На правах рукописи

ФРЕЙ АЛЕКСАНДР ИЛЬИЧ

ТЕОРЕТИКО-ГРУППОВОЙ ПОДХОД В КОМБИНАТОРНОЙ ТЕОРИИ ПЕРЕОБУЧЕНИЯ

05.13.17 — теоретические основы информатики

Автореферат диссертации на соискание ученой степени кандидата физико-математических наук

Москва — 2013

Работа выполнена на кафедре «Интеллектуальные системы» факультета управления и прикладной математики Федерального государственного образовательного учреждения высшего профессионального образования «Московский физико–технический институт (государственный университет)».

Научный руководитель: доктор физико-математических наук, профессор **Воронцов Константин Вячеславович**.

Официальные оппоненты:
доктор физико-математических наук, **Хачай Михаил Юрьевич**, Федеральное государственное бюджетное учреждение науки Институт математики и механики им. Н. Н. Красовского Уральского отделения Российской академии наук, зав. отделом математического программирования;
кандидат технических наук, **Игнатов Дмитрий Игоревич**, Федеральное государственное автономное образовательное учреждение высшего профессионального образования Национальный исследовательский университет «Высшая школа экономики», доцент.

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А. А. Харкевича Российской академии наук.

Защита диссертации состоится 19 декабря 2013 г. в 14:30 на заседании диссертационного совета Д 002.017.02 при Федеральном государственном бюджетном учреждении «Вычислительный центр им. А. А. Дородницына Российской академии наук», расположенном по адресу: 119333, г. Москва, ул. Вавилова, д. 40.
С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан 11 ноября 2013 г.

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

Общая характеристика работы

Диссертационная работа посвящена проблеме повышения точности комбинаторных оценок вероятности переобучения.

Актуальность темы. При решении задач обучения по прецедентам, восстановления зависимостей по эмпирическим данным, классификации, распознавания образов, прогнозирования часто возникает проблема переобучения. Она состоит в том, что решающая функция (алгоритм), построенная по конечной обучающей выборке, может допускать ошибки на объектах контрольной выборки существенно чаще, чем на объектах обучающей выборки. Для контроля переобучения на этапе построения алгоритма необходимо иметь оценки вероятности переобучения. Такие оценки известны в статистической теории обучения, однако они либо сильно завышены, либо имеют слишком узкую область применимости.

Степень разработанности темы. Основы статистической теории обучения были заложены в работах В. Н. Вапника и А. Я. Червоненкиса в конце 60-х годов. Ими была доказана состоятельность обучения по прецедентам и получены количественные оценки, связывающие обобщающую способность метода обучения с длиной обучающей выборки и сложностью семейства алгоритмов. Основной проблемой этих оценок является их завышенность. Для устранения завышенности предлагалось строить оценки, зависящие от выборки (D. Haussler, 1992); учитывать ширину зазора, разделяющего классы (P. Bartlett, 1998); строить оценки на основе локальной рademахеровской сложности семейства алгоритмов (V. Koltchinskii, 1998); учитывать априорные распределения на множестве алгоритмов (L. Valiant, 1982; D. McAllester, 1999; J. Langford, 2005); а также ряд других подходов.

Комбинаторная теория переобучения показала, что для повышения точности оценок и сокращения переобучения необходимо одновременно учитывать эффекты расслоения и сходства в се-

мействах алгоритмов (К. В. Воронцов, 2010). Была получена оценка расслоения-связности, справедливая для широкого класса семейств, представимых в виде связного графа (К. В. Воронцов, А. А. Ивахненко, И. М. Решетняк, 2010). Для некоторых модельных частных случаев было показано, что этого достаточно для получения неулучшаемых (точных) оценок. Таким образом, комбинаторная теория переобучения является новым перспективным подходом. Данная работа направлена на ее дальнейшее развитие: расширение границ применимости, разработку новых методов вывода оценок обобщающей способности и повышение точности этих оценок.

Цели и задачи работы: повышение точности комбинаторных оценок вероятности переобучения; переход от требования связности к более слабому требованию сходства алгоритмов; разработка новых методов получения оценок обобщающей способности, применимых к несвязным семействам алгоритмов высокой мощности.

Научная новизна. Впервые получены неулучшаемые оценки вероятности переобучения для рандомизированного метода минимизации эмпирического риска. Для их получения разработан новый теоретико-групповой подход, основанный на учете симметрий множества алгоритмов. С его помощью получены неулучшаемые оценки вероятности переобучения для девяти модельных семейств алгоритмов. Получена комбинаторная оценка вероятности переобучения, основанная на разложении множества алгоритмов на непересекающиеся подмножества (кластеры). Каждый кластер пополняется алгоритмами до объемлющего множества алгоритмов с известной точной оценкой вероятности переобучения. Итоговая оценка учитывает сходство алгоритмов внутри каждого кластера и расслоение алгоритмов по числу ошибок между разными кластерами. Данная оценка применима к широкому классу семейств, в том числе и к семействам, не обладающим свойством связности.

Теоретическая и практическая значимость. Данная работа вносит существенный вклад в развитие комбинаторной теории

переобучения и расширяет границы ее применимости на несвязные семейства алгоритмов высокой мощности.

Методы исследования. Для получения оценок вероятности переобучения использована слабая (перестановочная) вероятностная аксиоматика, комбинаторная теория переобучения, элементы комбинаторки, теории групп, теории вероятностей и теории графов. Для проверки точности комбинаторных оценок проведены вычислительные эксперименты на модельных данных и задачах из репозитория UCI.

Положения, выносимые на защиту.

1. Теоретико-групповой метод орбит, позволяющий выводить оценки вероятности переобучения для симметричных семейств алгоритмов и рандомизированного метода минимизации эмпирического риска.

2. Точные оценки вероятности переобучения рандомизированного метода минимизации эмпирического риска для модельных семейств: монотонной и унимодальной сетей, слоя хэммингова шара и ряда других.

3. Общая оценка вероятности переобучения, основанная на разложении и покрытии множества алгоритмов.

4. Экспериментальное подтверждение того, что новая оценка в некоторых случаях менее завышена по сравнению с другими комбинаторными оценками вероятности переобучения.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных оценок переобучения на реальных задачах классификации; публикациями результатов исследования в рецензируемых научных изданиях, в том числе рекомендованных ВАК РФ. Результаты работы докладывались, обсуждались и получили одобрение специалистов на следующих научных конференциях и семинарах:

- Всероссийская конференция «Математические методы распознавания образов» ММРО-14, 2009 г. [1];
- 52-я научная конференция МФТИ, 2009 г. [2];
- Международная конференция «Интеллектуализация обработки информации» ИОИ-8, 2010 г. [3];
- Всероссийская конференция «Математические методы распознавания образов» ММРО-15, 2011 г. [4];
- Международная конференция «25th European Conference on Operational Research», 2012 г.
- Международная конференция «Интеллектуализация обработки информации» ИОИ-9, 2012 г. [5];
- Научные семинары отдела Интеллектуальных систем Вычислительного центра РАН и кафедры «Интеллектуальные системы» МФТИ, 2010 – 2013 г.г.

Публикации по теме диссертации в изданиях из списка ВАК РФ — одна [7]. Другие публикации по теме диссертации: [1, 2, 3, 4, 5, 8, 6].

Структура и объем работы. Работа состоит из введения, пяти глав, заключения, списка использованных источников, включающего 78 наименований. Общий объем работы составляет 102 страницы.

Краткое содержание работы по главам

Глава 1. Теория статистического обучения

Глава содержит краткий обзор современного состояния теории статистического обучения (statistical learning theory). Обсуждается проблема завышенности классических оценок переобучения, приводятся мотивации перехода от классических теоретико-вероятностных постановок задач к комбинаторным.

Глава 2. Комбинаторный подход

2.1. Основные определения. Пусть задана конечная генеральная выборка $\mathbb{X} = \{x_1, \ldots, x_L\}$, состоящая из L объектов. Пусть A — некоторое множество алгоритмов. Каждый алгоритм классификации $a \in A$, примененный к выборке $\mathbb{X}$, порождает бинарный вектор ошибок $a \equiv \big(I(a, x_i)\big)_{i=1}^{L}$, где $I(a, x_i) \in \{0, 1\}$ — индикатор ошибки алгоритма a на объекте x_i. Для произвольной подвыборки $X \subset \mathbb{X}$ обозначим через $n(a, X) = \sum\limits_{x \in X} I(a, x)$ число ошибок, а через $\nu(a, X) = n(a, X)/|X|$ — частоту ошибок алгоритма a на выборке X.

Методом обучения называют отображение вида $\mu\colon 2^{\mathbb{X}} \to A$. Результатом обучения по выборке $X \subset \mathbb{X}$ называется алгоритм $a = \mu(X)$, обозначаемый также $a = \mu X$ или $a = \mu(A, X)$.

Метод обучения μ называется *минимизацией эмпирического риска* (МЭР), если

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

и методом *пессимистической минимизации эмпирического риска* (ПМЭР), если

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

Обозначим через $[\mathbb{X}]^{\ell}$ множество всех разбиений генеральной выборки $\mathbb{X}$ на обучающую выборку X длины ℓ и контрольную выборку $\bar{X}$ длины $k = L - \ell$. Если для выборки $X \in [\mathbb{X}]^{\ell}$ *уклонение частот* $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$ превосходит фиксированный порог $\varepsilon > 0$, то говорят, что алгоритм $a = \mu X$ является *переобученным*. Нашей целью является получение оценок *вероятности переобучения*:

$$Q_{\varepsilon}(\mu, \mathbb{X}) = \mathsf{P}\big[\delta(\mu X, X) \geqslant \varepsilon\big] \leqslant \eta(\varepsilon), \text{ где } \mathsf{P} \equiv \frac{1}{C_L^{\ell}} \sum_{X \in [\mathbb{X}]^{\ell}}, \qquad (2.1)$$

где квадратные скобки означают [истина] = 1, [ложь] = 0.

Коль скоро такая оценка получена, можно утверждать, что $\nu(\mu X, \bar{X}) \leqslant \nu(\mu X, X) + \varepsilon(\eta)$ с вероятностью $1 - \eta$, достаточно близкой к единице, где $\varepsilon(\eta)$ — функция, обратная к $\eta(\varepsilon)$.

2.2. Расслоение и связность. Определим естественное отношение порядка на алгоритмах: $a \leqslant b$ тогда и только тогда, когда $I(a,x) \leqslant I(b,x)$ для всех $x \in \mathbb{X}$. Определим метрику на алгоритмах как хэммингово расстояние между их векторами ошибок: $\rho(a,b) = \sum\limits_{i=1}^{L} |I(a,x) - I(b,x)|$. Определим отношение предшествования $a \prec b$ если $a < b$ и $\rho(a,b) = 1$.

Для каждого $a \in A$ рассмотрим *порождающее* и *запрещающее* множества:

$$X_a = \big\{x \in \mathbb{X} \;\big|\; \exists b \in A\colon I(a,x) < I(b,x),\ a \prec b\big\},$$
$$X'_a = \big\{x \in \mathbb{X} \;\big|\; \exists b \in A\colon I(b,x) < I(a,x),\ b \leqslant a\big\}.$$

Величину $u(a) = |X_a|$ называют *верхней связностью*, а величину $q(a) = |X'_a|$ — *неполноценностью* алгоритма a.

Теорема 2.1 (Воронцов, Ивахненко, Решетняк, 2010). *Если μ — метод пессимистичной минимизации эмпирического риска, то для вероятности переобучения справедлива оценка*

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant \sum_{a \in A} \frac{C^{\ell-u(a)}_{L-u(a)-q(a)}}{C^\ell_L} H^{\ell-u(a),\, n(a,\mathbb{X})-q(a)}_{L-u(a)-q(a)}\left(\tfrac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big)\right),$$

где $H^{\ell,m}_L(s) = \sum\limits_{i=0}^{\lfloor s \rfloor} \frac{C^i_m C^{\ell-i}_{L-m}}{C^\ell_L}$ — функция гипергеометрического распределения.

2.3. Постановка задачи. Приводятся результаты двух экспериментов, показывающих, в каких случаях оценка теоремы 2.1 сильно завышена. Эта оценка не учитывает сходство между алгоритмами с равным числом ошибок. Поэтому первая задача,

которая ставится в данной работе — получить оценки, одновременно учитывающие и расслоение алгоритмов по числу ошибок, и сходство алгоритмов внутри одного слоя. Вторая задача — разработать удобный математический инструментарий, позволяющий получать точные и вычислительно эффективные оценки вероятности переобучения для семейств, состоящих из большого числа алгоритмов.

Глава 3. Теоретико-групповой подход

3.1. Рандомизированный метод обучения и РМЭР. При минимизации эмпирического риска может возникать неоднозначность — несколько алгоритмов из $A(X) \equiv \operatorname{Arg}\min_{a \in A} n(a, X)$ могут иметь одинаковое число ошибок на обучающей выборке. Следующий метод обучения естественным образом устраняет эту неоднозначность.

Определение 3.1. *Путь $\mathbb{A} = \{0,1\}^L$ — множество всех бинарных векторов ошибок. Рандомизированный метод обучения произвольному множеству алгоритмов $A \subseteq \mathbb{A}$ и произвольной обучающей выборке $X \in [\mathbb{X}]^\ell$ ставит в соответствие функцию распределения весов на множестве алгоритмов:*

$$\mu : 2^{\mathbb{A}} \times [\mathbb{X}]^\ell \to \{f : \mathbb{A} \to [0,1]\}. \tag{3.1}$$

Естественно полагать, что эта функция нормирована и может быть интерпретирована как вероятность получить каждый из алгоритмов в результате обучения.

Методы обучения вида $\mu\colon 2^{\mathbb{X}} \to A$ далее будем называть *детерминированными.*

Рандомизированный метод минимизации эмпирического риска (РМЭР) выбирает произвольный алгоритм из множества $A(X)$ случайно и равновероятно:

$$\mu(A, X)(a) = \frac{[a \in A(X)]}{|A(X)|}.$$

Для РМЭР определение вероятности переобучения $Q_\varepsilon(A)$ соответствующим образом модифицируется:

$$Q_\varepsilon(A) = \mathsf{E}\frac{1}{|A(X)|}\sum_{a\in A(X)}\big[\delta(a,X)\geqslant\varepsilon\big],\ \text{где}\ \mathsf{E}\equiv\frac{1}{C_L^\ell}\sum_{X\in[\mathbb{X}]^\ell}. \quad (3.2)$$

Также возможна запись $Q_\varepsilon(A) = \sum\limits_{a\in A} Q_\varepsilon(a,A)$, где $Q_\varepsilon(a,A)$ — *вклад алгоритма* $a\in A$ в вероятность переобучения:

$$Q_\varepsilon(a,A) = \mathsf{E}\frac{\big[a\in A(X)\big]}{|A(X)|}\big[\delta(a,X)\geqslant\varepsilon\big].$$

3.2. Перестановки объектов. Пусть $S_L = \{\pi\colon \mathbb{X}\to\mathbb{X}\}$ — симметрическая группа из $L!$ элементов, действующая на генеральную выборку перестановками объектов. Действие произвольной перестановки $\pi\in S_L$ на алгоритм $a\in A$ определено перестановкой координат вектора ошибок: $(\pi a)(x_i) = a(\pi^{-1}x_i)$. Для произвольной выборки $X\in[\mathbb{X}]^\ell$ и множества алгоритмов $A\subset\{0,1\}^L$ действия πX и πA определены следующим образом: $\pi X = \{\pi x\colon x\in X\}$, $\pi A = \{\pi a\colon a\in A\}$.

Лемма 3.1. *Свойства действия произвольной перестановки* $\pi\in S_L$*:*

1) $I(\pi a,\pi x) = I(a,x)$ *для любых* $a\in A$ *и* $x\in\mathbb{X}$*;*

2) $n(\pi a,\mathbb{X}) = n(a,\mathbb{X})$ *для любого* $a\in A$*;*

3) $n(\pi a,\pi X) = n(a,X)$ *для любых* $a\in A$ *и* $X\subseteq\mathbb{X}$*;*

4) $\delta(\pi a,\pi X) = \delta(a,X)$ *для любых* $a\in A$ *и* $X\subseteq\mathbb{X}$*;*

5) $[a\in A(X)] = [\pi a\in(\pi A)(\pi X)]$ *для любых* $a\in A$ *и* $X\subseteq\mathbb{X}$*;*

6) $|A(X)| = |(\pi A)(\pi X)|$ *для любых* A *и* $X\subseteq\mathbb{X}$*;*

7) $\rho(a,a') = \rho(\pi a,\pi a')$ *для любых* $a,a'\in A$*.*

Для построения функций, инвариантных относительной действия S_L, введем следующую классификацию функций:

- Симметричной функцией *первого рода* будем называть $g\colon \mathbb{A} \times [\mathbb{X}]^\ell \to \mathbb{R}$, такую, что для всех $\pi \in S_L$ выполнено $g(a, X) = g(\pi a, \pi X)$;

- Симметричной функцией *второго рода* будем называть $G\colon 2^{\mathbb{A}} \times [\mathbb{X}]^\ell \to 2^{\mathbb{A}}$, такую, что для всех $\pi \in S_L$ выполнено $\pi G(A, X) = G(\pi A, \pi X)$;

- Симметричной функцией *третьего рода* будем называть $f\colon 2^{\mathbb{A}} \times [\mathbb{X}]^\ell \to \mathbb{R}$, такую, что для всех $\pi \in S_L$ выполнено $f(A, X) = f(\pi A, \pi X)$.

Лемма 3.1 утверждает, что функции $n(a, X)$ и $\nu(a, X)$ являются симметричными функциями первого рода, а $A(X)$, как функция A и X, является симметричной функцией второго рода.

Теорема 3.2. *Пусть $g_1, g_2, \ldots, g_p$ — симметричные функции первого рода, $f_1, f_2, \ldots, f_p$ — симметричные функции третьего рода, $F\colon \mathbb{R}^p \to \mathbb{R}$ — произвольная функция многих переменных. Тогда $F(g_1, g_2, \ldots, g_p)$ — вновь симметричная функция первого рода, $F(f_1, f_2, \ldots, f_p)$ — симметричная функция третьего рода.*

Теорема 3.3. *Пусть g — симметричная функция первого рода, G — симметричная функция второго рода. Тогда*

$$f(A, X) \equiv |G(A, X)| \text{ и } f(A, X) \equiv \sum_{a \in G(A,X)} g(a, X)$$

являются симметричными функциями третьего рода.

3.3. Группа симметрии множества алгоритмов.

Определение 3.2. *Группой симметрий* $\mathrm{Sym}(A)$ *множества алгоритмов A будем называть его стационарную подгруппу:*

$$\mathrm{Sym}(A) = \{\pi \in S_L\colon \pi A = A\}.$$

Пусть далее $G \subseteq \mathrm{Sym}(A)$ — произвольная подгруппа группы $\mathrm{Sym}(A)$. Для любой перестановки $\pi \in G$ и любого алгоритма

$a \in A$ алгоритм πa снова лежит в A. В таких случаях говорят, что группа G *действует* на множестве A.

Орбитой алгоритма $a \in A$ называется множество алгоритмов $Ga = \{\pi a \colon \pi \in G\}$. Множество A разбивается на непересекающиеся подмножества — орбиты:

$$A = \bigsqcup_{\omega \in \Omega(A)} \omega = \bigsqcup_{\omega \in \Omega(A)} Ga_\omega,$$

где $\Omega(A)$ — множество всех орбит в A, a_ω — произвольный представитель орбиты ω.

Лемма 3.4. *Алгоритмы из одной орбиты имеют равные вклады в вероятность переобучения:*

$$Q_\varepsilon(a, A) = Q_\varepsilon(\pi a, A) \quad \text{для всех } \pi \in G.$$

Теорема 3.5. *Для любой генеральной выборки $\mathbb{X}$, любого множества алгоритмов A с попарно различными векторами ошибок и любого $\varepsilon \in [0, 1]$ справедлива формула разложения вероятности переобучения по орбитам множества A:*

$$Q_\varepsilon(A) = \sum_{\omega \in \Omega(A)} |\omega| \, \mathsf{E} \frac{\big[a_\omega \in A(X)\big]}{|A(X)|} \big[\delta(a_\omega, X) \geqslant \varepsilon\big], \qquad (3.3)$$

где $\Omega(A)$ — множество всех орбит в A, a_ω — произвольный представитель орбиты ω.

По аналогии с действием группы $G \subset \mathrm{Sym}(A)$ на множестве алгоритмов рассматривается действие G на множестве $[\mathbb{X}]^\ell$ всех разбиений выборки на обучение и контроль.

Теорема 3.6 (Толстихин, Фрей, 2010). *В условиях теоремы 3.5 справедлива формула разложения $Q_\varepsilon(A)$ по орбитам множества $[\mathbb{X}]^\ell$:*

$$Q_\varepsilon(A) = \frac{1}{C_L^\ell} \sum_{\tau \in \Omega[\mathbb{X}]^\ell} \frac{|\tau|}{|A(X_\tau)|} \sum_{a \in A(X_\tau)} \big[\delta(a, X_\tau) \geqslant \varepsilon\big], \qquad (3.4)$$

где $\Omega[\mathbb{X}]^\ell$ *— множество всех орбит в* $[\mathbb{X}]^\ell$, X_τ *— произвольный представитель орбиты* τ.

Теоремы 3.5 и 3.6 являются основным инструментом для вывода оценок вероятности переобучения симметричных семейств алгоритмов. Оценки (3.3) и (3.4) являются точными равенствами и, следовательно, неулучшаемы.

3.4. Покрытия множества алгоритмов.

Лемма 3.7. *Пусть множество алгоритмов* A *представлено в виде разбиения* $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$. *Тогда для вероятности переобучения детерминированного метода ПМЭР справедлива верхняя оценка*

$$Q_\varepsilon(A) \leqslant \sum_{i=1}^{t} Q_\varepsilon(A_t). \tag{3.5}$$

Данная оценка аналогична неравенству Буля, широко используемому в статистической теории обучения. Преимущество данной леммы состоит в том, что вместо суммирования по всем алгоритмам $a \in A$ суммирование производится по подмножествам $A_1, \ldots, A_t$, называемых кластерами, что позволяет сократить число слагаемых и сделать оценку вычислительно эффективной.

Для вычисления $Q_\varepsilon(A_i)$ предлагается дополнить каждое множество A_i до множества $B_i \supset A_i$ с известной точной оценкой вероятности переобучения и применить неравенство[1]

$$Q_\varepsilon(A_i) \leqslant Q_\varepsilon(B_i),$$

справедливое для ПМЭР при условии равенства числа ошибок алгоритмов в множестве B_i.

3.5. Теоремы о порождающих и запрещающих множествах.

Метод порождающих и запрещающих множеств (ПЗМ), предложенный К. В. Воронцовым, основан на гипотезе, что для любого

[1] Толстихин И. О. Вероятность переобучения плотных и разреженных семейств алгоритмов // Межд. конф. Интеллектуализация обработки информации ИОИ-8. — М.: МАКС Пресс, 2010. — С. 83–86.

алгоритма $a \in A$ можно записать необходимое и достаточное условие того, что он будет выбран методом обучения μ по выборке X. В данном параграфе метод ПЗМ обобщается по двум направлениям: во-первых, на случай разложения множества алгоритмов на кластеры, во-вторых, на случай РМЭР.

Гипотеза 3.1. Пусть множество алгоритмов A представлено в виде разбиения на непересекающиеся подмножества $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$. Пусть выборка $\mathbb{X}$ и детерминированный метод обучения μ таковы, что для каждого $i = 1, \ldots, t$ можно указать пару непересекающихся подмножеств $X_i \subset \mathbb{X}$ и $X'_i \subset \mathbb{X}$, удовлетворяющую условию

$$[\mu(A, X) \in A_i] \leqslant [X_i \subset X][X'_i \subset \bar{X}] \text{ для всех } X \in [\mathbb{X}]^\ell.$$

Пусть, кроме этого, все алгоритмы $a \in A_i$ не допускают ошибок на X_i и ошибаются на всех объектах из X'_i.

Множество X_i будем называть *порождающим*, множество X'_i — *запрещающим* для A_i. Гипотеза 3.1 означает, что результат обучения может принадлежать A_i только в том случае, если в обучающей выборке X находятся все порождающие объекты и ни одного запрещающего. Все остальные объекты $\mathbb{Y}_i \equiv \mathbb{X} \backslash X_i \backslash X'_i$ будем называть *нейтральными* для A_i.

Пусть $L_i = L - |X_i| - |X'_i|$, $\ell_i = \ell - |X_i|$, $k_i = k - |X'_i|$. Обозначим через $Q'_\varepsilon(A_i, \mathbb{Y}_i)$ вероятность переобучения на множестве нейтральных объектов $\mathbb{Y}_i$:

$$Q'_\varepsilon(A_i, \mathbb{Y}_i) = \frac{1}{C^{\ell_i}_{L_i}} \sum_{Y \in [\mathbb{Y}_i]^{\ell_i}} [\max_{a \in A_i} \delta(a, Y) \geqslant \varepsilon],$$

где $[\mathbb{Y}_i]^{\ell_i}$ — множество разбиений $\mathbb{Y}_i$ на обучающую выборку Y длины ℓ_i и контрольную выборку $\bar{Y}$ длины $k_i = L_i - \ell_i$.

Теорема 3.8 (ПЗМ для кластеров). *Пусть выполнена гипотеза 3.1, а на разбиение $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ наложено дополнительное ограничение: внутри каждого кластера A_i все*

алгоритмы допускают равное число ошибок m_i. Тогда

$$Q_\varepsilon(A,\mathbb{X}) \leqslant \sum_{i=1}^{t} P_i\, Q'_{\varepsilon_i}(A_i,\mathbb{Y}_i), \tag{3.6}$$

где $P_i = C_{L_i}^{\ell_i}/C_L^\ell$, $\varepsilon_i = \frac{L_i}{\ell_i k_i}\frac{\ell k}{L}\varepsilon + \big(1 - \frac{\ell L_i}{L\ell_i}\big)\frac{m_i}{k_i} - \frac{|X'_i|}{k_i}$.

Лемма 3.9. *Пусть μ — детерминированный ПМЭР. Тогда множества*

$$X_i = \bigcap_{a\in A_i} \{x \in \mathbb{X}\colon \exists b \in A\colon a \prec b, I(a,x) < I(b,x)\},$$
$$X'_i = \bigcap_{a\in A_i} \{x \in \mathbb{X}\colon \exists b \in A\colon b < a, I(b,x) < I(a,x)\}$$

являются, соответственно, порождающим и запрещающим множествами для кластера A_i в смысле гипотезы 3.1.

Лемма 3.9 позволяет в явном виде построить множество порождающих и запрещающих объектов и применить теорему 3.8 к детерминированному ПМЭР. Аналогичный результат для РМЭР представлен гипотезой 3.2 и теоремой 3.10.

Гипотеза 3.2. Обозначим $\mathfrak{A}(A) = \big\{A(X)\colon X \in [\mathbb{X}]^\ell\big\}$. Пусть множество A и выборка $\mathbb{X}$ таковы, что для каждого $\alpha \in \mathfrak{A}(A)$ можно указать пару непересекающихся подмножеств $X_\alpha \subset \mathbb{X}$ и $X'_\alpha \subset \mathbb{X}$, удовлетворяющую условию

$$\big[A(X){=}\alpha\big] = \big[X_\alpha \subseteq X\big]\big[X'_\alpha \subseteq \bar{X}\big] \quad \text{для всех } X \in [\mathbb{X}]^\ell. \tag{3.7}$$

Теорема 3.10. *Если справедлива гипотеза 3.2, то вероятность переобучения рандомизированного метода минимизации эмпирического риска есть*

$$Q_\varepsilon(A) = \sum_{a\in A}\sum_{\alpha\in\mathfrak{A}(A)} \frac{[a\in\alpha]}{|\alpha|}\frac{C_{L_\alpha}^{\ell_\alpha}}{C_L^\ell} H_{L_\alpha}^{\ell_\alpha, m_\alpha^a}\big(s_\alpha^a(\varepsilon)\big),$$

где введены следующие обозначения:

$$L_\alpha = L - |X_\alpha| - |\bar{X}_\alpha|; \quad \ell_\alpha = \ell - |X_\alpha|;$$
$$m^a_\alpha = n(a, \mathbb{X}\backslash X_\alpha \backslash X'_\alpha); \quad s^a_\alpha(\varepsilon) = \tfrac{\ell}{L}\big(n(a,\mathbb{X}) - \varepsilon k\big) - n(a, X_\alpha).$$

Глава 4. Точные оценки вероятности переобучения для РМЭР

В данной главе получены точные оценки вероятности переобучения РМЭР для девяти модельных семейств алгоритмов:

- монотонная цепь;
- унимодальная цепь;
- пучок монотонных цепей;
- монотонная сеть;
- унимодальная сеть;
- разреженная монотонная сеть;
- разреженная унимодальная сеть;
- слой хэммингова шара;
- слой интервала булева куба.

Для вывода оценок используется разработанный выше математический инструментарий: разложение вероятности переобучения по орбитам действия группы симметрий на множестве алгоритмов или на множестве разбиений выборки, а также теорема о порождающих и запрещающих множеств для РМЭР.

Среди перечисленных семейств наибольшего внимания заслуживают следующие два.

Определение 4.1. *Центральным слоем шара радиуса r называют множество алгоритмов, заданное следующим условием:*

$$B_r(a_0) = \{a \in \mathbb{A}\colon n(a,\mathbb{X}) = n(a_0,\mathbb{X}) \text{ и } \rho(a,a_0) \leqslant r\},$$

где a_0 — фиксированный алгоритм, $\rho(a,a')$ — расстояние Хэмминга между векторами ошибок алгоритмов a, a'.

Центральный слой хэммингова шара является множеством из большого числа попарно близких алгоритмов. Это позволяет использовать данное множество в качестве объемлющего множества B_i в оценках вида $Q_\varepsilon(A_i) \leqslant Q_\varepsilon(B_i)$. Следующее множество также обладает указанным свойством, но кроме этого различает несколько классов объектов: надежно классифицируемые, ошибочно классифицируемые, и пограничные.

Определение 4.2. *Пусть все объекты генеральной выборки $\mathbb{X}$ разделены на три непересекающихся множества: надежно классифицируемые объекты X_0, ошибочно классифицируемые объекты X_1 и пограничные объекты X_r. Пусть $|X_r| = r$ и $|X_1| = m$, ρ — целочисленный параметр, $\rho \leqslant r$. Слоем интервала булева куба будем называть множество $\hat{B}^m_{r,\rho}$, удовлетворяющее следующим условиям:*

- *$\hat{B}^m_{r,\rho}$ содержит все алгоритмы, допускающие ровно ρ ошибок на объектах из X_r,*
- *ни один алгоритм из $\hat{B}^m_{r,\rho}$ не ошибается на объектах из X_0,*
- *все алгоритмы из $\hat{B}^m_{r,\rho}$ ошибаются на всех объектах из X_1.*

Теорема 4.1 (Толстихин, 2010). *Вероятность переобучения ПМЭР для центрального слоя шара $B_r(a_0)$:*

$$Q_\varepsilon(B_r(a_0)) = H_L^{\ell,m}\big(\tfrac{\ell}{L}(m - \varepsilon k) + \lfloor r/2 \rfloor\big) \cdot [m \geqslant \varepsilon k]. \qquad (4.1)$$

В настоящей работе приводится более простое доказательство Теоремы 4.1.

Теорема 4.2. *Вероятность переобучения ПМЭР для слоя интервала булева куба $\hat{B}^m_{r,\rho}$:*

$$Q_\varepsilon(\hat{B}^m_{r,\rho}) = \frac{1}{C_L^\ell} \sum_{i=0}^{\min(m,\ell)} \sum_{j=0}^{\min(r,\ell-i)} C_m^i C_r^j C_{L-m-r}^{\ell-i-j} \Big[\delta(i,j) \geqslant \varepsilon\Big], \quad (4.2)$$

где $t(i,j) = i + \max(0, \rho - r - j)$ и $\delta(i,j) = \frac{m+\rho-t(i,j)}{k} - \frac{t(i,j)}{\ell}$.

Глава 5. Вычислительные эксперименты на реальных данных

В данной главе описываются вычислительные эксперименты на реальных данных из репозитория UCI. В экспериментах сравнивается завышенность уже известных комбинаторных оценок вероятности переобучения и новой оценки (3.6), в которой вероятность переобучения каждого кластера оценивается сверху с помощью (4.2). Показывается, что построенная таким образом оценка в ряде случаев оказывается менее завышенной по сравнению с другими комбинаторными оценками. Кроме того, новая оценка эффективно вычислима для семейств с существенно большим числом алгоритмов, т.к. сумма (3.6) содержит меньшее число слагаемых из-за кластеризации алгоритмов с близкими векторами ошибок.

Заключение

Основные результаты диссертационной работы.

1. Предложен теоретико-групповой метод вывода оценок вероятности переобучения для рандомизированного метода минимизации эмпирического риска.

2. Доказаны точные оценки вероятности переобучения рандомизированного метода минимизации эмпирического риска для девяти модельных семейств, включая монотонные и унимодальные сети, слой хэммингова шара и ряд других.

3. Получена общая оценка вероятности переобучения, основанная на разложении и покрытии множества алгоритмов.

4. В экспериментах на реальных данных показано, что новая оценка вероятности переобучения является более точной по сравнению с другими оценками вероятности переобучения.

Возможным направлением дальнейших исследований является повышение точности комбинаторных оценок вероятности переобучения путем более аккуратного учета эффекта расслоения, а также применение полученных в данной работе оценок для улучшения обобщающей способности логических алгоритмов классификации.

Публикации по теме диссертации

1. Фрей А. И. Точные оценки вероятности переобучения для симметричных семейств алгоритмов // Всеросс. конф. Математические методы распознавания образов-14. — М.: МАКС Пресс, 2009. — С. 66–69.
2. Фрей А. И. Точные оценки вероятности переобучения для симметричных семейств алгоритмов // Труды 52-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук». Часть II. Управление и прикладная математика. — М.: МФТИ, 2009. — С. 106–109.
3. Фрей А. И. Вероятность переобучения плотных и разреженных многомерных сеток алгоритмов // Межд. конф. Интеллектуализация обработки информации ИОИ-8. — М.: МАКС Пресс, 2010. — С. 87–90.
4. Фрей А. И. Метод порождающих и запрещающих множеств для рандомизированного метода минимизации эмпирического риска // Всеросс. конф. Математические методы распознавания образов-15. — М.: МАКС Пресс, 2011. — С. 60–63.
5. Фрей А. И., Ивахненко А. А., Решетняк И. М. Применение комбинаторных оценок вероятности переобучения в простом голосовании пороговых конъюнкций // Межд. конф. Интеллектуализация обработки информации ИОИ-9. — М.: МАКС Пресс, 2012. — С. 86–89.
6. Фрей А. И., Толстихин И. О. Комбинаторные оценки вероятности переобучения на основе кластеризации и покрытий множества алгоритмов // Machine learning and data analysis. — 2013. — Т. 1(6). — С. 751–767.

7. Frei A. I. Accurate estimates of the generalization ability for symmetric set of predictors and randomized learning algorithms // Pattern Recognition and Image Analysis. — 2010. — V. 20, no. 3. — P. 241–250.
8. Vorontsov K., Frey A. I., Sokolov E. Computable combinatorial overfitting bounds // Machine learning and data analysis. — 2013. — V. 1(6). — P. 724–733.

В работах с соавторами лично соискателем сделано следующее:

- проведены эксперименты по исследованию обобщающей способности логических алгоритмов классификации [5];
- получена оценка вероятности переобучения, обобщающая оценку метода порождающих и запрещающих множеств на случай произвольного разбиения множества алгоритмов на кластеры [6];
- проведены эксперименты по сравнению комбинаторных и PAC-Bayes оценок переобучения; экспериментально исследованы кривые обучения логистической регрессии [8].

Фрей Александр Ильич

ТЕОРЕТИКО-ГРУППОВОЙ ПОДХОД В КОМБИНАТОРНОЙ ТЕОРИИ ПЕРЕОБУЧЕНИЯ

АВТОРЕФЕРАТ

Подписано в печать: 11.11.2013. Формат $60 \times 84\,{}^{1}/_{16}$.

Печать трафаретная. Объем: 1,0 усл. печ. л.

Тираж – 100 экз. Заказ № 339.

Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Московский физико-технический институт (государственный университет)»

Отдел оперативной полиграфии «Физтех-полиграф»

141700, Московская обл., г. Долгопрудный,
Институтский пер., 9.