отклонения частот в двух выборках для отдельного алгоритма. Гипотетическое точное значение функции роста $\Delta = Q_\varepsilon/\eta(\varepsilon)$ предлагается называть *эффективным локальным коэффициентом разнообразия* (ЭЛКР). Возможна ещё одна интерпретация ЭЛКР — он показывает, во сколько раз снижается надёжность предсказания качества обучения по сравнению с предсказанием частоты ошибок отдельного алгоритма, которое даёт закон больших чисел.

§3.2. Эксперименты с логическими алгоритмами классификации на реальных задачах из репозитория UCI показали, что среди всех факторов завышенности наиболее существенными являются два — это игнорирование свойств расслоения и связности семейства алгоритмов. Каждый из этих двух факторов завышает оценку Q_ε в 10^3–10^5 раз. На практике ЭЛКР принимает значения порядка 10^0–10^2, тогда как функция роста обычно имеет порядок 10^5–10^{10} и выше.

Расслоение является следствием универсальности применяемых на практике семейств. Как правило, лишь ничтожная доля алгоритмов в семействе подходит для решения фиксированной задачи, но именно эти алгоритмы имеют наибольшие шансы быть полученными в результате обучения. Распределение вероятностей на множестве алгоритмов оказывается существенно неравномерным, однако этот факт никак не учитывается классическими VC-оценками.

Связность является следствием непрерывности применяемых на практике семейств. Как правило, для любого алгоритма в семействе существует большое число похожих на него алгоритмов. В частности, связные семейства порождаются методами классификации с непрерывной по параметрам разделяющей поверхностью: линейные классификаторы, машины опор-

ных векторов с непрерывными ядрами, нейронные сети с непрерывными функциями активации, решающие деревья с пороговыми условиями ветвления, и многие другие. Чем больше в семействе схожих алгоритмов, тем сильнее завышено неравенство Буля, используемое при выводе VC-оценки. Классические VC-оценки ориентированы на «худший случай», когда все алгоритмы существенно различны, но этот случай практически никогда не реализуется в приложениях.

§3.3. Следующая серия экспериментов на модельных данных подтверждает необходимость совместного учёта эффектов расслоения и связности. Рассматривается простейшее семейство с расслоением и связностью — монотонная цепочка алгоритмов. Строятся его естественные аналоги, не обладающие либо свойством расслоения, либо свойством связности. В обоих случаях вероятность переобучения оказывается значительной уже при нескольких десятках алгоритмов в семействе. Отсюда следуют два важных вывода. Во-первых, все реальные семейства с необходимостью расслоены и связны (а если и не связны, то обладают какой-либо иной структурой сходства алгоритмов). Во-вторых, только при совместном учёте обоих свойств возможно получение точных оценок вероятности переобучения.

§3.3.1. Третий эксперимент на модельном семействе, состоящем только из двух алгоритмов, показывает, что даже в этом простейшем случае появляется переобучение, а эффекты расслоения и сходства снижают вероятность переобучения.

Поскольку матрицы ошибок модельных семейств обладают вполне определённой структурой, для них возможно получать точные формулы вероятности переобучения, используя чисто комбинаторные методы. На этом этапе исследования точные формулы были получены для нескольких «искусственных» се-

мейств простой структуры. Затем эти оценки были обобщены и обнаружен общий механизм их вывода, основанный на порождающих и разрушающих множествах объектов.

Глава 4.
Точные оценки вероятности переобучения

В данной главе выводятся точные оценки вероятности переобучения, основанные на предположении, что для каждого алгоритма $a \in A$ возможно в явном виде выписать условия, при которых a является результатом обучения, $\mu X = a$.

§4.1.1. Обозначим через $A(X) = \operatorname{Arg}\min_{a \in A} n(a, X)$ множество алгоритмов с минимальным числом ошибок на обучении X.

Определение 4.1. Метод μ называется *минимизацией эмпирического риска* (МЭР), если $\mu X \in A(X)$ при всех $X \subset \mathbb{X}$.

Если множество $A(X)$ содержит более одного элемента, то в методе μ возникает проблема неоднозначности выбора алгоритма. Рассмотрим два крайних случая.

Определение 4.2. Минимизация эмпирического риска μ называется *оптимистичной*, если $\mu X = \arg\min_{a \in A(X)} n(a, \bar{X})$.

Определение 4.3. Минимизация эмпирического риска μ называется *пессимистичной*, если $\mu X = \arg\max_{a \in A(X)} n(a, \bar{X})$.

Оба варианта на практике не реализуемы, так как контрольная выборка $\bar{X}$ скрыта на этапе обучения. Теоретически они интересны тем, что дают точные нижние и верхние оценки вероятности переобучения. Большинство получаемых в работе оценок основаны на пессимистичной МЭР.

§4.1.2. Вводится предположение, что получение каждого из алгоритмов в результате обучения связано с наличием или отсутствием некоторых объектов в обучающей выборке.

Гипотеза 4.1. Множество A, выборка $\mathbb{X}$ и метод μ таковы, что для каждого $a \in A$ можно указать пару непересекающихся подмножеств $X_a \subset \mathbb{X}$ и $X'_a \subset \mathbb{X}$, удовлетворяющую условию

$$\big[\mu X{=}a\big] = \big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^\ell. \tag{16}$$

Множество X_a называется *порождающим*, множество X'_a — *запрещающим* алгоритм a. Остальные объекты $\mathbb{X}\backslash X_a\backslash X'_a$ называются *нейтральными* для алгоритма a, их наличие или отсутствие в обучающей выборке не влияет на результат обучения.

Для произвольного $a \in A$ введём обозначения:

$$L_a = |\mathbb{X}\backslash X_a\backslash X'_a|; \qquad m_a = n(a, \mathbb{X}\backslash X_a\backslash X'_a);$$
$$\ell_a = |X\backslash X_a|; \qquad s_a(\varepsilon) = \tfrac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big) - n(a, X_a).$$

Лемма 4.2. Если гипотеза 4.1 справедлива, то

$$P(a) = \mathsf{P}\big[\mu X{=}a\big] = C^{\ell_a}_{L_a}/C^\ell_L.$$

Теорема 4.3. Если гипотеза 4.1 справедлива, то вероятность переобучения вычисляется по формуле

$$Q_\varepsilon = \sum_{a\in A} P(a) H^{\ell_a, m_a}_{L_a}\left(s_a(\varepsilon)\right).$$

Гипотеза 4.1 накладывает настолько сильные ограничения на $\mathbb{X}$, A и μ, что теорему 4.3 удаётся применять лишь в специальных случаях. Рассмотрим её естественное обобщение.

Гипотеза 4.2. Множество A, выборка $\mathbb{X}$ и метод μ таковы, что для каждого $a \in A$ можно указать конечное множество индексов V_a, и для каждого индекса $v \in V_a$ можно указать *порождающее* множество $X_{av} \subset \mathbb{X}$, *запрещающее* множество $X'_{av} \subset \mathbb{X}$ и коэффициент $c_{av} \in \mathbb{R}$, удовлетворяющие условиям

$$\big[\mu X{=}a\big] = \sum_{v\in V_a} c_{av}\big[X_{av} \subseteq X\big]\big[X'_{av} \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^\ell. \tag{17}$$

Гипотеза 4.1 является частным случаем гипотезы 4.2, когда все множества V_a одноэлементные и $c_{av} = 1$.

Теорема 4.4. Гипотеза 4.2 верна всегда: для любых $\mathbb{X}$, A и μ существуют множества V_a, X_{av}, X'_{av}, при которых имеет место (17), причём $c_{av} = 1$ для всех $a \in A,\ v \in V_a$.

Теорема 4.4 является типичной теоремой существования. Использованный при её доказательстве способ построения индексных множеств V_a требует явного перебора всех разбиений выборки, что приводит к вычислительно неэффективным оценкам Q_ε. В общем случае представление (17) не единственно. Отдельной проблемой является поиск представлений с минимальными по мощности множествами V_a, X_{av}, X'_{av}.

Введём для каждого $a \in A$ и каждого $v \in V_a$ обозначения:

$$L_{av} = |\mathbb{X}\backslash X_{av}\backslash X'_{av}|; \qquad m_{av} = n(a, \mathbb{X}\backslash X_{av}\backslash X'_{av});$$
$$\ell_{av} = |X\backslash X_{av}|; \qquad s_{av}(\varepsilon) = \tfrac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big) - n(a, X_{av}).$$

Лемма 4.5. Если гипотеза 4.2 верна, то для всех $a \in A$

$$P(a) = \mathsf{P}\big[\mu X{=}a\big] = \sum_{v \in V_a} c_{av} P_{av}; \tag{18}$$

$$P_{av} = \mathsf{P}\big[X_{av} \subseteq X\big]\big[X'_{av} \subseteq \bar{X}\big] = C^{\ell_{av}}_{L_{av}} / C^{\ell}_{L}. \tag{19}$$

Теорема 4.6. Если гипотеза 4.2 справедлива, то вероятность переобучения вычисляется по формуле

$$Q_\varepsilon = \sum_{a \in A} \sum_{v \in V_a} c_{av} P_{av} H^{\ell_{av}, m_{av}}_{L_{av}} \left(s_{av}(\varepsilon)\right). \tag{20}$$

Формула (20) сильно упрощается, если в семействе A содержится корректный алгоритм a_0, не допускающий ошибок на генеральной выборке $\mathbb{X}$.

Теорема 4.7. Пусть гипотеза 4.2 справедлива, метод μ минимизирует эмпирический риск, множество A содержит алгоритм a_0 такой, что $n(a_0, \mathbb{X}) = 0$. Тогда

$$Q_\varepsilon = \sum_{a \in A} \big[n(a, \mathbb{X}) \geqslant \varepsilon k \big] P(a). \tag{21}$$

§4.1.3. Получены также оценки, не требующие задания порождающих и запрещающих множеств. Они основаны на разбиении генеральной выборки на блоки, имеют довольно громоздкий вид и эффективны только для семейств малой мощности. В частности, с помощью блочного метода легко выводится оценка вероятности переобучения для пары алгоритмов.

§4.2. Рассматриваются модельные семейства алгоритмов, для которых получаются точные оценки вероятности переобучения. Модельные семейства задаются непосредственно бинарной матрицей ошибок, а не какими-либо реальными данными и методами обучения. Они отличаются определённой «регулярностью» или симметрией, которой реальные семейства, как правило, не обладают. Модельные семейства хорошо иллюстрируют эффекты расслоения и связности. Кроме того, рассмотрение большого числа разнообразных частных случаев ведёт к постепенному обобщению модельных семейств и получению оценок, неплохо аппроксимирующих реальные семейства. Такой путь развития комбинаторной теории переобучения представляется наиболее реалистичным.

§4.2.1. Рассматривается множество из двух алгоритмов $A = \{a_1, a_2\}$. Пусть в выборке $\mathbb{X}$ имеется m_{11} объектов, на которых оба алгоритма допускают ошибку; m_{10} и m_{01} объектов, на которых, соответственно только a_1 или только a_2 допускает ошибку.

Теорема 4.9. Пусть μ — пессимистичная минимизация эмпирического риска, $A = \{a_1, a_2\}$. Тогда при любом $\varepsilon \in [0, 1)$

$$Q_\varepsilon = \sum_{s_{11}=0}^{m_{11}} \sum_{s_{10}=0}^{m_{10}} \sum_{s_{01}=0}^{m_{01}} \frac{C_{m_{11}}^{s_{11}} C_{m_{10}}^{s_{10}} C_{m_{01}}^{s_{01}} C_{L-m_{11}-m_{10}-m_{01}}^{\ell-s_{11}-s_{10}-s_{01}}}{C_L^\ell} \times$$
$$\times \Big(\big[s_{10} < s_{01}\big] \big[s_{11} + s_{10} \leqslant \tfrac{\ell}{L}(m_{11} + m_{10} - \varepsilon k)\big] +$$
$$+ \big[s_{10} \geqslant s_{01}\big] \big[s_{11} + s_{01} \leqslant \tfrac{\ell}{L}(m_{11} + m_{01} - \varepsilon k)\big] \Big). \qquad (22)$$

§4.2.2. Рассматривается множество A, состоящее из всех C_L^m алгоритмов с попарно различными векторами ошибок и ровно m ошибками на полной выборке $\mathbb{X}$. Поскольку все возможные векторы ошибок образуют булев куб размерности L, то векторы ошибок множества A — это m-й слой булева куба.

Теорема 4.10. Пусть μ — произвольный метод минимизации эмпирического риска, A — m-й слой булева куба. Тогда $Q_\varepsilon = \big[\varepsilon k \leqslant m \leqslant \ell - \varepsilon\ell\big]$ для любого $\varepsilon \in [0, 1]$.

Хотя оценка вырождена и принимает только два значения, 0 или 1, из неё следуют два важных вывода. Во-первых, алгоритмы нижних слоёв $m < \lceil \varepsilon k \rceil$ не вносят вклад в переобучение. Во-вторых, никакой слой выше $m = \lceil \varepsilon k \rceil$ не должен целиком содержаться в семействе алгоритмов.

§4.2.3. Рассматривается множество A, образующее интервал ранга m в L-мерном булевом кубе. Объекты делятся на три категории: m_0 «внутренних» объектов, на которых ни один из алгоритмов не допускает ошибок; m_1 «шумовых» объектов, на которых все алгоритмы допускают ошибки; и m «пограничных» объектов, на которых реализуются все 2^m вариантов допустить ошибки. Интервал булева куба обладает свойствами расслоения и связности и может рассматриваться как модель реального семейства размерности m.

Теорема 4.11. Пусть μ — пессимистичная МЭР, A — интервал булева куба. Тогда для любого $\varepsilon \in [0,1]$

$$Q_\varepsilon = \sum_{s=0}^{m} \sum_{s_1=0}^{m_1} \frac{C_m^s C_{m_1}^{s_1} C_{L-m-m_1}^{\ell-s-s_1}}{C_L^\ell} \big[s_1 + \tfrac{\ell}{L} s \leqslant \tfrac{\ell}{L}(m_1 + m - \varepsilon k) \big].$$

В работе получена также точная оценка Q_ε для интервала булева куба при рандомизированной МЭР.

§4.2.4. Рассматривается множество A, образованное t нижними слоями интервала ранга m в L-мерном булевом кубе. Это семейство интересно тем, что оно позволяет исследовать влияние эффекта расслоения на вероятность переобучения, построив зависимость Q_ε от числа нижних слоёв t.

Теорема 4.13. Пусть μ — пессимистичная МЭР, A — нижние t слоёв интервала булева куба. Тогда для любого $\varepsilon \in [0,1]$

$$Q_\varepsilon = \sum_{s=0}^{m} \sum_{s_1=0}^{m_1} \frac{C_m^s C_{m_1}^{s_1} C_{L-m-m_1}^{\ell-s-s_1}}{C_L^\ell} \big[s_1 \leqslant \tfrac{\ell}{L}(m_1 + \min\{t, m-s\} - \varepsilon k) \big].$$

Для множества t нижних слоёв интервала булева куба также получена точная оценка Q_ε при рандомизированной МЭР.

§4.2.5. Рассматривается *монотонная цепочка алгоритмов*, которую можно интерпретировать как простейшую модель однопараметрического связного семейства алгоритмов.

Определим расстояние Хэмминга между алгоритмами:

$$\rho(a, a') = \sum_{i=1}^{L} \big| I(a, x_i) - I(a', x_i) \big|, \quad \forall a, a' \in A.$$

Множество алгоритмов $A = \{a_0, a_1, \ldots, a_D\}$ называется *цепочкой алгоритмов*, если $\rho(a_{d-1}, a_d) = 1$ для всех $d = 1, \ldots, D$.

Цепочка алгоритмов $A = \{a_0, a_1, \ldots, a_D\}$ называется *монотонной*, если $n(a_d, \mathbb{X}) = m + d$ при некотором $m \geqslant 0$.

Алгоритм a_0 называется *лучшим в цепочке*.

Теорема 4.16. Пусть $A = \{a_0, a_1, \ldots, a_D\}$ — монотонная цепочка; $L \geqslant m + D$. Тогда в случае $D \geqslant k$

$$Q_\varepsilon = \sum_{d=0}^{k} P_d H^{\ell-1,m}_{L-d-1}\left(s_d(\varepsilon)\right); \qquad P_d = \frac{C^{\ell-1}_{L-d-1}}{C^\ell_L},\ d = 0, \ldots, k;$$

в случае $D < k$

$$Q_\varepsilon = \sum_{d=0}^{D-1} P_d H^{\ell-1,m}_{L-d-1}\left(s_d(\varepsilon)\right) \ + \ P_D H^{\ell,m}_{L-D}\left(s_D(\varepsilon)\right);$$
$$P_d = \frac{C^{\ell-1}_{L-d-1}}{C^\ell_L},\ d = 0, \ldots, D-1; \qquad P_D = \frac{C^\ell_{L-D}}{C^\ell_L},$$

где $P_d = \mathsf{P}\big[\mu X{=}a_d\big]$, $s_d(\varepsilon) = \frac{\ell}{L}(m + d - \varepsilon k)$.

§4.2.6. *Унимодальная цепочка алгоритмов* является более реалистичной моделью однопараметрического связного семейства, по сравнению с монотонной цепочкой.

Множество алгоритмов $A = \{a_0, a_1, \ldots, a_D, a'_1, \ldots, a'_{D'}\}$ называется *унимодальной цепочкой*, если *левая ветвь* $a_0, a_1, \ldots, a_D$ и *правая ветвь* $a_0, a'_1, \ldots, a'_{D'}$ являются монотонными цепочками с общим лучшим алгоритмом a_0.

В работе получены точные оценки вероятности переобучения для унимодальной цепочки.

§4.2.7. *Единичная окрестность лучшего алгоритма* является «экстремальным» частным случаем, когда алгоритмы максимально близки друг к другу, и классические оценки, основанные на неравенстве Буля, наиболее завышены.

Множество алгоритмов $A = \{a_0, a_1, \ldots, a_D\}$ называется *единичной окрестностью* алгоритма a_0, если все векторы ошибок a_d попарно различны, $n(a_d, \mathbb{X}) = n(a_0, \mathbb{X}) + 1$ и $\rho(a_0, a_d) = 1$ для

всех $d = 1, \ldots, D$. Алгоритм a_0 называется *лучшим в окрестности* или *центром окрестности*.

В работе получены точные оценки вероятности переобучения для единичной окрестности.

§4.2.8. Краткий обзор точных и верхних оценок вероятности переобучения, полученных другими авторами в рамках предлагаемого в данной работе комбинаторного подхода.

§4.3. Пусть в A есть корректный на $\mathbb{X}$ алгоритм и μ — пессимистичный метод МЭР. Задача вычисления Q_ε сводится к тому, чтобы найти информацию $\mathfrak{I}(a) = \langle X_{av}, X'_{av}, c_{av} \rangle_{v \in V_a}$ для каждого алгоритма $a \in A$, где V_a — индексное множество, X_{av} — множество порождающих объектов, X'_{av} — множество запрещающих объектов, $c_{av} \in \mathbb{R}$.

Обозначим через μ_d метод обучения μ, выбирающий алгоритмы только из подмножества $A_d = \{a_0, \ldots, a_d\}$. Рассмотрим переход от метода μ_{d-1} к методу μ_d при добавлении в A_{d-1} алгоритма a_d. Допустим, что для всех алгоритмов $a_t,\ t < d$, информация $\mathfrak{I}(a_t)$ относительно метода μ_{d-1} уже известна. Поставим задачу: вычислить информацию $\mathfrak{I}(a_d)$ и скорректировать информацию $\mathfrak{I}(a_t),\ t < d$, относительно метода μ_d. Необходимость коррекции вызвана тем, что алгоритм a_d может «отбирать некоторые разбиения» у каждого из предыдущих алгоритмов a_t.

Лемма 4.19. Метод μ_d выбирает алгоритм a_d тогда и только тогда, когда все объекты, на которых a_d допускает ошибку, попадают в контрольную выборку:

$$\big[\mu_d X{=}a_d\big] = \big[X'_d \subseteq \bar{X}\big], \quad X'_d = \big\{x_i \in \mathbb{X}\colon I(a_d, x_i) = 1\big\}.$$

Лемма 4.20. Корректировка информации $\mathfrak{I}(a_t),\ t < d$ при добавлении алгоритма a_d сводится к проверке для каждого $v \in V_t$ такого, что $X_{tv} \cap X'_d = \varnothing$, трёх условий:

1) если $X'_d \setminus X'_{tv} = \{x_i\}$ — одноэлементное множество, то x_i присоединяется к X_{tv};
2) если $|X'_d \setminus X'_{tv}| > 1$, то множество индексов V_t пополняется элементом w, полагая $c_{tw} = -c_{tv},\ X_{tw} = X_{tv},\ X'_{tw} = X'_{tv} \cup X'_d$;
3) если $|X'_d \setminus X'_{tv}| = 0$, то из множества индексов V_t удаляется индекс v, а из $\mathfrak{I}(a_t)$ удаляется вся тройка $\langle X_{tv}, X'_{tv}, c_{tv}\rangle$.

Леммы 4.19, 4.20 и теорема 4.7 позволяют рекуррентно вычислять вероятность переобучения Q_ε. На каждом d-м шаге, $d = 0, \ldots, D$, добавляется алгоритм a_d, вычисляется информация $\mathfrak{I}(a_d)$; затем для каждого $t = 0, \ldots, d-1$ корректируется информация $\mathfrak{I}(a_t)$ и вероятности P_{tv}. На основе скорректированной информации обновляется текущая оценка Q_ε. По окончании последнего D-го шага текущая оценка Q_ε совпадает с точным значением вероятности переобучения. Процедура вычисления точной оценки Q_ε описана с помощью псевдокода, облегчающего программную реализацию.

Вычисления могут оказаться неэффективными по времени, если условие 2) леммы 4.20 будет выполняться слишком часто. Каждое его выполнение приводит к добавлению ещё одного слагаемого в сумму (20), причём к новым слагаемым в свою очередь может применяться условие 2), в результате объём вычислений может расти экспоненциально по D. В работе доказывается, что если в описанной вычислительной процедуре в определённые моменты пропускать проверку условия 2), то можно получать верхние или нижние оценки вероятности переобучения, существенно сокращая объём вычислений. Таким образом, время вычислений удаётся обменивать на точность оценки.

Доказывается, что если проверка условия 2) пропускается всегда, то значение Q_ε, вычисляемое такой максимально упрощённой процедурой, будет верхней оценкой вероятности пере-

обучения. В следующей теореме эта оценка выписывается в явном виде через профиль расслоения и связности множества A.

Связностью $q(a)$ алгоритма $a \in A$ назовём число алгоритмов в следующем слое, допускающих ошибки на тех же объектах, что и a: $q(a) = \#\bigl\{a' \in A_{n(a,\mathbb{X})+1} \colon I(a,x) \leqslant I(a',x),\ x \in \mathbb{X}\bigr\}$.

Определение 4.12. *Профилем расслоения и связности* множества A называется матрица $\bigl(\Delta_{mq}\bigr)_{m=0,L}^{q=0,L}$, где Δ_{mq} — число алгоритмов в m-м слое со связностью q.

Теорема 4.22. Пусть векторы ошибок всех алгоритмов множества A попарно различны, в A есть корректный на $\mathbb{X}$ алгоритм. Тогда справедлива верхняя оценка

$$Q_\varepsilon \leqslant \sum_{m=\lceil \varepsilon k \rceil}^{L} \sum_{q=0}^{L} \Delta_{mq} \frac{C_{L-m-q}^{\ell-q}}{C_L^\ell}. \tag{23}$$

Согласно оценке (23) наибольший вклад в вероятность переобучения вносят алгоритмы с малым числом ошибок, начиная с $m = \lceil \varepsilon k \rceil$. По мере увеличения m комбинаторный множитель $C_{L-m-q}^{\ell-q}/C_L^\ell$ убывает экспоненциально. Увеличение связности q улучшает оценку. Есть основания полагать, что среднее значение связности q определяется размерностью пространства. В общем случае при увеличении размерности пространства возникают два противонаправленных эффекта: с одной стороны, увеличивается число алгоритмов в каждом слое, что приводит к росту Q_ε; с другой стороны, увеличивается связность q, что приводит к уменьшению Q_ε. Второй эффект совершенно не учитывается в классической VC-теории.

Предварительные эксперименты с линейными классификаторами и методом ближайших соседей показали, что профиль расслоения и связности Δ_{mq} с высокой точностью является сепарабельным: $\Delta_{mq} \lessapprox \Delta_m \lambda_q$, где Δ_m — коэффициент разнооб-

разия m-го слоя, λ_q — доля алгоритмов m-го слоя, имеющих связность q. Вектор $(\Delta_m)_{m=0,L}$ предлагается называть *профилем расслоения*, а вектор $(\lambda_q)_{q=0,L}$ — *профилем связности* множества алгоритмов A. В этих терминах немного ослабленная оценка (23) принимает следующий вид:

$$Q_\varepsilon \leqslant \underbrace{\sum_{m=\lceil \varepsilon k \rceil}^{k} \Delta_m \frac{C_{L-m}^{\ell}}{C_L^{\ell}}}_{\text{VC-оценка}} \underbrace{\sum_{q=0}^{L} \lambda_q \left(\frac{\ell}{L-m} \right)^q}_{\text{поправка на связность}}. \tag{24}$$

Первая часть (24) совпадает с VC-оценкой (14). Вторая часть представляет собой «поправку на связность», экспоненциально убывающую с ростом q, что и делает данную оценку существенно более точной, чем классические VC-оценки (11), (14).

Если профиль Δ_{mq} представлен в виде разложения $\Delta_m \lambda_q$, то вычисления Q_ε занимают $O(L)$ операций, что вполне приемлемо для практического применения.

Глава 5. Комбинаторные оценки полного скользящего контроля

§5.1. Функционал *полного скользящего контроля* (CCV) определяется как средняя частота ошибок на контроле по всевозможным разбиениям генеральной выборки:

$$\mathrm{CCV}(\mu, \mathbb{X}) = \frac{1}{C_L^{\ell}} \sum_{(X,\bar{X})} \nu(\mu X, \bar{X}).$$

В терминах слабой аксиоматики CCV есть математическое ожидание частоты ошибок на контроле, $\mathrm{CCV} = \mathsf{E}\nu(\mu X, \bar{X})$. Недостаток CCV в том, что он ничего не говорит о возможном разбросе (дисперсии) частоты ошибок $\nu(\mu X, \bar{X})$. Поэтому его нельзя использовать для получения верхних оценок.

§5.2. Рассматриваются задачи классификации с множеством классов $\mathbb{Y}$. Индикатор ошибки имеет вид $I(a,x) = \big[y(x) \neq a(x)\big]$, где $y\colon \mathbb{X} \to \mathbb{Y}$ — восстанавливаемая зависимость, $a\colon \mathbb{X} \to \mathbb{Y}$ — алгоритм классификации. Решение задач классификации часто основано на *гипотезе компактности* — эмпирическом предположении, что классы образуют локализованные «компактные» подмножества объектов, и схожие объекты, как правило, лежат в одном классе. Простейшим методом обучения, построенным на его основе, является метод ближайших соседей.

§5.2.1. Пусть на множестве $\mathbb{X}$ определена функция расстояния $\rho(x,x')$. *Метод ближайшего соседа* — это метод обучения μ, который запоминает обучающую выборку $X \subset \mathbb{X}$ и строит алгоритм $a = \mu X$, работающий следующим образом:

$$a(x;X) = y\big(\arg\min_{x' \in X} \rho(x,x')\big) \text{ для всех } x \in \mathbb{X}.$$

Для каждого объекта x_i, $i = 1,\ldots,L$ выборки $\mathbb{X}$ расположим остальные $L-1$ объектов в порядке возрастания расстояния до x_i, пронумеровав их двойными индексами: $x_i = x_{i0}$, $x_{i1}, x_{i2}, \ldots, x_{i,L-1}$. Таким образом,

$$0 = \rho(x_i, x_{i0}) \leqslant \rho(x_i, x_{i1}) \leqslant \cdots \leqslant \rho(x_i, x_{i,L-1}).$$

Обозначим через $r_m(x_i)$ ошибку, возникающую при замене правильного ответа $y(x_i)$ ответом на m-ом соседе объекта x_i:

$$r_m(x_i) = \big[y(x_i) \neq y(x_{im})\big]; \quad i = 1,\ldots,L;\ m = 1,\ldots,L-1.$$

Определение 5.1. *Профилем компактности* выборки $\mathbb{X}$ называется доля объектов выборки, для которых правильный ответ не совпадает с правильным ответом на m-ом соседе:

$$K(m,\mathbb{X}) = \frac{1}{L}\sum_{i=1}^{L} r_m(x_i); \quad m = 1,\ldots,L-1.$$

Профиль компактности является формальным выражением гипотезы компактности. Чем проще задача, тем чаще близкие объекты лежат в одном классе, тем сильнее «прижимается к нулю» начальный участок профиля. И наоборот, в трудных задачах, где ближайшие объекты практически не несут информации о классе, профиль вырождается в константу, близкую к 0.5.

§5.2.2. Существует связь профиля компактности с качеством классификации методом ближайшего соседа.

Теорема 5.2. Для метода ближайшего соседа μ и любой $\mathbb{X}$

$$\mathrm{CCV}(\mu, \mathbb{X}) = \sum_{m=1}^{k} K(m, \mathbb{X}) \frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^{\ell}}. \tag{25}$$

Комбинаторный множитель $\gamma_m = C_{L-1-m}^{\ell-1} / C_{L-1}^{\ell}$ убывает с ростом m быстрее геометрической прогрессии. Для обеспечения малого значения функционала CCV достаточно потребовать, чтобы профиль $K(m)$ принимал малые значения при малых m.

§5.2.3. В методе ближайшего соседа имеет смысл запоминать не всю обучающую выборку, а только подмножество наиболее типичных объектов — *эталонов*. Отбор эталонов обычно преследует несколько целей: сокращение объёма хранимых данных, повышение скорости классификации, повышение качества классификации за счёт удаления шумовых объектов.

Предлагается итерационный процесс минимизации CCV путём последовательного отсева неэталонных объектов. Эксперименты на модельных данных показывают, что в этом процессе сначала удаляются все шумовые выбросы, при этом функционал CCV убывает. Затем удаляются неинформативные «внутренние» объекты классов, при этом функционал либо не изменяется, либо увеличивается на ничтожно малую величину. Когда функционал начинает заметно возрастать, процесс уда-

ления объектов останавливается, и все оставшиеся объекты принимаются за эталоны. Эксперименты показывают, что данный метод вообще не переобучается. Важным для приложений побочным результатом является разделение всех объектов на три категории: эталонные, неинформативные и шумовые.

§5.3. Рассматриваются задачи классификации, в которых множество X частично упорядочено, $\mathbb{Y} = \{0, 1\}$, индикатор ошибки имеет вид $I(a, x) = [y(x) \neq a(x)]$, и имеется априорная информация о монотонности или почти-монотонности целевой зависимости $y(x)$. Ограничения монотонности могут возникать в различных контекстах. Во-первых, они могут быть результатом формализации экспертных знаний вида «чем больше значение признака $f(x)$ тем больше значение $y(x)$». Во-вторых, они возникают в алгоритмических композициях с нелинейной корректирующей операцией, которые являются естественным обобщением линейных выпуклых композиций (взвешенного голосования) алгоритмов классификации.

Допустим, что метод обучения μ выбирает алгоритмы из множества A всех монотонных отображений $F\colon \mathbb{X} \to \mathbb{Y}$.

Степенью немонотонности выборки $\mathbb{X}$ называется наименьшая частота ошибок, допускаемых на ней монотонными алгоритмами: $\theta(\mathbb{X}) = \min\limits_{a \in A} \nu(a, \mathbb{X})$.

Выборка $\mathbb{X}$ называется монотонной, если из $x_i \leqslant x_j$ следует $y(x_i) \leqslant y(x_j)$ для всех $i, j = 1, \ldots, L$. Выборка монотонна тогда и только тогда, когда $\theta(\mathbb{X}) = 0$.

Верхним и нижним клином объекта $x_i \in \mathbb{X}$ называются, соответственно, множества

$$W_0(x_i) = \{x \in \mathbb{X}\colon x_i < x \text{ и } y(x) = 0\};$$
$$W_1(x_i) = \{x \in \mathbb{X}\colon x < x_i \text{ и } y(x) = 1\}.$$

Введём сокращённое обозначение $W_i = W_{y(x_i)}(x_i)$.

Мощность клина $w_i = |W_i|$ характеризует глубину погружения объекта x_i в тот класс, которому он принадлежит. Чем меньше w_i, тем ближе объект к границе класса. Объекты, не имеющие своего клина ($w_i = 0$) будем называть *граничными*.

Определение 5.5. *Профилем монотонности* выборки $\mathbb{X}$ называется функция $M(m, \mathbb{X})$, выражающая долю объектов выборки с клином мощности m:

$$M(m, \mathbb{X}) = \frac{1}{L} \sum_{i=1}^{L} [w_i = m]; \quad m = 0, \ldots, L-1.$$

Теорема 5.4. Если метод μ минимизирует эмпирический риск в классе всех монотонных функций и степень немонотонности выборки $\mathbb{X}$ равна θ, то

$$\mathrm{CCV}(\mu, \mathbb{X}) \leqslant \sum_{m=0}^{\theta L+k-1} M(m, \mathbb{X}) \sum_{s=\max\{0, m-k+1\}}^{\min\{\theta L, \ell, m\}} \frac{C_m^s C_{L-1-m}^{\ell-s}}{C_{L-1}^{\ell}}. \quad (26)$$

Оценка (26) монотонно не убывает по θ, достигая наименьшего значения при $\theta = 0$, когда выборка монотонна и метод μ является корректным на генеральной выборке $\mathbb{X}$:

$$\mathrm{CCV}(\mu, \mathbb{X}) \leqslant \sum_{m=0}^{k-1} M(m, \mathbb{X}) \frac{C_{L-1-m}^{\ell}}{C_{L-1}^{\ell}}. \quad (27)$$

Оценка (26), в отличие от завышенных сложностных оценок, никогда не превышает 1. Наибольшее значение 1 достигается при $w_i = 0,\ i = 1, \ldots, L$. Это тот случай, когда оба класса состоят из попарно несравнимых объектов, и вся выборка распадается на две антицепи. Наименьшее значение достигается, когда выборка монотонна и линейно упорядочена. Тогда $\mathrm{CCV} \leqslant 2/\ell$.

Комбинаторный множитель в (26) убывает с ростом m быстрее геометрической прогрессии. Чтобы обеспечить малое значение функционала CCV, достаточно потребовать, чтобы функция $M(m, \mathbb{X})$ принимала малые значения при малых m. При больших m её рост компенсируется комбинаторным множителем. Таким образом, качество монотонного классификатора тем выше, чем меньше объектов имеют клинья небольшой мощности. Для этого отношение порядка на множестве объектов X должно быть близко к линейному вблизи границы классов.

Интересно отметить структурное сходство оценок (25) и (27), полученных для таких различных, на первый взгляд, априорных ограничений, как компактность и монотонность.

Публикации по теме диссертации

[1] *Воронцов К. В.* Качество восстановления зависимостей по эмпирическим данным // Математические методы распознавания образов: 7-ая Всерос. конф. Тезисы докл. — Пущино, 1995. — С. 24–26.

[2] *Рудаков К. В., Воронцов К. В.* О методах оптимизации и монотонной коррекции в алгебраическом подходе к проблеме распознавания // **Доклады РАН**. — 1999. — Т. 367, № 3. — С. 314–317.

[3] *Воронцов К. В.* Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания // **ЖВМ и МФ**. — 2000. — Т. 40, № 1. — С. 166–176.

[4] *Воронцов К. В.* Оценка качества монотонного решающего правила вне обучающей выборки // Интеллектуализация обработки информации: Тез. докл. — Симферополь, 2002. — С. 24–26.

[5] *Воронцов К. В.* О комбинаторном подходе к оценке качества обучения алгоритмов // Математические методы распознавания образов: 11-ая Всерос. конф. Тезисы докл. — Пущино, 2003. — С. 47–49.

[6] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // **Математические вопросы кибернетики** / Под ред. О. Б. Лупанов. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

[7] *Воронцов К. В.* Комбинаторные обоснования обучаемых алгоритмов // **ЖВМиМФ**. — 2004. — Т. 44, № 11. — С. 2099–2112.

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

[9] *Воронцов К. В.* Обзор современных исследований по проблеме качества обучения алгоритмов // Таврический вестник информатики и математики. — 2004. — № 1. — С. 5–24.

[10] *Воронцов К. В.* Комбинаторный подход к повышению качества логических классификаторов // Интеллектуализация обработки информации: Тезисы докл. — Симферополь, 2004. — С. 44.

[11] *Кочедыков Д. А., Ивахненко А. А., Воронцов К. В.* Система кредитного скоринга на основе логических алгоритмов классификации // Докл. всеросс. конф. Математические методы распознавания образов-12. — М.: МАКС Пресс, 2005. — С. 349–353.

[12] *Воронцов К. В., Ивахненко А. А.* Эмпирические оценки локальной функции роста в задачах поиска логических закономерностей // Искусственный Интеллект. — 2006. — С. 281–284.

[13] *Воронцов К. В., Колосков А. О.* Профили компактности и выделение опорных объектов в метрических алгоритмах классификации // Искусственный Интеллект. — 2006. — С. 30–33.

[14] *Воронцов К. В.* Слабая вероятностная аксиоматика и надёжность эмпирических предсказаний // Докл. всеросс. конф. Ма-

тематические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 21–25.

[15] *Ивахненко А. А., Воронцов К. В.* Верхние оценки переобученности и профили разнообразия логических закономерностей // Докл. всеросс. конф. Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 33–37.

[16] *Кочедыков Д. А., Ивахненко А. А., Воронцов К. В.* Применение логических алгоритмов классификации в задачах кредитного скоринга и управления риском кредитного портфеля банка // Докл. всеросс. конф. Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 484–488.

[17] *Венжега А. В., Ументаев С. А., Орлов А. А., Воронцов К. В.* Проблема переобучения при отборе признаков в линейной регрессии с фиксированными коэффициентами // Докл. всеросс. конф. Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 90–93.

[18] *Ульянов Ф. М., Воронцов К. В.* Проблема переобучения функций близости при построении алгоритмов вычисления оценок // Докл. всеросс. конф. Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 105–108.

[19] *Воронцов К. В., Инякин А. С., Лисица А. В.* Система эмпирического измерения качества алгоритмов классификации // Докл. всеросс. конф. Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 577–580.

[20] *Цюрмасто П. А., Воронцов К. В.* Анализ сходства алгоритмов классификации в оценках обобщающей способности // Интеллектуализация обработки информации (ИОИ-2008): Тез. докл. — Симферополь: КНЦ НАН Украины, 2008. — С. 232–234.

[21] *Vorontsov K. V.* On the influence of similarity of classifiers on the probability of overfitting // Pattern Recognition and Image Analysis: new information technologies (PRIA-9). — Vol. 2. —

Nizhni Novgorod, Russian Federation, 2008. — Pp. 303–306.

[22] *Vorontsov K. V.* Combinatorial probability and the tightness of generalization bounds // **Pattern Recognition and Image Analysis**. — 2008. — Vol. 18, no. 2. — Pp. 243–259.

[23] *Vorontsov K. V.* Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting // **Pattern Recognition and Image Analysis**. — 2009. — Vol. 19, no. 3. — Pp. 412–420.

[24] *Воронцов К. В.* Точные оценки вероятности переобучения // **Доклады РАН**. — 2009. — Т. 429, № 1. — С. 15–18.

[25] *Воронцов К. В.* Методы машинного обучения, основанные на индукции правил // Труды семинара «Знания и онтологии ELSEWHERE 2009», ассоциированного с 17-й международной конференцией по понятийным структурам ICCS-17, Москва, 21–26 июля. — Высшая школа экономики, 2009. — С. 57–71.

[26] *Воронцов К. В.* Комбинаторный подход к проблеме переобучения // Докл. всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 18–21.

[27] *Иванов М. Н., Воронцов К. В.* Отбор эталонов, основанный на минимизации функционала полного скользящего контроля // Докл. всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 119–122.

[28] *Воронцов К. В., Ивахненко А. А., Инякин А. С., Лисица А. В., Минаев П. Ю.* «Полигон» — распределённая система для эмпирического анализа задач и алгоритмов классификации // Докл. всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 503–506.