На правах рукописи

КОЧЕДЫКОВ ДЕНИС АЛЕКСЕЕВИЧ

ОЦЕНКИ ОБОБЩАЮЩЕЙ СПОСОБНОСТИ НА ОСНОВЕ ХАРАКТЕРИСТИК РАССЛОЕНИЯ И СВЯЗНОСТИ СЕМЕЙСТВ ФУНКЦИЙ

05.13.17 — теоретические основы информатики

Автореферат диссертации на соискание ученой степени
кандидата физико-математических наук

Москва, 2011

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

Научный руководитель:	доктор физико-математических наук, **Воронцов Константин Вячеславович**
Официальные оппоненты:	доктор физико-математических наук, **Дьяконов Александр Геннадьевич**
	кандидат физико-математических наук, **Червоненкис Алексей Яковлевич**
Ведущая организация:	Московский физико-технический институт (государственный университет)

Защита диссертации состоится « ____ » _____ 2011 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 « ____ » _____ 2011 г.

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

Общая характеристика работы

Работа посвящена проблеме повышения точности оценок обобщающей способности алгоритмов классификации в рамках комбинаторной теории надёжности обучения по прецедентам.

Актуальность темы. В задаче обучения по прецедентам рассматривается *генеральная совокупность* объектов на которой задана некоторая *целевая* функция. Из совокупности случайным образом извлекается *обучающая выборка* объектов (прецедентов). *Метод обучения* получает на вход обучающую выборку со значениями целевой функции на ее объектах и на выходе дает функцию, которая должна аппроксимировать целевую функцию на оставшейся (скрытой) части генеральной совокупности, называемой *контрольной выборкой*. Функцию, возвращаемую методом, традиционно называют *алгоритмом*, имея в виду, что процедура вычисления значения функции на объектах совокупности должна допускать эффективную компьютерную реализацию. Множество всех алгоритмов, которые может выдать метод обучения, называют *семейством алгоритмов*.

Задача оценивания обобщающей способности метода обучения состоит в том, чтобы, имея в распоряжении лишь наблюдаемую обучающую выборку, определить, насколько хорошо алгоритм, выданный методом, аппроксимирует целевую зависимость на скрытой части совокупности. Качество алгоритма на множестве объектов обычно характеризуется числом или частотой ошибок. Если частота ошибок на скрытой части (частота на контроле) существенно выше, чем на обучающей выборке (частота на обучении), то говорят, что метод переобучился или что выбранный им алгоритм переобучен.

В предположении, что все обучающие выборки заданного размера равновероятны, ставится задача оценивания *вероятности переобучения* метода. В англоязычной литературе такая постановка для бесконечной генеральной совокупности носит название *PAC-обучения* (probably approximately correct learning, Valiant 1984; Boucheron, Bousquet, Lugosi, 2004). Случай конечной генеральной совокупности рассматривается в *комбинаторной теории надежности обучения по прецедентам* (Воронцов, 2010).

Чтобы исключить зависимость от метода обучения, имеющего обычно довольно сложную структуру, рассматривают *вероятность равномерного отклонения частот* — вероятность того, что в семействе *возможно* выбрать алгоритм, у которого частота ошибок на контрольной выборке существенно больше его частоты ошибок на обучающей выборке.

Классические оценки в PAC-теории крайне завышены, поскольку ориентированы на худший возможный случай целевой зависимости. Одним из наиболее актуальных направлений исследований в связи с этим является получение оценок, зависящих от свойств целевой функции, семейства и обучающей выборки.

Одними из основных факторов завышенности классических оценок является пренебрежение *расслоением* семейства по частоте ошибок и *сходством* алгоритмов в семействе. Учет обоих факторов приводят к существенному уточнению оценок вероятности переобучения в комбинаторной теории (Воронцов, 2009). Однако точные оценки к настоящему моменту получены лишь для некоторых модельных семейств алгоритмов и довольно узкого класса методов обучения.

Цель работы. Разработка новых методов получения оценок обобщающей способности, учитывающих расслоение и сходство для произвольных семейств и методов обучения, в рамках комбинаторной теории надежности обучения по прецедентам.

Научная новизна. В работе развивается два метода получения оценок обобщающей способности. Оценки, использующие расслоение семейства по частоте ошибок, рассматривались ранее в контексте классической PAC-теории. В данной работе выводятся их комбинаторные аналоги с некоторыми улучшениями. Второй метод — оценки, учитывающие сходство алгоритмов в смысле расстояния Хэмминга между векторами ошибок алгоритмов. Он основан на неравенствах типа Бонферрони, оценивающих вероятность конъюнкции большого числа событий через вероятности дизъюнкции их различных комбинаций и технику производящих функций. Данный метод является новым. Основная оценка параграфа 4.5 вводит понятие *степени связности* алгоритма a — числа алгоритмов на единичном расстоянии от a и понятие *профиля связности* семейства — распределение степени связности в семействе. Оценка улучшает базовую оценку Вапника-Червоненкиса на множитель, экспоненциальный по средней степени связности алгоритмов в семействе (для линейных классификаторов — по размерности пространства параметров). Для семейства линейных классификаторов в работе получены оценки среднего значения и дисперсии профиля связности.

Методы исследования. Основными методами исследования в работе являются комбинаторная теория надежности обучения по прецедентам, оценки концентрации вероятностной меры,

неравенства типа Бонферрони-Галамбоса, используемые для оценивания вероятности равномерного отклонения частот, метод производящих функций перечислительной комбинаторики, используемый для вычисления отдельных слагаемых неравенств типа Бонферрони. Для анализа профиля связности семейства линейных классификаторов в работе используется теория геометрических конфигураций, применяемая в теории обучения для решения гораздо более простой задачи — оценивания числа алгоритмов в семействе. Для экспериментального вычисления и сравнения оценок используется метод Монте-Карло.

Результаты, выносимые на защиту.

1. Получены комбинаторные аналоги shell-оценок Лэнгфорда-МакАллистера, показано, что они являются аналогом классических оценок Вапника-Червоненкиса и «бритвы Оккама» Блумера, и имеют ту же степень завышенности.

2. Предложен новый способ учета сходства алгоритмов в оценках вероятности переобучения, основанный на Бонферрони-оценках вероятности равномерного отклонения частот и методе производящих функций.

3. Получены оценки вероятности переобучения для случаев связного семейства, семейства с известным профилем расслоения-связности, семейства, состоящего из множества монотонных максимальных цепей алгоритмов.

4. Для семейства линейных классификаторов получены оценки среднего и дисперсии профиля связности.

Теоретическая и практическая значимость. Работа носит в основном теоретический характер и вносит существенный вклад в развитие комбинаторной теории надежности обучения по прецедентам. Предложенные методы учета сходства алгоритмов могут применяться для конкретных семейств как в рамках комбинаторного подхода, так и в рамках классического PAC-подхода для уточнения оценок, использующих неравенство Буля.

Основным практическим приложением оценок обобщающей способности является разработка и обоснование новых методов обучения. Они также могут служить источником качественных соображений при выборе семейства. К примеру, основная оценка настоящей работы показывает, что при повышении сложности семейства существенным фактором, уменьшающим вероятность переобучения, является увеличение степени сходства алгоритмов в семействе, что может служить обоснованием для применения семейств, непрерывных по параметрам.

Апробация работы. Результаты работы докладывались на российских и международных научных конференциях:

- всероссийская конференция «Математические методы распознавания образов» ММРО-12, 2005 г. [8];
- международная конференция «Интеллектуализация обработки информации» ИОИ-6, 2006 г. [7];
- всероссийская конференция «Математические методы распознавания образов» ММРО-13, 2007 г. [6];
- научная конференция МФТИ 50 «Современные проблемы фундаментальных и прикладных наук» 2007 г. [5];

- научная конференция МФТИ 51 «Современные проблемы фундаментальных и прикладных наук» 2008 г. [4];

- всероссийская конференция «Математические методы распознавания образов» ММРО-14, 2009 г. [3].

Результаты неоднократно докладывались на семинарах отдела Интеллектуальных систем ВЦ РАН (руководитель — чл.-корр. РАН Константин Владимирович Рудаков).

Публикации по теме диссертации в изданиях из списка ВАК: [2, 1]. Другие публикации по теме диссертации: [8, 7, 6, 5, 4, 3]. Текст диссертации доступен на странице автора `www.MachineLearning.ru/wiki`, «Участник:Denis Kochedykov».

Структура и объём работы. Работа состоит из оглавления, введения, пяти глав, заключения, списка обозначений, списка литературы (66 пунктов). Общий объём работы — 101 стр.

Краткое содержание работы по главам

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

Введение

§1.1. Пусть $\mathcal{X}$ — множество объектов, Y — множество допустимых ответов, $\mathcal{F}$ — параметрическое семейство отображений из $\mathcal{X}$ в Y, называемых *алгоритмами*, $\mathbb{X} = \{x_1, \ldots, x_L\} \subset \mathcal{X}$ — фик-

сированная конечная генеральная совокупность из L объектов, называемая *полной выборкой*.

Будем называть подмножества $X \subset \mathbb{X}$, $|X| = \ell$ *обучающими* выборками. *Метод обучения* μ есть отображение $\mu\colon 2^{\mathbb{X}} \to \mathcal{F}$.

Пусть задана бинарная *функция потерь* $I\colon \mathcal{F} \times \mathbb{X} \to \{0,1\}$. Если $I(f,x) = 1$, то говорят, что алгоритм $f \in \mathcal{F}$ допускает ошибку на объекте $x \in \mathbb{X}$.

Вводится множество L-мерных бинарных *векторов ошибок* алгоритмов из $\mathcal{F}$ на полной выборке $\mathbb{X}$:

$$A = \left\{ \big(I(f,x_i)\big)_{i=1}^{L} \;\middle|\; f \in \mathcal{F} \right\} \;\subseteq\; \{0,1\}^L . \tag{1}$$

Будем для краткости обозначать (1) как $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ и называть векторы из A также «алгоритмами», имея ввиду произвольный алгоритм из соответствующего класса эквивалентности на $\mathcal{F}$.

Число ошибок алгоритма $a \in A$ на $X \subseteq \mathbb{X}$ есть $n(a,X) = \\ = \operatorname{card}\{x \in X\colon I(a,x) = 1\}$, *частота ошибок* есть $\nu(a,X) = \\ = n(a,X)/|X|$. Будем пользоваться сокращенными обозначениями $n(a,\mathbb{X}) \equiv n_a$, $n(a,X) \equiv \hat{n}_a$, $\nu(a,\mathbb{X}) \equiv \nu_a$, $\nu(a,X) \equiv \hat{\nu}_a$.

Будем обозначать через ν, $\hat{\nu}$ (без индекса a) допустимые значения частоты на полной/обучающей выборке, через m, s — допустимые значения числа ошибок на полной/обучающей выборке.

Допустим, что все разбиения полной выборки $\mathbb{X}$ на две подвыборки, наблюдаемую обучающую X длины ℓ и скрытую контрольную $\mathbb{X}\backslash X$ длины $L-\ell$, равновероятны. *Вероятность события* φ есть доля разбиений, для которых предикат φ истинен:

$$\mathbf{P}[\,\varphi(X)] \;\stackrel{\text{def}}{=}\; \frac{1}{\binom{L}{\ell}} \sum_{X \in [\mathbb{X}]^\ell} [\,\varphi(X)]\,,$$

где $[\mathbb{X}]^\ell$ — множество всех ℓ-элементных подмножеств $\mathbb{X}$.

Нашей основной задачей будет получение как можно более точных доверительных оценок $\bar{\nu}$ истинной частоты ошибок:

$$\mathbf{P}[\,\nu_{\hat{a}} < \bar{\nu}(\hat{a}, X, \mathcal{F}, \mu, \alpha)] \;\geqslant\; 1 - \alpha,$$

где $\bar{\nu}$ — некоторая оценочная функция, значение $\alpha \in (0,1)$ достаточно близко к нулю. Назначение таких оценок состоит в том, чтобы, минимизируя оценочную функцию $\bar{\nu}(\hat{a}, X, \mathcal{F}, \mu, \alpha)$ по $\mathcal{F}, \mu$, сконструировать метод обучения μ и семейство $\mathcal{F}$ с наилучшей обобщающей способностью.

§1.2. Вводится общий критерий переобучения $U(n_a, \hat{n}_a) \geqslant \varepsilon$, рассматриваются некоторые его частные случаи: $n_a/L - \hat{n}_a/\ell \geqslant \varepsilon$, $(n_a - \hat{n}_a)/(L-\ell) - \hat{n}_a/\ell \geqslant \varepsilon$ и $1 - H_{n_a}(\hat{n}_a) \geqslant 1 - \eta$. Определяется вероятность переобучения $\mathbf{P}[\,U(n_{\hat{a}}, \hat{n}_{\hat{a}}) \geqslant \varepsilon]$ и процедура обращения оценки вероятности переобучения:

$$\mathbf{P}[\,U(n_{\hat{a}}, \hat{n}_{\hat{a}}) \geqslant \varepsilon] \leqslant P(\varepsilon) \;\Leftrightarrow\; \mathbf{P}\big[\,n_{\hat{a}} \geqslant U^{-1}\big(\hat{n}_{\hat{a}}, P^{-1}(\alpha)\big)\big] \leqslant \alpha.$$

§1.3. Приводится определение и доказываются некоторые свойства гипергеометрического распределения. Пусть имеется L объектов, из которых равновероятно выбирается без возвращения ℓ объектов. Если среди L объектов на m объектах алгоритм a делает ошибку, то вероятность того, что в выборку попадут s таких объектов, подчиняется гипергеометрическому распределению: $h_m(s) = \binom{m}{s}\binom{L-m}{\ell-s} \,/\, \binom{L}{\ell}$. Вводится функция гипергеометрического распределения $H_m(s) = \sum_{t=0}^{s} h_m(t)$.

§1.4. В классических работах критерий переобучения определяется как $\nu_a - \hat{\nu}_a \geqslant \varepsilon$, однако можно выбрать и другую меру уклонения $\hat{\nu}_a$ от ν_a, в частности, *квантильный критерий*:

$$H_{n_a}(\hat{n}_a) \leqslant \eta \quad\Leftrightarrow\quad n_a \geqslant \bar{n}\,(\hat{n}_a, \eta) \quad\Leftrightarrow\quad \hat{n}_a \leqslant s_{n_a}(\eta), \qquad (2)$$

$\bar{n}\,(\hat{n}_a, \eta) = \min\{m\colon H_m(\hat{n}_a) \leqslant \eta\}$ и $s_{n_a}(\eta) = \max\{s\colon H_{n_a}(s) \leqslant \eta\}$.

Третий вариант в (2) интерпретируется следующим образом: алгоритм a переобучен, если число его ошибок $\hat{n}_a$ на обучающей выборке меньше η-квантили распределения $H_{n_a}(s)$. Параметр η здесь играет ту же роль, что ε в критерии $\nu_a - \hat{\nu}_a \geqslant \varepsilon$, и также называется в работе *порогом переобучения*. Чем *меньше* η, тем *более* переобучен алгоритм a. Квантильный критерий удобен тем, что вероятность переобучения для одного алгоритма равна η независимо от его полного числа ошибок n_a.

Лемма 1.4.1. Для любого $a \in A$ и любого $\eta \in (0,1)$ справедлива доверительная оценка: $\mathbf{P}[\, n_a < \bar{n}\,(\hat{n}_a, \eta)] \;\geqslant\; 1 - \eta$.

Лемма 1.4.2. Для любого $a \in A$ и любого $\eta \in (0,1)$ справедлива доверительная оценка $\mathbf{P}[\, n_a > \underline{n}\,(\hat{n}_a, \eta)] \;\geqslant\; 1 - \eta$, где $\underline{n}\,(\hat{n}_a, \eta) = \max\{m\colon 1 - H_m(\hat{n}_a) \leqslant \eta\}$.

§1.5. Приводятся комбинаторные аналоги классических оценок Вапника-Червоненкиса и «бритвы Оккама» (Блумер, 1987).

Теорема 1.5.1. Для любого семейства $\mathcal{F}$, любой полной выборки $\mathbb{X}$, $|\mathbb{X}| = L$, метода обучения $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$, индикатора ошибки $\mathrm{I}\colon \mathbb{X} \times \mathcal{F} \to \{0,1\}$, и любого $\alpha \in (0,1)$ верна оценка

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta)] \;\leqslant\; |A|\,\eta$$

и доверительная оценка

$$\mathbf{P}\Big[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \tfrac{\alpha}{|A|})\Big] \;\leqslant\; \alpha.$$

Отметим, что в этой и последующих оценках величина α имеет смысл вероятности переобучения, а величина η — порога переобучения и вероятности переобучения для отдельного алгоритма; для VC-оценки $\alpha = |A|\,\eta$.

Теорема 1.5.2. Для любого семейства $\mathcal{F}$, любой полной выборки $\mathbb{X}$, $|\mathbb{X}| = L$, метода обучения $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$, индикатора

ошибки I: $\mathbb{X} \times \mathcal{F} \to \{0,1\}$, любого $\alpha \in (0,1)$ и *вектора* порогов переобучения $\boldsymbol{\eta} = (\eta_a \colon a \in A)$ имеет место оценка вероятности переобучения

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta_{\hat{a}})] \;\leqslant\; \sum_{a \in A} \eta_a.$$

При условии $\sum\limits_{a \in A} \eta_a \;=\; \alpha$, верна также доверительная оценка

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta_{\hat{a}})] \;\leqslant\; \alpha. \tag{3}$$

Оценка «бритвы Оккама» позволяет получать *более точные в среднем (по обучающим выборкам) доверительные оценки* для $n_{\hat{a}}$, задавая бóльшие пороги η_a (и, соответственно, получая меньшие оценки $\bar{n}\,(\hat{n}_a, \eta_a)$) для тех a, которые чаще являются результатом обучения.

§1.6. Показывается, что оптимальные (в некотором смысле) пороги переобучения в оценке «бритвы Оккама» должны быть пропорциональны вероятностям получения алгоритмов в результате обучения.

Лемма 1.6.1. Для любого семейства $\mathcal{F}$, полной выборки $\mathbb{X}$, $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ и метода μ, пусть $\boldsymbol{p} = \big(\mathbf{P}[\, \mu(X) = a] : \; a \in A\big)$ есть распределение вероятностей получения различных алгоритмов в результате обучения и $\boldsymbol{\eta} = \boldsymbol{q}\alpha$, $\sum\limits_{a \in A} q_a = 1$ есть нормированный вектор порогов переобучения. Тогда минимум $\min\limits_{\boldsymbol{q}} \mathbf{E}\,(-\ln \eta_{\hat{a}})$ достигается при $\boldsymbol{q} = \boldsymbol{p}$.

В параграфе показывается, что $-\ln \eta_{\hat{a}}$ имеет смысл квадрата уклонения $\big(\bar{\nu}\,(\hat{n}_a, \eta_a) - \hat{\nu}_a\big)^2$ и характеризует точность оценки «бритвы Оккама» для алгоритма $\hat{a}$.

§1.7. Приводится процедура оценивания вероятности переобучения методом Монте-Карло для экспериментального сравнения различных оценок вероятности переобучения.

Глава 2. Обзор литературы

Глава содержит краткий обзор методов и результатов теории статистического обучения (statistical learning theory).

Глава 3. Оценки на основе характеристик расслоения семейства

Оценки обобщающей способности, учитывающие расслоение семейства, называемые «shell-оценками», были предложены в работах (Лэнгфорд, МакАллестер 2000, Лэнгфорд 2002). В данной главе выводятся аналогичные комбинаторные оценки. Оценки выводятся более общим и простым образом, показывается, что shell-оценки являются частным случаем (или вариантом) оценок Вапника-Червоненкиса и «бритвы Оккама».

§3.1. Дается определение профиля расслоения и наблюдаемого профиля расслоения семейства. Приводятся примеры оценок профилей методом Монте-Карло для семейства линейных классификаторов, обсуждаются их свойства.

Определение 1. *Профиль расслоения* множества A есть $\Delta_m = \operatorname{card}\{a \in A\colon n_a = m\}$. Будем называть соответствующее подмножество $A_m = \{a \in A\colon n_a = m\}$ m-ым *слоем* множества A. *Профиль наблюдаемых частот* ошибок на обучающей выборке X есть случайная величина $\hat{\Delta}_s = \operatorname{card}\{a \in A\colon \hat{n}_a = s\}$.

§3.2. Краткий обзор работ, развивающих идею shell-оценок.

§3.3. Выводятся комбинаторные аналоги двух shell-оценок, зависящих от полной выборки $\mathbb{X}$. Показывается, что обе оценки являются вариантом или частным случаем оценок Вапника-Червоненкиса или «бритвы Оккама», хотя исходно они позиционировались как принципиально более точные благодаря учёту

существенно большего количества информации о задаче (профиля расслоения). Основной результат параграфа представлен следующей теоремой.

Теорема 3.3.3. Для любого семейства $\mathcal{F}$, полной выборки $\mathbb{X}$, $|\mathbb{X}| = L$, метода обучения μ, индикатора ошибки I, если профиль расслоения множества $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ есть Δ_m, то имеет место оценка вероятности переобучения

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta(\hat{n}_{\hat{a}})] \;\leqslant\; \sum_{s=0}^{\ell} P(s, \eta(s)),$$

где $P(s,\eta) \stackrel{\text{def}}{=} \sum\limits_{m \geqslant \bar{n}\,(s,\eta)} \Delta_m h_m(s)$ — верхняя оценка вероятности того, что какой-либо алгоритм из A имеет на обучающей выборке s ошибок и при этом переобучен.

Кроме того, справедлива доверительная оценка:

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta(\hat{n}_{\hat{a}}, \alpha)] \;\leqslant\; \alpha,$$

где $\eta(s,\alpha) = \max\left\{\eta\colon P(s,\eta) \leqslant \frac{\alpha}{\ell}\right\}$.

§3.4. Выводится комбинаторный вариант shell-оценки, зависящей от обучающей выборки X. Оценка мотивируется более простым и наглядным способом, чем исходная оценка Лэнгфорда — через верхнюю оценку $\Delta^*_m(\hat{\Delta}_s)$ для истинного профиля расслоения Δ_m по наблюдаемому профилю расслоения $\hat{\Delta}_s$. Исправляется ошибка в доказательстве исходной shell-оценки — пропущенное неравенство Буля по ℓ всевозможным значениям наблюдаемой частоты ошибки $\hat{\nu}_{\hat{a}}$. Основной результат параграфа представлен следующей теоремой.

Теорема 3.4.3. Для любого семейства $\mathcal{F}$, полной выборки $\mathbb{X}$, $|\mathbb{X}| = L$, метода обучения μ, индикатора ошибки I, справедлива доверительная оценка

$$\mathbf{P}[\, n_{\hat{a}} < \bar{n}\,(\hat{n}_{\hat{a}}, \hat{\eta}(\hat{n}_{\hat{a}}, \alpha))] \;\geqslant\; 1 - \alpha,$$

где $\hat{\eta}(s,\alpha) = \max\left\{\eta\colon 2\hat{P}(s,\eta,\frac{\delta}{4\ell}) \leqslant \frac{\alpha}{2\ell}\right\}$, $\hat{P}(s,\eta,\delta) = \sum\limits_m \Delta^*_m h_m(s)$ — пессимистичная оценка функции $P(s,\eta)$, $\Delta^*_m = \sum\limits_{s\colon n^*(s)=m} \hat{\Delta}_s$ — пессимистичная оценка профиля расслоения, $n^*(s)$ — пессимистичная оценка полного числа ошибок алгоритма с s ошибками на обучающей выборке,

$$n^*(s) = \min\left\{\bar{n}\,(s,\delta/2),\ \max\left\{\bar{n}\,(s,\eta),\ \underline{n}\,(s,\delta/2)\right\}\right\}.$$

§3.5. Как показывают эксперименты, точность shell-оценок не сильно отличается от точности VC-оценок. В эксперименте основная масса алгоритмов в семействе действительно концентрируется в области наихудшей частоты ошибок $\hat{\nu} = \frac{1}{2}$ и при этом метод обучения μ в основном выбирает алгоритмы из области малых частот ошибок. Таким образом, сложность эффективно используемой части семейства существенно меньше сложности всего семейства $\mathcal{F}$. Именно эти факты приводятся в качестве исходной мотивации shell-оценок. Однако фактически в shell-оценках они учитываются не в полной мере. Shell-оценки, как и VC-оценка, основаны на вероятности равномерного отклонения частот *по всему* семейству $\mathcal{F}$, а не по его части с малыми частотами ошибок. Основной причиной завышенности по-прежнему остаётся неравенство Буля, в котором суммирование вероятностей производится *по всему* семейству.

Преимущество shell-оценок в том, что они позволяют балансировать точность оценки для разных частот ошибок, делая оценку точнее для одних частот за счет ухудшения оценки для других, аналогично тому, что делается в оценке «бритвы Оккама» для отдельных алгоритмов. Эта идея представляется плодотворной, но выигрыш в точности, который она может дать, полностью нивелируется завышенностью оценки равномерной сходимости и неравенства Буля.

Глава 4. Оценки на основе характеристик сходства алгоритмов в семействе

В настоящей главе выводятся оценки обобщающей способности, учитывающие сходство алгоритмов в семействе.

§4.1. Приводится несколько точных разложений вероятности равномерного отклонения частот, включая разложение по принципу включения-исключения, как альтернатив неравенству Буля. Для удобства вычисления отдельных слагаемых таких разложений вводится понятие графа 1-сходства множества A.

Определение 2. Пусть $\rho(a, a') = \sum_{x \in \mathbb{X}} [\,\mathrm{I}(a, x) \neq \mathrm{I}(a', x)]$ — хэммингово расстояние между алгоритмами a и a'. Графом 1-сходства множества $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ будем называть граф $\mathrm{G}_A^1 = (A,\, \mathrm{E})$ со множеством ребер $\mathrm{E} = \{\{a, a'\} \in A \times A \colon \rho(a, a') = 1\}$, соединяющих алгоритмы, векторы ошибок которых отличаются на одном объекте.

§4.2. Приводится краткий обзор известных способов оценивания вероятности конъюнкции событий (или, в контексте обучения, — вероятности равномерного отклонения частот) и их приложений. Приводятся другие распространенные способы учета сходства алгоритмов в оценках обобщающей способности,

§4.3. Предлагается два способа вычисления вероятностей вида $\mathbf{P}[\,\mathrm{U}_{a_1} \ldots \mathrm{U}_{a_k}]$, где U_a есть $\hat{n}_a \leqslant s_{n_a}(\eta)$ — условие переобучения алгоритма a. Первый способ опирается на производящую функцию множества подмножеств из $\mathbb{X}$ с заданными свойствами. Второй способ основан на представлении условий $\mathrm{U}_{a_1} \ldots \mathrm{U}_{a_k}$ в виде системы линейных неравенств с вектором бинарных неизвестных. Между этими двумя способами устанавливается соответствие.

§4.4. Выводится оценка обобщающей способности, учитывающая связность графа 1-сходства A.

Теорема 4.4.3. Для любого семейства $\mathcal{F}$, любой полной выборки $\mathbb{X}$, метода обучения $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$ и индикатора ошибки I, если граф G^1_A множества $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ связный, то выполняется оценка вероятности переобучения:

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta)] \;\leqslant\; P(\eta) \;\stackrel{\text{def}}{=}\; \eta + |A| \max_m h_m(s_m(\eta))\,.$$

Также верна доверительная оценка

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta(\alpha))] \;\leqslant\; \alpha,$$

где $\eta(\alpha) \;=\; \max\{\eta\colon P(\eta) \leqslant \alpha\}$.

§4.5. Вводятся понятия полустепени связности алгоритма $\rho_+(a)$ и профиля расслоения-связности $\Delta_{m,q}$ множества A, доказывается оценка обобщающей способности, учитывающая степени связности алгоритмов в A.

Определение 3. *Верхняя полустепень связности* алгоритма a есть есть число алгоритмов в его верхней единичной окрестности: $\rho_+(a) = \operatorname{card}\{a' \in A\colon \rho(a,a') = 1,\ n(a',\mathbb{X}) = n(a,\mathbb{X}) + 1\}$. *Профиль расслоения-связности* множества $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ есть матрица чисел:

$$\Delta_{m,q} = \operatorname{card}\{a \in A\colon n(a,\mathbb{X}) = m,\ \rho_+(a) = q\}\,,\ m,q \in \{1,\ldots,L\}\,.$$

Основной результат параграфа представлен следующей теоремой.

Теорема 4.5.7. Для любого семейства $\mathcal{F}$, полной выборки $\mathbb{X}$, индикатора ошибки I, если профиль расслоения-связности множества $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ есть $\Delta_{m,q}$, то для любого метода обучения $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$ вероятность переобучения оценивается как

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta)] \;\leqslant\; P(\eta) \;\stackrel{\text{def}}{=}\; \eta N_0 \;+\; \sum_{q=1}^{L} \sum_{m=0}^{L} h_m(s_m)\,\Delta_{m,q}\alpha_m^q,$$

где N_0 — число алгоритмов в A с пустой верхней 1-окрестностью, $\alpha_m = \frac{\ell - s_m}{L-m}$ $\left[\, \frac{\ell - s_m}{L-m} < 1\right]$. Также верна доверительная оценка

$$\mathbf{P}\left[\, n_{\hat{a}} \geqslant \bar{n}\left(\hat{n}_{\hat{a}}, \eta(\alpha)\right)\right] \;\leqslant\; \alpha, \tag{4}$$

где $\eta(\alpha) \;=\; \max\{\eta\colon P(\eta) \leqslant \alpha\}$.

Поскольку VC-оценка может быть записана как $\eta\,|A| \;=\; \eta \sum_{m,q} \Delta_{m,q}$, то в последней теореме имеем множители α_m^q, $\alpha < 1$ к слагаемым VC-оценки. В эксперименте с семейством линейных классификаторов Главы 6, число N_0 алгоритмов без верхних связей в A пренебрежимо мало́ в сравнении с общим числом алгоритмов, что дает экспоненциальное улучшение оценки последней теоремы относительно неравенства Буля с ростом среднего q. Для модельного случая $A = \{0,1\}^L$, $N_0 = 1$, точное вычисление показывает, что последняя оценка улучшает оценку Вапника-Червоненкиса приблизительно в $2^{0.4L}$ раз.

§4.6. Выводится оценка, учитывающая наличие исходящей монотонной цепи для каждого алгоритма в семействе.

Определение 4. *Цепь алгоритмов* есть последовательность алгоритмов $a_1, \ldots, a_K$, таких, что $\rho(a_k, a_{k-1}) = 1$. Будем называть цепь *монотонной*, если $n_{a_k} = n_{a_{k-1}} + 1$. Будем называть монотонную цепь *максимальной*, если $a_K \in A_M$, $M = \max\limits_{a \in A} n_a$.

Наличие цепей в A — достаточно естественное предположение для семейств $\mathcal{F}$, непрерывных по параметрам. Цепь может возникать, если выбирается некоторая «начальная» функция в $\mathcal{F}$ и ее параметры изменяются вдоль некоторого непрерывного пути в пространстве параметров.

Основной результат параграфа представлен следующей теоремой.

Теорема 4.6.3. Для любого семейства $\mathcal{F}$, полной выборки $\mathbb{X}$, $|\mathbb{X}| = L$, индикатора ошибки $\mathrm{I}\colon \mathcal{F}\times\mathbb{X} \to \{0,1\}$, если в множестве $A = \mathrm{I}(\mathcal{F},\mathbb{X})$ для любого алгоритма a можно найти максимальную цепь, то для любого метода обучения $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$ верна верхняя оценка вероятности переобучения:

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta)] \;\leqslant\; P(\eta) \;\stackrel{\text{def}}{=}\; N_0\eta \;+\; \sum_{m=0}^{M} h_m(s_m)\,\Delta_m\beta_m,$$

где $\beta_m = \binom{L-m}{\ell-s_m}_{\boldsymbol{u}} \Big/ \binom{L-m}{\ell-s_m} < 1$ и $\binom{L-m}{\ell-s_m}_{\boldsymbol{u}}$ — усеченный биномиальный коэффициент с вектором ограничений

$$\boldsymbol{u} = ((s_{m+1} - s_m),\, \ldots,\, (s_M - s_m),\, \ldots,\, (s_M - s_m))$$

длины $L - m$. Также верна доверительная оценка

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta(\alpha))] \;\leqslant\; \alpha,$$

где $\eta(\alpha) = \max\{\eta\colon P(\eta) \leqslant \alpha\}$.

Усеченный биномиальный коэффициент определяется по аналогии с обычным: $\binom{n}{k}_{\boldsymbol{u}} = \left(\binom{n-1}{k}_{\boldsymbol{u}} + \binom{n-1}{k-1}_{\boldsymbol{u}}\right) \times [\, k > u_n]$ с граничными условиями $\binom{n}{0}_{\boldsymbol{u}} = 1$.

§4.7. Экспериментальное вычисление оценок настоящей главы в главе 6 позволяет предположить, что для существенного улучшения оценки необходим учет сходства алгоритмов не столько «в глубину» (крайним случаем которого является максимальная цепь), сколько «в ширину» (крайним случаем которого является единичная окрестность), то есть, учет окрестностей радиуса $\rho > 1$ в A и, в пределе, учет сходства каждого алгоритма со всеми алгоритмами, в которые из него идут монотонные цепи.

Глава 5. Характеристики связности семейства линейных классификаторов

В настоящей главе исследуются среднее и дисперсия профиля связности для семейства линейных классификаторов:

$$\mathcal{F} = \left\{ a_w(x) \stackrel{\text{def}}{=} \operatorname{sign}\langle w, x\rangle \colon w \in \mathbb{W} \equiv \mathbb{R}^{p+1},\ x \in \{1\} \times \mathbb{R}^p \right\}. \quad (5)$$

с индикатором ошибки $I(a_w, x) = [\, a_w(x) \neq y(x)]$, где y — целевая функция.

Точная форма профиля связности $\mathcal{F}$ зависит от полной выборки $\mathbb{X}$, однако среднее значение профиля и оценка его дисперсии, полученные в Теоремах 5.2.2, 5.4.1 настоящей главы, не зависят от $\mathbb{X}$ и являются комбинаторными свойствами семейства линейных классификаторов.

§5.1. Вводится понятие конфигурации гипеплоскостей, ячейки и грани конфигурации и определяется их взаимосвязь с множеством A и его свойствами.

d-мерной конфигурацией $\mathcal{H}(\mathbb{X})$ однородных гиперплоскостей называется множество из L проходящих через 0 гиперплоскостей в $\mathbb{W}$, взаимнооднозначно соответствующих объектам в $\mathbb{X}$:

$$\mathcal{H}(\mathbb{X}) = \{h(x_i) \colon x_i \in \mathbb{X}\}, \quad h(x_i) = \{w \in \mathbb{W} \colon \langle w, x_i\rangle = 0\}.$$

Гиперплоскость $h(x_i)$ разделяет $\mathbb{W}$ на два полупространства, соответствующие классификаторам, дающим правильный и неправильный ответ на объекте x_i. Будем называть первое полупространство положительным, второе — отрицательным.

Гиперплоскости $\mathcal{H}$ разбивают пространство $\mathbb{W}$ на множество *ячеек* $\{c(a), a \in A\}$, взаимнооднозначно соответствующих алгоритмам в A. Каждая ячейка представляет собой d-мерный многогранный бесконечный конус с вершиной в 0, включающий

в себя все свои грани. Грани размерности $d-1$ взаимнооднозначно соответствуют парам смежных $\rho(a,a')=1$ алгоритмов в A.

Будем говорить, что $(d-1)$-мерная грань ячейки $c(a), a \in A$ положительна, если ячейка лежит в положительном полупространстве соответствующей гиперплоскости, то есть алгоритм a не допускает ошибки на x, а смежный с ним алгоритм — допускает. Тогда значение профиля связности Δ_q^+ равно числу ячеек в $\mathcal{H}$, имеющих ровно q положительных граней, а значение профиля расслоения Δ_m — числу ячеек, лежащих ровно в m отрицательных полупространствах.

Известно, что для $\mathbb{X}$ в общем положении полное число ячеек и граней в конфигурации $\mathcal{H}$ есть, соответственно (Эдельсбруннер, 1987):

$$C_0(L,d) = 2\sum_{k=0}^{d-1} \tbinom{L-1}{k}, \qquad C_1(L,d) = 2L\sum_{k=0}^{d-2} \tbinom{L-2}{k}. \tag{6}$$

§5.2. Выводится точное значение для средней степени связности алгоритмов в A. Для небольших ($< L/2$) размерностей пространства параметров, средняя степень связности равна размерности семейства. Это примерно соответствует уменьшению оценки Теоремы 4.5.7 в 2^{-p} раз в сравнении с VC-оценкой, и согласуется с результатами экспериментального вычисления оценок в Разделе 6.

Теорема 5.2.2. Пусть $\mathcal{F}$ — семейство линейных классификаторов (5) с индикатором ошибки (5.2) и объекты выборки $\mathbb{X} \subset \mathbb{R}^p$ находятся в общем положении. Тогда средняя полустепень связности алгоритмов во множестве $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ есть

$$\bar{\rho}_\pm \;=\; |A|^{-1} \sum_{a\in A} \rho_\pm(a) \;=\; \frac{L \cdot \sum_{k=0}^{p-1} \binom{L-2}{k}}{\sum_{k=0}^{p} \binom{L-1}{k}}.$$

§5.3. Дается определение зоны гиперплоскости в конфигурации гиперплоскостей, приводится лемма о максимальной сложности зоны в неоднородной конфигурации и доказывается лемма о максимальной сложности в однородной конфигурации. Последняя лемма используется в следующем параграфе для получения оценки дисперсии связности алгоритмов в A.

§5.4. Выводится верхняя оценка для дисперсии степени связности в множестве A. Дисперсия связности определяется суммой квадратов степеней связности или, иначе говоря, суммарным числом пар положительных связей по алгоритмам в A.

Идея оценки состоит в том, чтобы для каждого объекта $x \in \mathbb{X}$ рассмотреть подмножество таких алгоритмов в A, что для каждого из них в A существует алгоритм, отличающийся от него только на x, и оценить суммарное число связей этих алгоритмов при помощи теоремы о сложности зоны гиперплоскости из предыдущего параграфа. Это число связей в свою очередь равно числу тех из интересующих нас *пар* связей, в которых одна из связей идет через объект x. Суммируя такие оценки по объектам $x \in \mathbb{X}$ получаем требуемое число пар связей.

Теорема 5.4.1. Пусть $\mathcal{F}$ есть семейство линейных классификаторов (5) с индикатором ошибки (5.2) и объекты выборки $\mathbb{X} \subset \mathbb{R}^p$ находятся в общем положении. Тогда дисперсия полустепени связности алгоритмов во множестве $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$

$$\mathbf{Var}_a(\rho_\pm(a)) = |A|^{-1} \sum_{a \in A} (\rho_\pm(a) - \bar{\rho_\pm})^2$$

оценивается сверху как

$$\mathbf{Var}_a(\rho_\pm(a)) \leqslant L \cdot \frac{C_0(L-1,d-1)+Z_1(L-1,d)-C_1(L-1,d-1)}{C_0(L,d)} - \frac{C_1(L,d)^2}{C_0(L,d)^2},$$

где $Z_1(L-1,d)$ — максимальная сложность зоны в конфигурации $L-1$ однородных гиперплоскостей.

Глава 6. Эксперименты с семейством линейных классификаторов

Экспериментально оцениваются профили расслоения и связности для семейства линейных классификаторов, вычисляются оценки обобщающей способности из предшествующих глав.

§6.1. Приводится процедура Монте-Карло оценки профилей $\Delta_m, \Delta_q^+, \Delta_{m,q}$ и примеры профилей для случайной полной выборки $\mathbb{X}$. Выдвигается гипотеза о возможности представления профиля $\Delta_{m,q}$ в виде произведения профилей Δ_m и Δ_q^+.

§6.2. Для оценивания профилей методом Монте-Карло требуется равномерный сэмплинг большого числа векторов из множества A. Выбор вектора из $\{0,1\}^L$ и определение его принадлежности A требует решения системы L неравенств на каждом шаге и не может использоваться в методе Монте-Карло.

Предлагается процедура Π_A неравномерного сэмплинга из A и доказывается лемма, позволяющая вычислять вероятность извлечения процедурой Π_A заданного алгоритма $a \in A$ одновременно с вычислением степени связности $\rho_+(a)$.

§6.3. Эксперименты показывают, что shell-оценки незначительно улучшают оценки Вапника-Червоненкиса и «бритвы Оккама» и подтверждают, что оценка, учитывающая степени связности алгоритмов, меньше VC-оценки на множитель, экспоненциальный по средней степени связности в A.

Публикации по теме диссертации

[1] **Kochedykov D. A. A combinatorial approach to hypothesis similarity in generalization bounds // Pattern Recognition and Image Analysis, December 2011 — Vol. 21 no. 4.**

[2] **Kochedykov D. A. Combinatorial shell bounds for generalization ability // Pattern Recognition and Image Analysis, December 2010 — Vol. 20 no. 4. — Pp. 459–473.**

[3] Кочедыков Д. А. Структуры сходства в семействах алгоритмов классификации и оценки обобщающей способности // Докл. 14-й Всеросс. конф. «Математические методы распознавания образов» — М.: МАКС Пресс, 2009. — С. 45–48.

[4] Кочедыков Д. А. Комбинаторные оценки обобщающей способности методов обучения по прецедентам с расслоением по наблюдаемой частоте ошибок // Труды 51-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук»: Часть VII. Управление и прикладная математика. — М.: МФТИ, 2008.

[5] Кочедыков Д. А., К определению понятия информативности логических закономерностей в задачах классификации // Труды 50-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук»: Часть VII. Управление и прикладная математика. — М.: МФТИ, 2007. — С. 279–281.

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

[7] Кочедыков Д. А., Воронцов К. В. О поиске оптимальных сочетаний управляющих параметров в логических алгоритмах классификации // Тезисы докл. межд. конф. «Интеллектуализация обработки информации», Симферополь, 2006. — С. 117–118.

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