На правах рукописи

Толстихин Илья Олегович

Неравенства концентрации вероятностной меры в трансдуктивном обучении и PAC-Байесовском анализе

Специальность 05.13.17 —
«Теоретические основы информатики»

Автореферат

Диссертация на соискание учёной степени
кандидата физико-математических наук

Москва — 2014

Работа выполнена в отделе Интеллектуальных систем Федерального государственного бюджетного учреждения науки Вычислительный центр имени А. А. Дородницына Российской академии наук.

Научный руководитель:	доктор физико-математических наук, **Воронцов Константин Вячеславович**.
Официальные оппоненты:	доктор физико-математических наук, зав. отделом математического программирования Федерального государственного бюджетного учреждения науки Института математики и механики им. Н. Н. Красовского Уральского отделения Российской академии наук, **Хачай Михаил Юрьевич**; доктор физико-математических наук, главный научный сотрудник лаборатории №3 Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А. А. Харкевича Российской академии наук, **Голубев Георгий Ксенофонтович**.
Ведущая организация:	Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В. А. Трапезникова Российской академии наук.

Защита состоится «16» октября 2014 г. в 15:00 на заседании диссертационного совета Д 002.017.02 при Федеральном государственном бюджетном учреждении науки Вычислительный центр имени А. А. Дородницына Российской академии наук, расположенном по адресу: 119333, г. Москва, ул. Вавилова, д. 40, конференц-зал.

С диссертацией и авторефератом можно ознакомиться в библиотеке и на официальном сайте (`http://www.ccas.ru/`) ВЦ РАН.

Автореферат разослан «____» _____ 2014 г.

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

Общая характеристика работы

Диссертационная работа посвящена теории статистического обучения, изучающей свойства процедур обучения в рамках строгого математического формализма.

Актуальность темы. Задачи поиска закономерностей или восстановления функциональных зависимостей в наблюдаемых данных сегодня играют ключевую роль во многих прикладных областях. Методы машинного обучения, позволяющие во многих случаях эффективно решать задачи распознавания образов, классификации, восстановления регрессии, оценивания неизвестной плотности и другие задачи предсказания, стали неотъемлемой частью различных аспектов современной жизни. С теоретической точки зрения, важным вопросом является выявление факторов, влияющих на качество работы найденных на основе обучающей выборки закономерностей на новых данных, что позволило бы разрабатывать новые и более качественные алгоритмы обучения.

Теория статистического обучения (или *VC-теория*), предложенная в работах В. Н. Вапника и А. Я. Червоненкиса в конце 1960-х годов и позже получившая мировую известность, впервые позволила строго описать соотношение между необходимым для успешного обучения числом наблюдаемых данных и сложностью используемого класса отображений. Подобные результаты формулируются обычно в виде верхних границ (или *оценок*) на вероятность того, что найденное на основе обучающей выборки отображение даст ошибочный ответ на новых данных. Проблемой первых оценок была их сильная завышенность, обусловленная попыткой получения результатов, справедливых в чересчур общих постановках. Дальнейшее развитие VC-теории было связано с попытками улучшения точности оценок на основе учета различных свойств рассматриваемых задач (Boucheron S., Lugosi G., Bousquet O, 2005). Среди предложенных за последние 45 лет подходов VC-теории можно выделить результаты, основанные на покрытиях класса отображений (Kearns M. J., Schapire R. E., 1990; Bartlett P. L., Long P. M., Williamson R. C., 1996), на учете отступов объектов (Schapire R., Freund Y., Bartlett P., W. L., 1998; Koltchinskii V., Panchenko D., 2002), на понятии стабильности процедуры обучения (Bousquet O., Elisseeff A., 2002), на глобальной Радемахеровской сложности класса (Koltchinskii V., Panchenko D., 1999; Bartlett P. L., Mendelson S., 2001), на локальных мерах сложности класса (Koltchinskii V., Panchenko D., 1999; Massart P., 2000; Bartlett P., Bousquet O., Mendelson S., 2005; Koltchinskii V., 2006) и на изучении рандомизированных отображений (Shawe-Taylor J., Williamson R. C., 1997; McAllester D., 1998; Langford J., Shawe-Taylor J., 2002; Seeger M., 2002).

Несмотря на многочисленные попытки улучшения точности оценок, которые продолжаются до настоящего времени (Mendelson S., 2014), она остается по-прежнему не достаточной для применения оценок на практике. Поэтому актуальной проблемой является получение более точных оценок, с одной стороны достаточно общих, но с другой стороны учитывающих специфику решаемой прикладной задачи.

Цель диссертационной работы. Улучшение точности существующих оценок в теории статистического обучения на основе современных результатов теории неравенств концентрации вероятностной меры. Получение новых неравенств концентрации для выборок случайных величин без возвращения, учитывающих их дисперсии.

Методы исследования. В первой части настоящей работы используется *энтропийный* подход в *теории неравенств концентрации вероятностной меры* (Boucheron S., Lugosi G., Massart P., 2013), предложенный М. Леду и развитый П. Массаром, С. Бушроном, Г. Лугоши и О. Буске в работах (Ledoux M., 1996; Massart P., 2000; Boucheron S., Lugosi G., Massart P., 2013). Данный подход позволяет получать неравенства концентрации, сравнимые по точности с более сильными результатами *индуктивного подхода* (Talagrand M., 1995) М. Талаграна, избегая при этом чересчур громоздких доказательств. В частности, ключевую роль будут играть субгауссовское неравенство концентрации для функций, определенных на срезах Булева куба, полученное С. Г. Бобковым в работе (Bobkov S., 2004), и неравенство Талаграна для супремумов эмпирических процессов (Talagrand M., 1996), позже усиленное О. Буске (Bousquet O., 2002) на основе энтропийного подхода.

Вторая часть работы будет использовать подход в *теории статистического обучения* (Vapnik V. N., 1998; Boucheron S., Lugosi G., Bousquet O., 2005), существенно основанный на результатах теории неравенств концентрации вероятностной меры и *теории эмпирических процессов* (van der Vaart A. W., Wellner J., 2000; van de Geer S., 2000). Предложенный впервые в конце 60-х годов в работах В. Н. Вапника и А. Я. Червоненкиса (Vapnik V. N., Chervonenkis A. Y., 1968; Vapnik V. N., Chervonenkis A. Y., 1971), данный подход продолжает активно развиваться и на сегодняшний день. В частности, ряд важных результатов настоящей работы будет основан на так называемом *локальном* подходе, развитом в начале 2000-х годов В. Колчинским, Д. Панченко, П. Массаром, П. Бартлетом, О. Буске, Ш. Мендельсоном, Г. Лугоши и рядом других авторов в серии работ (Koltchinskii V., Panchenko D., 1999; Massart P., 2000; Bartlett P., Bousquet O., Mendelson S., 2005; Koltchinskii V., 2006). В отличие от большинства других подходов теории Вапника-Червоненкиса, локальный подход позволяет эффективно

учитывать свойства конкретной решаемой задачи при оценке качества процедуры обучения, что часто ведет к существенно более точным результатам.

Третья часть работы основана на *комбинаторной теории переобучения*, предложенной К. В. Воронцовым (Воронцов К. В., 2011; Vorontsov K. V., 2010; Vorontsov K. V., Ivahnenko. A. A., 2011; Vorontsov K. V., Frey A. I., Sokolov E. A., 2013). В частности, мы будем использовать *теоретико-групповой подход*, развитый в работах (Фрей А. И., 2009; Фрей А. И., 2010).

Наконец, четвертая часть работы использует *PAC-Байесовский анализ* — относительно новый подход в теории статистического обучения, предложенный Д. МакАллистером и Дж. Шоу-Тейлором в работах (Shawe-Taylor J., Williamson R. C., 1997; McAllester D., 1998) и далее развитый Дж. Лэнгфордом, М. Зигером (Langford J., Shawe-Taylor J., 2002; Seeger M., 2002) и рядом других авторов. Известно, что оценки PAC-Байесовского анализа в ряде прикладных задач ведут к наиболее точным на сегодняшний день результатам.

Основные положения, выносимые на защитузащиту:

1. Получено два новых неравенства концентрации типа Талаграна для супремумов эмпирических процессов и выборок без возвращения, которые учитывают дисперсии случайных величин.
2. Получена новая оценка избыточного риска в трансдуктивной постановке теории статистического обучения, основанная на локальных мерах сложности рассматриваемого класса отображений и впервые в трансдуктивном подходе ведущая к быстрой скорости сходимости в общих предположениях.
3. В рамках теоретико-группового подхода комбинатонной теории переобучения предложено учитывать орбиты множества выборок при вычислении точного значения вероятности переобучения. На основе этого подхода получены новые точные (не завышенные) оценки вероятности переобучения для трех модельных семейств отображений, бинарные векторы ошибок которых являются различными подмножествами шара в Булевом кубе.
4. В рамках PAC-Байесовского анализа теории статистического обучения получено новое PAC-Байесовское эмпирическое неравенство Бернштейна, полностью вычислимое на основе обучающей выборки и во многих случаях ведущее к существенно более точным оценкам по сравнению с известными ранее результатами.

Научная новизна. В диссертационной работе впервые показано, что в трансдуктивной постановке теории статистического обучения быстрая скорость сходимости может достигаться в достаточно общих предположениях. В частности, продемонстрировано, что избыточный риск при использовании ме-

тода минимизации эмпирического риска определяется величиной неподвижной точки модуля непрерывности эмпирического процесса в окрестностях оптимального на генеральной выборке отображения. Подобные результаты до этого были известны в задачах М-оценивания в теории эмпирических процессов и позже в индуктивной постановке теории статистического обучения. Однако, все они были основаны на неравенстве Талаграна для эмпирических процессов, которое, в свою очередь, существенно опирается на предположении о независимости случайных величин и, следовательно, не может быть использовано в трансдуктивной постановке теории статистического обучения.

Для преодоления этой трудности в диссертационной работе были впервые получены аналоги неравенства Талаграна, которые справедливы для случайных величин, выбранных равномерно без возвращения из произвольного конечного множества. До этого в литературе были известны лишь неравенства типа Мак-Диармида для супремумов эмпирических процессов и выборок без возвращения, не учитывающие дисперсии случайных величин.

Результаты, полученные в рамках теоретико-группового подхода в комбинаторной теории переобуения, основаны на новой идее учета при вычислении вероятности переобучения *орбит разбиений* генеральной выборки.

Новое PAC-Байесовское эмпирическое неравенство Бернштейна является первым примером PAC-Байесовских неравенств, одновременно учитывающих дисперсии потерь и вычислимых на основе обучающей выборки.

Теоретическая значимость. Полученные в диссертационной работе неравенства концентрации типа Талаграна являются достаточно общими и могут быть использованы в теоретическом анализе большого числа современных прикладных задач (в том числе выходящих за рамки теории обучения), где важную роль играют выборки без возвращения. Одним из примеров таких задач является неасимптотический анализ свойств процедуры скользящего контроля, широко применяемой на практике.

Новые оценки избыточного риска и обобщающей способности, полученные в диссертационной работе, улучшают точность известных ранее в теории статистического обучения результатов, давая более глубокое понимание процесса обучения на основе эмпирических данных в трансдуктивной постановке. В частности, новые оценки позволяют заключить, что сложность задач трансдуктивного обучения, по крайней мере, не превосходит сложность задач индуктивного обучения. Более того, они показывают, что свойства задач трансдуктивного обучения в ряде случаев могут выгодно отличаться от свойств задач индуктивного обучения.

Новые результаты комбинаторного подхода, полученные в диссертационной работе, расширяют класс задач и семейств отображений, для которых возмож-

но эффективное (полиномиальное по длине выборки) вычисление вероятности переобучения.

Практическая значимость. PAC-Байесовского эмпирическое неравенство Бернштейна, полученное в настоящей работе, во многих случаях ведет к существенно более точным оценкам обобщающей способности по сравнению с известными ранее результатами PAC-Байесовского анализа. Кроме того, новая оценка полностью вычислима на основе наблюдаемых данных. Это дает возможность эффективно применять ее на практике при решении задач обучения по прецедентам для оценивания качества получаемых решений или настройки гиперпараметров, избегая при этом больших вычислительных затрат процедуры скользящего контроля. Наконец, минимизация полученной оценки может вести к новым более точным методам обучения, имеющим гарантированную обобщающую способность.

Полученные в диссертационной работе оценки избыточного риска для трансдуктивного обучения могут вести к применимым на практике методам выбора моделей, основанным на использовании всех объектов генеральной совокупности. В частности, вместе со Следствием 15 (стр. 110) они могут вести к новым алгоритмам выбора ядер, поскольку собственные значения матрицы Грамма спрямляющего ядра, определяющие скорость сходимости метода минимизации эмпирического риска, в этом случае могут быть вычислены на основе наблюдаемых данных.

Степень достоверности. Достоверность результатов обеспечивается математическими доказательствами теорем и серией подробно описанных вычислительных экспериментов, результаты которых согласуются с теоретическими результатами настоящей работы.

Апробация работы. Результаты диссертационной работы неоднократно докладывались и обсуждались на следующих конференциях и научных семинарах:

1. Международная конференция «Ломоносов-2010», 2010 г. [4];
2. Международная конференция «Интеллектуализация обработки информации», 2010 г. [5];
3. Международная конференция «Интеллектуализация обработки информации», 2012 г. [6];
4. Международная конференция “Neural Information Processing Systems (NIPS)”, озеро Тахо, США, Декабрь 2013 г. [1];
5. Научный семинар группы проф. F. Laviolette и M. Marchand, Лавальский Университет, Квебек, Канада, Декабрь 2013 г.;
6. Три доклада на совместном НМУ–МФТИ семинаре «Стохастический анализ в задачах», Москва, Декабрь 2013 г. и Апрель 2014 г.;

7. Научный семинар группы профессора B. Schoelkopf, Max Planck Institute for Intelligent Systems, Тюбинген, Германия, Май 2014 г.;
8. Научный семинар Лаборатории 7 Института Проблем Управления РАН, Москва, Июнь 2014 г.;
9. Международная конференция “Conference on Learning Theory (COLT)”, Барселона, Испания, Июнь 2014 г. [3];
10. Научные семинары отдела Интеллектуальных систем Вычислительного Центра им. А. А. Дородницына РАН.

Публикации. Основные результаты настоящей диссертационной работы опубликованы в 7 работых [1–7], 3 из которых входят в список изданий, рекомендованных ВАК [1–3].

Личный вклад диссертанта заключается в выполнении основного объёма теоретических и экспериментальных исследований, изложенных в диссертационной работе. Все результаты, приведенные в настоящей работе, относятся к личному вкладу диссертанта, за исключением отдельно оговоренных случаев.

Подготовка к публикации полученных в работах [1–3, 7] результатов проводилась совместно с соавторами. Все экспериментальные и основная часть теоретических результатов работы [1] получены лично автором. В работах [2, 7] к личному вкладу автора относится разработка техники учета орбит разбиений при вычислении вероятности переобучения, использовавшаяся в доказательстве всех основных результатов данных работ, а также теоремы о вероятности переобучения центрального слоя Хэммингова шара и монотонного роста вероятности переобучения множеств бинарных векторов ошибок, лежащих в одном слое Булева куба.

Объем и структура работы Диссертация состоит из оглавления, введения, шести глав, заключения, списка иллюстраций (13 п.), списка таблиц (1 п.), списка литературы (124 п.) и списка обозначений. Общий объём работы составляет 201 страницу.

Краткое содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель работы, приводится список положений, выносимых на защиту, а также формулируется теоретическая и практическая значимость работы.

Первая часть диссертационной работы (**первая и вторая главы**) посвящена *теории неравенств концентрации вероятностной меры*[1]. Пусть дана функция

[1] Boucheron S., Lugosi G., Massart P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.

$g\colon \mathcal{X}^n \to \mathbb{R}$, зависящая от большого числа аргументов, принимающих значения в некотором множестве $\mathcal{X}$. Пусть также дано большое число случайных величин $X_1, \ldots, X_n$, принимающих значения в $\mathcal{X}$. Неравенства концентрации вероятностной меры позволяют ограничивать сверху вероятности отклонений $\mathsf{P}\{Q - \mathbb{E}[Q] \geqslant t\}$ и $\mathsf{P}\{\mathbb{E}[Q] - Q \geqslant t\}$ для $t > 0$, где $Q = g(X_1, \ldots, X_n)$.

В **первой главе** приводится подробный обзор классических и современных результатов теории неравенств концентрации вероятностной меры для *независимых случайных величин* $X_1, \ldots, X_n$. В **разделе 1.1** рассматриваются суммы независимых и ограниченных случайных величин, и для них приводятся неравенства Хефдинга, Бернштейна и Беннетта. В **разделе 1.2** рассматриваются функции с ограниченными разностями, и для них приводятся неравенства Азумы-Хефдинга и МакДиармида. В **разделе 1.3** формулируются основные результаты *энтропийного метода*[2] в теории неравенств концентрации, включая неравенства для самоограничивающих функций, а также приводятся эмпирическое неравенство Бернштейна[3] и неравенство Талаграна для супремумов эмпирических процессов[4].

Вторая глава посвящена результатам теории неравенств концентрации вероятностной меры для случайных величин $X_1, \ldots, X_n$, выбранных *без возвращения* из некоторого конечного множества $\mathcal{C} = \{c_1, \ldots, c_N\}$. Первая часть главы содержит подробный обзор известных результатов: в **разделе 2.1** рассматриваются суммы ограниченных случайных величин, и для них приводятся неравенство Стирлинга и метод редукции Хефдинга; в **разделе 2.2** рассматриваются функции, определенные на множестве разбиений генеральной выборки $\mathcal{C}$, и для них приводятся неравенство Эль-Янива-Печиони[5], которое является аналогом неравенства МакДиармида, и субгауссовское неравенство С. Г. Бобкова[6].

Неравенство Талаграна для супремумов эмпирических процессов, рассмотренное в разделе 1.3, играет важную роль во многих современных прикладных задачах. К сожалению, до сих пор в литературе не было известно его аналога для выборок *без возвращения*. В **разделе 2.3** приводятся основные результаты второй главы: два новых неравенства концентрации типа Талаграна для супремумов эмпирических процессов и выборок без возвращения, учитывающие дисперсии случайных величин (Теоремы 20 и 22).

---

[2] Ledoux M. *On Talagrand's Deviation Inequalities for Product Measures*, ESAIM: Probability and Statistics, 1996.

[3] Maurer A., Pontil M. *Empirical Bernstein Bounds and Sample Variance Penalization*, Proceedings of the International Conference on Computational Learning Theory (COLT), 2009.

[4] Talagrand M. *New concentration inequalities in product spaces*, Inventiones Mathematicae, 1996, Vol. 126.

[5] El-Yaniv R., Pechyony D. *Transductive Rademacher Complexity and its Applications*, Journal of Artificial Intelligence Research, 2009.

[6] Bobkov S. *Concentration of normalized sums and a central limit theorem for noncorrelated random variables*, Annals of Probability, 2004, Vol. 32.

Пусть $\mathcal{C} = \{c_1, \ldots, c_N\}$ — некоторое конечное множество. Для $n \leqslant N$ рассмотрим последовательности случайных величин $\eta_1, \ldots, \eta_n$ и $\xi_1, \ldots, \xi_n$, выбранные равномерно из $\mathcal{C}$ без возвращения и с возвращением соответственно. Пусть $\mathcal{F}$ — некоторое *счетное*[7] множество отображений $f\colon \mathcal{C} \to \mathbb{R}$, таких что $\mathbb{E}[f(\xi_1)] = 0$ и $f(x) \in [-1, 1]$ для всех $f \in \mathcal{F}$ и $x \in \mathcal{C}$. *Супремумами эмпирических процессов* для выборок без возвращения и с возвращением соответственно называются следующие случайные величины:

$$Q'_n = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(\eta_i), \quad Q_n = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(\xi_i).$$

Теорема 20. *Введем обозначение* $\sigma^2_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \mathbb{D}[f(\xi_1)]$. *Тогда для любого* $\varepsilon \geqslant 0$ *справедливо:*

$$\mathsf{P}\left\{Q'_n - \mathbb{E}[Q'_n] \geqslant \varepsilon\right\} \leqslant \exp\left(-\frac{(N+2)\varepsilon^2}{8N^2\sigma^2_{\mathcal{F}}}\right).$$

Аналогичное неравенство справедливо и для $\mathsf{P}\left\{\mathbb{E}[Q'_n] - Q'_n \geqslant \varepsilon\right\}$. *Кроме того, для любого* $\delta \in (0, 1]$ *с вероятностью не меньше* $1 - \delta$ *справедливо:*

$$Q'_n \leqslant \mathbb{E}[Q'_n] + 2\sqrt{2N\sigma^2_{\mathcal{F}} \log\frac{1}{\delta}}.$$

Теорема 22. *Положим* $\sigma^2_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \mathbb{D}[f(\xi_1)]$, $v = n\sigma^2_{\mathcal{F}} + 2\mathbb{E}[Q_n]$ *и для* $u > -1$ *определим* $\phi(u) = e^u - u - 1$, $h(u) = (1+u)\log(1+u) - u$. *Тогда для* $\varepsilon \geqslant 0$ *справедливо:*

$$\mathsf{P}\left\{Q'_n - \mathbb{E}[Q_n] \geqslant \varepsilon\right\} \leqslant \exp\left(-v \cdot h\left(\frac{\varepsilon}{v}\right)\right) \leqslant \exp\left(-\frac{\varepsilon^2}{2v + 2\varepsilon/3}\right).$$

Кроме того, для любого $\delta \in (0, 1]$ *с вероятностью не менее* $1 - \delta$ *справедливо:*

$$Q'_n \leqslant \mathbb{E}[Q_n] + \sqrt{2vt} + \frac{1}{3}\log\frac{1}{\delta}.$$

Подробный теоретический анализ, представленный в конце раздела 2.3, показывает, что во многих интересных случаях новые неравенства Теорем 20 и 22 существенно превосходят точность всех известных в литературе аналогов. Обратим внимание, что Теорема 22 контролирует отклонения случайной величины Q'_n не от своего математического ожидания $\mathbb{E}[Q'_n]$, а от величины $\mathbb{E}[Q_n]$. В разделе 2.3 показано, что этот факт в общем случае может вести к потерям точности оценок, поскольку всегда справедливо $\mathbb{E}[Q_n] - \mathbb{E}[Q'_n] \geqslant 0$. Однако, при определенном соотношении размеров выборок n и N ухудшения оказываются пренебрежимо малыми, поскольку также справедливо $\mathbb{E}[Q_n] - \mathbb{E}[Q'_n] \leqslant 2n^3/N$.

[7] Все результаты могут быть обобщены на случай *несчетного* множества $\mathcal{F}$ в том случае, если эмпирический процесс *сепарабелен*, то есть множество $\mathcal{F}$ содержит счетное и всюду плотное подмножество.

Вторая часть диссертационной работы (**главы 3, 4, 5 и 6**) посвящена *теории статистического обучения*[8]. В **третьей главе** приведен подробный обзор ряда классических и современных результатов *индуктивной* постановки *теории статистического обучения*[9]. **Раздел 3.1** посвящен введению определений и обсуждению различных постановок задач. В **разделе 3.2** приводится подробный обзор и сравнение ряда подходов к получению оценок избыточного риска и обобщающей способности. Сначала в **подразделе 3.2.1** рассмотрены подходы, существенно опирающиеся на *неравенство Буля*, включая оценку Вапника-Червоненкиса, оценку «бритвы Оккама»[10] и оценку, основанную на покрытиях множества отображений[11]. Затем в **подразделе 3.2.2** рассматриваются оценки, основанные на *глобальной Радемахеровской сложности*[12], а также приводятся неравенства симметризации и сжатия. Наконец, в **подразделе 3.2.3** приводится обзор так называемого *локального анализа*[13], включая обсуждение условий ограниченного шума Маммена-Цыбакова и Массара, быстрых скоростей сходимости и *локальных Радемахеровских сложностей*.

В **четвертой главе** рассматривается *трансдуктивная* постановка[8] теории статистического обучения, и в рамках нее впервые применяется локальный анализ, описанный в третьей главе. В **разделе 4.1** приводится формальная постановка задачи. Рассмотрим произвольное конечное множество $\boldsymbol{X}_N$, содержащее N объектов из *пространства объектов* $\mathcal{X}$. Выберем $n \leqslant N$ объектов $\boldsymbol{X}_n \subseteq \boldsymbol{X}_N$ равномерно *без возвращения* из множества $\boldsymbol{X}_N$ и получим ответы $\boldsymbol{Y}_n$ для объектов $\boldsymbol{X}_n$ с помощью неслучайной целевой функции[14] $\varphi\colon \mathcal{X} \to \mathcal{Y}$, где $\mathcal{Y}$ — *пространство ответов*. Обозначим *обучающую выборку* $(\boldsymbol{X}_n, \boldsymbol{Y}_n)$ с помощью S_n. Оставшиеся объекты без ответов на них $\boldsymbol{X}_u = \boldsymbol{X}_N \setminus \boldsymbol{X}_n$, где $u = N - n$, формируют *контрольную выборку*. Рассмотрим произвольную ограниченную и неотрицательную *функцию потерь* $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ и обозначим величину потерь при использовании отображения $h\colon \mathcal{X} \to \mathcal{Y}$ для предсказания ответа на объекте X с помощью $\ell_h(X) = \ell\big(h(X), \phi(X)\big)$. Определим соответственно *эмпирический риск*, *потери на контрольной выборке* и *потери на генеральном*

[8] Vapnik V. N. *Statistical Learning Theory*, John Wiley & Sons, 1998.

[9] Boucheron S., Lugosi G., Bousquet O. *Theory of Classification: a Survey of Recent Advances*, ESAIM: Probability and Statistics, 2005.

[10] Blumer A., Ehrenfueucht A., Haussler D., Warmuth M. *Occam's razor*, Information Processing Letters, 1987, Vol. 24, P. 377–380.

[11] Mendelson S. *A Few Notes on Statistical Learning Theory*, Lecture Notes in Computer Science, 2003.

[12] Bartlett P. L., Mendelson S. *Rademacher and Gaussian complexities: Risk bounds and structural results*, Proceedings of the International Conference on Computational Learning Theory (COLT), 2001.

[13] Koltchinskii V. *Local Rademacher Complexities and Oracle Inequalities in Risk Minimization*, The Annals of Statistics, 2006, Vol. 34, no. 6, p. 2593–2656.

[14] Такая постановка задачи известна как «задача без шума». Результаты четвертой главы могут быть обобщены на случай с шумом, когда ответ Y для объекта $x \in \mathcal{X}$ выбирается из неизвестного условного распределения $P(Y|X = x)$ на множестве $\mathcal{Y}$.

множестве произвольного отображения $h\colon \mathcal{X} \to \mathcal{Y}$ следующим образом:

$$\hat{L}_n(h) = \frac{1}{n} \sum_{X \in \boldsymbol{X}_n} \ell_h(X), \quad L_u(h) = \frac{1}{u} \sum_{X \in \boldsymbol{X}_u} \ell_h(X), \quad L_N(h) = \frac{1}{N} \sum_{X \in \boldsymbol{X}_N} \ell_h(X).$$

Цель *процедуры обучения* заключается в поиске на основе обучающей выборки S_n и объектов контрольной выборки $\boldsymbol{X}_u$ отображения h_u^* в заранее выбранном классе отображений $\mathcal{H} = \{h\colon \mathcal{X} \to \mathcal{Y}\}$ (не обязательно содержащем целевую функцию φ), имеющего минимальные потери на контрольной выборке:

$$h_u^* \in \operatorname{Arg}\min_{h \in \mathcal{H}} L_u(h).$$

Поскольку ошибка на контроле $L_u(h)$ нам неизвестна, мы воспользуемся методом *минимизации эмпирического риска* (МЭР) и будем приближать оптимальное отображение h_u^* отображением $\hat{h}_n \in \mathcal{H}$, имеющим наименьший в заданном классе $\mathcal{H}$ эмпирический риск:

$$\hat{h}_n \in \operatorname{Arg}\min_{h \in \mathcal{H}} \hat{L}_n(h).$$

Найдя такое отображение $\hat{h}_n$, мы хотели бы оценить потери на контроле найденного отображения $L_u(\hat{h}_n)$ и понять, насколько эта величина превосходит оптимальное значение $L_u(h_u^*)$. В частности, нас будут интересовать *обобщающая способность* $L_u(\hat{h}_n) - \hat{L}_n(\hat{h}_n)$ метода МЭР и *избыточный риск* $L_u(\hat{h}_n) - L_u(h_u^*)$ отображения $\hat{h}_n$. Поскольку эти величины являются случайными (так как отображения h_u^* и $\hat{h}_n$ зависят от обучающей выборки), нашей главной задачей[15] является получение верхних оценок на вероятности отклонений $\mathsf{P}\{L_u(\hat{h}_n) - \hat{L}_n(\hat{h}_n) \geqslant t\}$ и $\mathsf{P}\{L_u(\hat{h}_n) - L_u(h_u^*) \geqslant t\}$ для $t \geq 0$, убывающих с ростом размера обучающей выборки n и учитывающих свойства семейства $\mathcal{H}$. В **разделе 4.1** приводится обзор известных результатов.

В **разделе 4.2** приводятся основные результаты четвертой главы: новые оценки избыточного риска (Теоремы 41 и 42, Следствия 13 и 14), основанные на *локальных мерах сложности* семейства отображений $\mathcal{H}$. Результаты приводятся в двух вариантах в зависимости от того, какая из Теорем 20 и 22 использовалась в их доказательстве.

Обозначим с помощью h_N^* отображение из $\mathcal{H}$ с наименьшим значением $L_N(h)$ и для $r \geqslant 0$ рассмотрим отображения $h \in \mathcal{H}$ с малыми дисперсиями

[15]Отметим, что *индуктивная* постановка отличается от трансдуктивной главным образом способом получения обучающей выборки. В индуктивной постановке объекты обучающей выборки $\{(X_i, Y_i)\}_{i=1}^n$ выбираются *независимо* из неизвестного распределения P на декартовом произведении $\mathcal{X} \times \mathcal{Y}$. Задача заключается в поиске отображения $h \in \mathcal{H}$ с наименьшим значением *среднего риска* $\mathbb{E}_{(X,Y)\sim P}\big[\ell\big(h(X), Y\big)\big]$. Кроме того, процедура обучения индуктивной постановки не получает доступа к объектам, не содержащимся в обучающей выборке.

избыточных потерь:

$$\mathcal{H}(r) = \left\{ h \in \mathcal{H} \colon \frac{1}{N} \sum_{X \in \boldsymbol{X}_N} \big(\ell_h(X) - \ell_{h_N^*}(X)\big)^2 \leqslant r \right\}.$$

Субкоренной мы будем называть неубывающую и неотрицательную функцию $\psi \colon [0, \infty) \to [0, \infty)$, такую что отображение $r \to \psi(r)/\sqrt{r}$ является невозрастающим для $r > 0$. Можно показать, что у любой субкоренной функции существует единственная положительная неподвижная точка.

Следующие результаты оценивают сверху величину $L_N(\hat{h}_n) - L_N(h_N^*)$ в терминах *локальной меры сложности* класса $\mathcal{H}$:

Теорема 41. *Пусть существует константа $B > 0$, такая что для всех $h \in \mathcal{H}$:*

$$\frac{1}{N} \sum_{X \in \boldsymbol{X}_N} \big(\ell_h(X) - \ell_{h_N^*}(X)\big)^2 \leqslant B \cdot \big(L_N(h) - L_N(h^*)\big).$$

Кроме того, пусть существует субкоренная функция $\psi_n(r)$, такая что

$$B \cdot \mathbb{E}\left[\sup_{h \in \mathcal{H}(r)} L_N(h) - \hat{L}_n(h) - \big(L_N(h_N^*) - \hat{L}_n(h_N^*)\big) \right] \leqslant \psi_n(r).$$

Обозначим с помощью r_n^* неподвижную точку функции $\psi_n(r)$. Тогда для любого $\delta \in (0, 1]$ с вероятностью не меньше $1 - \delta$ справедливо:

$$L_N(\hat{h}_n) - L_N(h_N^*) \leqslant 51 \frac{r_n^*}{B} + 17B \left(\frac{N}{n^2} \right) \log \frac{1}{\delta}.$$

Теорема 42. *Пусть существует константа $B > 0$, такая что для всех $h \in \mathcal{H}$:*

$$\frac{1}{N} \sum_{X \in \boldsymbol{X}_N} \big(\ell_h(X) - \ell_{h_N^*}(X)\big)^2 \leqslant B \cdot \big(L_N(h) - L_N(h^*)\big).$$

Кроме того, пусть существует субкоренная функция $\tilde{\psi}_n(r)$, такая что:

$$B \cdot \mathbb{E}\left[\sup_{h \in \mathcal{H}(r)} L_N(h) - \tilde{L}_n(h) - \big(L_N(h_N^*) - \tilde{L}_n(h_N^*)\big) \right] \leqslant \tilde{\psi}_n(r),$$

где $\tilde{L}_n(h) = \frac{1}{n} \sum_{i=1}^n \ell_h(\xi_i)$ для случайных величин $\xi_1, \ldots, \xi_n$, выбранных равномерно с возвращением *из $\boldsymbol{X}_N$. Обозначим с помощью $\tilde{r}_n^*$ неподвижную точку функции $\tilde{\psi}_n(r)$. Тогда для любого $\delta \in (0, 1]$ с вероятностью не меньше $1 - \delta$ справедливо:*

$$L_N(\hat{h}_n) - L_N(h_N^*) \leqslant 901 \frac{\tilde{r}_n^*}{B} + \frac{(16 + 25B)}{3n} \log \frac{1}{\delta}.$$

Теоремы 41 и 42 обобщают результаты локального анализа индуктивной постановки теории статистического обучения[16] на трансдуктивную постановку и показывают, что порядок убывания $L_N(\hat{h}_n) - L_N(h_N^*)$ определяется величиной неподвижной точки *модуля непрерывности эмпирического процесса*, посчитанного в окрестности лучшего на генеральном множестве отображения h_N^*. Наконец, справедливы следующие оценки избыточного риска:

Следствие 13. *Пусть выполнены предположения Теоремы 41. Тогда для любого* $\delta \in (0, 1]$ *с вероятностью не меньше* $1 - \delta$ *справедливо:*

$$L_u(\hat{h}_n) - L_u(h_u^*) \leqslant \frac{51N}{B}\left(\frac{r_n^*}{u} + \frac{r_u^*}{n}\right) + \frac{17BN^2}{nu}\left(\frac{1}{u} + \frac{1}{n}\right)\log\frac{2}{\delta}.$$

Следствие 14. *Пусть выполнены предположения Теоремы 42. Тогда для любого* $\delta \in (0, 1]$ *с вероятностью не меньше* $1 - \delta$ *справедливо:*

$$L_u(\hat{h}_n) - L_u(h_u^*) \leqslant \frac{901N}{B}\left(\frac{\tilde{r}_n^*}{u} + \frac{\tilde{r}_u^*}{n}\right) + \frac{(32 + 50B)N}{3nu}\log\frac{2}{\delta}.$$

Оказывается, в ряде интересных случаев неподвижные точки r_n^* и $\tilde{r}_n^*$ имеют порядок $o(n^{-1/2})$. Примером являются задачи с бинарной функцией потерь и множеством отображений $\mathcal{H}$, имеющем конечную VC-размерность $d < \infty$. Известно[17], что в этом случае $r_n^* = O\big(\frac{d\log n}{n}\big)$. В теории статистического обучения скорости сходимости оценок принято разделять на два типа: *быстрые скорости* порядка $o(n^{-1/2})$ и *медленные скорости* порядка $O(n^{-1/2})$. Новые результаты впервые позволяют получать оценки избыточного риска в трансдуктивном обучении, имеющие быстрые скорости сходимости в общих предположениях на рассматриваемый класс задач.

В разделе 4.2 также приводятся новые оценки обобщающей способности, которые, в отличие от известных в литературе аналогов, учитывают дисперсию потерь и справедливы для *произвольных* классов отображений и ограниченных функций потерь. Подробные доказательства результатов приводятся в **разделе 4.3**.

В **пятой главе** рассматривается *комбинаторная теория переобучения*[18], тесно связанная с трансдуктивным обучением. В **разделе 5.1** вводятся определения и ставится формальная постановка задачи. Несмотря на то, что постановка комбинаторной теории переобучения чрезвычайно похожа на введенную ранее постановку трансдуктивного обучения, нам будет удобно ввести ряд новых

[16]Bartlett P., Bousquet O., Mendelson S. *Local Rademacher Complexities*, The Annals of Statistics, 2005, Vol. 33, no. 4, p. 1497–1537.

[17]Massart P. *Some applications of concentration inequalities to statistics*, Ann. Fac. Sci. Toulouse Math, 2000, Vol. 9, no. 6, P. 245–303.

[18]Воронцов К. В. *Комбинаторная теория переобучения: результаты, приложения и открытые проблемы*, Математические методы распознавания образов: 15-ая Всеросс. конф.: Докл, 2011.

обозначений. Мы также продолжим пользоваться обозначениями, введенными в четвертой главе.

Рассмотрим *конечную* генеральную выборку *упорядоченных* пар «объект-ответ» $\mathbb{S} = \{(X_i, Y_i)\}_{i=1}^N$ и *бинарную* функцию потерь. Конечное множество *различных* векторов ошибок отображений из $\mathcal{H}$ на генеральной выборке $\mathbb{S}$ мы будем обозначать $\mathbb{A}$:

$$\mathbb{A} = \left\{ \left(\mathbb{1}_{\{h(X_1) \neq Y_1\}}, \ldots, \mathbb{1}_{\{h(X_1) \neq Y_1\}} \right) \colon h \in \mathcal{H} \right\} \subseteq \{0,1\}^N.$$

Всюду далее мы будем отождествлять отображения $h \in \mathcal{H}$ с их векторами ошибок $a \in \mathbb{A}$. *Индикатором ошибки* вектора $a \in \mathbb{A}$ на паре $(X, Y) \in \mathbb{S}$ назовем $I_a(X, Y) = \mathbb{1}_{\{h_a(X) \neq Y\}}$, где h_a — произвольное отображение из $\mathcal{H}$, вектор ошибок которого равен a. Из генеральной выборки $\mathbb{S}$ *равномерно без возвращения* выбирается $n \leqslant N$ пар «объект-ответ» $S_n \subseteq \mathbb{S}$, которые образуют *обучающую выборку*, и становятся доступны алгоритму обучения. Оставшиеся $u = N - n$ пар S_u остаются неизвестными и образуют *контрольную выборку*. Определим среднее число ошибок вектора $a \in \mathbb{A}$ на подвыборке $S \subseteq \mathbb{S}$:

$$L(a, S) := \frac{1}{|S|} \sum_{(X,Y) \in S} I_a(X, Y).$$

Для произвольной подвыборки $S \subseteq \mathbb{S}$ обозначим с помощью $A(S)$ множество векторов в $\mathbb{A}$, имеющих минимальное число ошибок на ней:

$$A(S) = \operatorname{Arg}\min_{a \in \mathbb{A}} L(a, S).$$

Также обозначим с помощью $[\mathbb{S}]^n$ всевозможные подвыборки $\mathbb{S}$, состоящие в точности из n пар «объект-ответ».

В комбинаторной теории переобучения рассматриваются различные *методы обучения* — отображения $\mu \colon [\mathbb{S}]^n \to \mathbb{A}$, ставящие в соответствие обучающим выборкам векторы ошибок из $\mathbb{A}$. Основная задача комбинаторной теории переобучения заключается в получении *точных* (не завышенных) оценок *вероятности переобучения* метода μ:

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \mathsf{P}\left\{ L\big(\mu(S_n), S_u\big) - L\big(\mu(S_n), S_n\big) \geqslant \varepsilon \right\}.$$

В диссертационной работе рассматривается *рандомизированный метод минимизации эмпирического риска* (РМЭР), который для заданной обучающей выборки S_n выбирает случайный вектор a равномерно из множества $A(S_n)$. Обратим внимание, что в этом случае вектор $\mu(S_n)$ в определении вероятности переобучения $Q_{\mu,\varepsilon}(\mathbb{A})$ также становится случайным. В разделе 5.1 также приводится сравнение постановки кобинаторной теории переобучения с постановкой трансдуктивного обучения.

В **разделе 5.2** рассматривается *теоретико-групповой подход* к комбинаторной теории переобучения, и приводятся основные результаты пятой главы. В **подразделе 5.2.1** вводятся определения и обозначения, а также приводится обзор известных результатов. Рассмотрим симметрическую группу перестановок Π_N. Элементы группы Π_N очевидным образом действует на генеральной выборке, переставляя местами пронумерованные пары «объект-ответ», составляющие ее. Для пары $(X, Y) \in \mathbb{S}$ и $\pi \in \Pi_N$ с помощью $\pi(X, Y)$ обозначим ту пару, в которую пара (X, Y) переходит под действием перестановки π. На основе этого можно определить действие элементов $\pi \in \Pi_N$ на множестве подвыборок $S \subseteq \mathbb{S}$:

$$\pi(S) = \{\pi(X, Y) \colon (X, Y) \in S\},$$

на множестве бинарных векторов ошибок $a \in \mathbb{A}$:

$$\pi(a) = \big(I_{\pi(a)}(X_i, Y_i))\big)_{i=1}^N = \Big(I_a\big(\pi^{-1}(X_i, Y_i)\big)\Big)_{i=1}^N$$

и на всевозможных подмножествах $A \subseteq \mathbb{A}$:

$$\pi(A) = \{\pi(a) \colon a \in \mathbb{A}\}.$$

Группой симметрий $S_{\mathbb{A}}$ множества векторов ошибок $\mathbb{A} \subseteq \{0, 1\}^N$ будем называть его стационарную подгруппу:

$$S_{\mathbb{A}} = \{\pi \in \Pi_N : \pi(\mathbb{A}) = \mathbb{A}\}.$$

Орбитой действия группы симметрий $S_{\mathbb{A}}$ *на вектор ошибок* $a \in \mathbb{A}$ будем называть множество $\{\pi(a) : \pi \in S_{\mathbb{A}}\}$. Множество всех орбит действия группы $S_{\mathbb{A}}$ на элементы $\mathbb{A}$ будем обозначать $\Omega(\mathbb{A})$.

Теоретико-групповой подход основан на следующем наблюдении: в том случае, когда семейство векторов $\mathbb{A}$ обладает определенными симметриями (когда группа $S_{\mathbb{A}}$ содержит достаточное число элементов), вычисление точного значения вероятности переобучения существенно упрощается[19].

В **подразделе 5.2.2** предлагается при вычислении вероятности переобучения помимо орбит векторов ошибок учитывать также *орбиты выборок*, и приводится новая формула вычисления точного значения вероятности переобучения. *Орбитой действия группы симметрий* $S_{\mathbb{A}}$ *на подвыборку* $S \in [\mathbb{S}]^n$ мы будем называть множество $\{\pi(S) \colon \pi \in S_{\mathbb{A}}\}$. Множество всех таких орбит мы будем обозначать $\Omega\left([\mathbb{S}]^n\right)$.

[19] Фрей А. И. *Точные оценки вероятности переобучения для симметричных семейств алгоритмов*, Математические методы распознавания образов: 14-ая Всеросс. конф.: Докл, 2009.

Теорема 44. *При использовании РМЭР для произвольного множества попарно различных векторов ошибок $\mathbb{A} \in \{0,1\}^N$ и любых $\varepsilon \geqslant 0$ справедливо:*

$$Q_{\mu,\varepsilon}(\mathbb{A}) = \sum_{\omega\in\Omega(\mathbb{A})} \frac{|\omega|}{C_N^n} \sum_{\tau\in\Omega([\mathbb{S}]^n)} \frac{|\{S\in\tau\colon a_\omega \in A(S)\}|}{|A(S_\tau)|} \mathbb{1}_{\{L(a_\omega,\mathbb{S}\setminus S_\tau)-L(a_\omega,S_\tau)\geqslant\varepsilon\}},$$

где a_ω — произвольный вектор ошибок из орбиты $\omega \in \Omega(\mathbb{A})$, а S_τ — произвольная выборка из орбиты $\tau \in \Omega([\mathbb{S}]^n)$. Результат остается справедлив, если мы заменим множества $\Omega(\mathbb{A})$ и $\Omega([\mathbb{S}]^n)$ соответствующими множествами орбит действия произвольной подгруппы группы симметрий $S_\mathbb{A}$.

В **подразделе 5.2.3** обсуждаются свойства *сходства* и *расслоения* семейств векторов ошибок и их влияние на вероятность переобучения[20].

Число ошибок вектора a на подвыборке $S \in \mathbb{S}$ мы будем обозначать $n(a,S) = |S|\cdot L(a,S)$. Рассмотрим множество бинарных векторов, допускающих ровно m ошибок на генеральной выборке: $\mathbb{B}_m^N = \{a \in \{0,1\}^N\colon n(a,\mathbb{S}) = m\}$. Введем расстояние Хэмминга между двумя векторами $a', a'' \in \mathbb{A}$:

$$d(a',a'') = \sum_{(X,Y)\in\mathbb{S}} |I_{a'}(X,Y) - I_{a''}(X,Y)|.$$

На основе Теоремы 44 в **подразделе 5.2.4** получены новые точные (не завышенные) оценки вероятности переобучения для следующих трех модельных семейств векторов ошибок $\mathbb{A}$:

1. *Шар* радиуса r с центром в векторе $a_0 \in \{0,1\}^N$:

$$B_r(a_0) = \{a \in \{0,1\}^N\colon d(a,a_0) \leqslant r\}.$$

2. *Центральный слой шара*:

$$B_r^c(a_0) = \mathbb{B}_m^N \cap B_r(a_0),$$

   где $m = n(a_0,\mathbb{S})$.

3. *d нижних слоёв шара*:

$$B_r(a_0,d) = \bigcup_{j=1,\ldots,d} \left(\mathbb{B}_{m_j}^N \cap B_r(a_0)\right),$$

   где $m_j = \max\{n(a_0,\mathbb{S}) - r, 0\} + (d-1)$.

Новые оценки еще раз подтверждают хорошо известный в комбинаторной теории переобучения факт о том, что получение не сильно завышенных оценок невозможно без *одновременного* учета свойств сходства и расслоения множества векторов ошибок.

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

Шестая глава посвящена *PAC-Байесовскому анализу*[21], который в рамках теории статистического обучения изучает рандомизированные отображения. В **разделах 6.1** и **6.2** вводятся основные определения, описывается общий подход к получению PAC-Байесовских неравенств, и приводится достаточно подробный обзор известных результатов, включая неравенство МакАллистера (также известное как PAC-Байесовское неравенство Хефдинга), PAC-Байесовское неравенство Бернштейна и PAC-Байесовское kl-неравенство. Также в этих разделах приводится подробное сравнение трех описанных неравенств, и обсуждаются способы их применения в индуктивном обучении, рассмотренном в третьей главе. В **разделе 6.3** представлены главные результаты шеской главы (Теоремы 54 и 55, а также результаты экспериментов).

Напомним, что в рамках индуктивной постановки теории статистического обучения нам дана обучающая выборка $S_n = \{(X_i, Y_i)\}_{i=1}^n$, которая состоит из n пар «объект-ответ», выбранных независимо из неизвестного распределения P на $\mathcal{X} \times \mathcal{Y}$. Задача процедуры обучения заключается в поиске на основе S_n отображения $h \in \mathcal{H}$, имеющего малый *средний риск* $L(h) = \mathbb{E}_{(X,Y)\sim P}\big[\ell\big(h(X), Y\big)\big]$ для неотрицательной и ограниченной функции потерь $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. Обозначим с помощью $L_n(h) = \frac{1}{n}\sum_{i=1}^n \ell\big(h(X_i), Y_i\big)$ *эмпирический риск* отображения $h \in \mathcal{H}$.

Рассмотрим произвольное вероятностное распределение ρ на множестве отображений $\mathcal{H}$. PAC-Байесовский анализ изучает *рандомизированное отображение* G_ρ, которое для получения ответа на новом объекте $X \in \mathcal{X}$ выбирает случайное отображение h из распределения ρ (независимо от объекта X) и возвращает ответ $h(X)$. *Средний* и *эмпирический риск* рандомизированного отображения G_ρ определяется следующим образом:

$$L(G_\rho) = \mathbb{E}_{h\sim\rho}[L(h)], \quad L_n(G_\rho) = \mathbb{E}_{h\sim\rho}\left[L_n(h)\right].$$

Задача PAC-Байесовского обучения заключается в поиске распределения ρ с малым значением среднего риска $L(G_\rho)$. PAC-Байесовские неравенства позволяют оценивать неизвестный средний риск стохастического отображения $L(G_\rho)$ с помощью эмпирического риска $L_n(G_\rho)$. Кроме того, в ряде случаев при определенном выборе распределения ρ величина $L(G_\rho)$ позволяет оценивать сверху средний риск $L(h)$ обыкновенных детерминированных отображений[22].

[21] McAllester D. *Some PAC-Bayesian Theorems*, Proceedings of the International Conference on Computational Learning Theory (COLT), 1998.

[22] Langford J., Shawe-Taylor J. *PAC-Bayes & Margins*, Advances in Neural Information Processing Systems (NIPS), 2002.

Одним из наиболее точных на сегодняшний день PAC-Байесовских неравенств является PAC-Байесовское неравенство Бернштейна[23]. Однако, его оценка зависит от неизвестной усредненной дисперсии потерь $\mathbb{E}_{h\sim\rho}[\mathbb{D}(h)]$, где $\mathbb{D}(h) = \mathbb{D}_{(X,Y)\sim P}\big[\ell\big(h(X),Y\big)\big]$.

В **подразделе 6.3.1** представлен новый результат, позволяющий оценивать неизвестную величину $\mathbb{E}_{h\sim\rho}[\mathbb{D}(h)]$ с помощью несмещенной выборочной оценки дисперсии:

$$\mathbb{D}_n(h) = \frac{1}{n-1}\sum_{i=1}^{n}\big(\ell_h(X_i,Y_i) - L_n(h)\big)^2.$$

Обозначим дивергенцию Кульбака-Лейблера для двух вероятностных распределений ρ и π на $\mathcal{H}$ с помощью $\mathrm{KL}(\rho,\pi) = \mathbb{E}_{h\sim\rho}\left[\log\frac{\rho(h)}{\pi(h)}\right]$.

Теорема 54. *Для любого распределения π на множестве $\mathcal{H}$, не зависящего от обучающей выборки, для любого $\delta\in(0,1]$ и для любого $c>1$ с вероятностью не меньше $1-\delta$ (относительно случайной реализации обучающей выборки) следующее справедливо одновременно для всех распределений ρ на $\mathcal{H}$:*

$$\mathbb{E}_{h\sim\rho}[\mathbb{D}(h)] \le c\,\mathbb{E}_{h\sim\rho}[\mathbb{D}_n(h)] + 2c\sqrt{\big(\mathbb{E}_{h\sim\rho}[\mathbb{D}_n(h)] + \eta_n(\rho)\big)\eta_n(\rho)} + 2c\,\eta_n(\rho), \quad (1)$$

где

$$\eta_n(\rho) = \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{\nu}{\delta}}{2(n-1)}, \quad \nu = \left\lceil \frac{1}{\ln c}\ln\left(\frac{1}{2}\sqrt{\frac{n-1}{\ln 1/\delta}+1}+\frac{1}{2}\right)\right\rceil.$$

В **подразделе 6.3.2** представлено новое *PAC-Байесовское эмпирическое неравенство Бернштейна*, основанное на Теореме 54 и PAC-Байесовском неравенстве Бернштейна:

Теорема 55. *Для любого распределения π на $\mathcal{H}$, не зависящего от обучающей выборки, любых $\delta_1\in(0,1]$ и $c_1,c_2>1$, обозначив $\hat{\mathbb{D}}_n(\rho)$ правую часть неравенства* (1) *(с $\delta=\frac{\delta_1}{2}$) и положив $\bar{\mathbb{D}}_n(\rho) = \min\left(\hat{\mathbb{D}}_n(\rho), \frac{1}{4}\right)$, мы получим, что с вероятностью не меньше $1-\delta_1$ (относительно случайной реализации обучающей выборки) следующее:*

$$L(G_\rho) \leqslant L_n(G_\rho) + (1+c_1)\sqrt{\frac{(e-2)\bar{\mathbb{D}}_n(\rho)\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{2\nu_1}{\delta_1}\right)}{n}}$$

выполнено одновременно для всех распределений ρ на $\mathcal{H}$, которые удовлетворяют условию

$$\sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{2\nu_1}{\delta_1}}{(e-2)\bar{\mathbb{D}}_n(\rho)}} \leqslant \sqrt{n},$$

[23] Seldin Y., Laviolette F., Cesa-Bianchi N., Shawe-Taylor J., Auer P. *PAC-Bayesian Inequalities for Martingales*, IEEE Transactions on Information Theory, 2012, Vol. 58.

где $\nu_1 = \left\lceil \frac{1}{\ln c_1} \ln \left(\sqrt{\frac{(e-2)n}{4\ln(2/\delta_1)}} \right) \right\rceil + 1$. *Для всех остальных распределений* ρ *одновременно справедливо:*

$$L(G_\rho) \leqslant L_n(G_\rho) + 2\frac{\mathrm{KL}(\rho\|\pi) + \ln \frac{2\nu_1}{\delta_1}}{n}.$$

Обратим внимание, что оценка последнего результата полностью вычислима на основе обучающей выборки. Кроме того, теоретический анализ показывает, что во многих случаях она существенно точнее известных в литературе аналогов. В **подразделе 6.3.2** приводятся численные эксперименты с модельными выборками и реальными данными из репозитория UCI, результаты которых согласуются с представленным теоретическим анализом.

Заключение

Основные результаты диссертационной работы заключаются в следующем:

1. Получено два новых неравенства концентрации для супремумов эмпирических процессов и выборок без возвращения. Оба неравенства учитывают дисперсию случайных величин, что отличает их от всех известных ранее аналогов. Подробный теоретический анализ показывает, что во многих интересных случаях новые неравенства существенно превосходят по точности все известные аналоги.
2. Получены новые неравенства обобщающей способности для трансдуктивного обучения, учитывающие дисперсии потерь и, в отличие от известных в литературе аналогов, справедливые для произвольных классов отображений и огрониченных функций потерь. Также получены первые оценки избыточного риска в трансдуктивном обучении, основанные на локальных мерах сложности семейств отображений и во многих интересных случаях имеющие быстрые скорости сходимости в общих предположениях о рассматриваемом классе задач.
3. В рамках теоретико-группового подхода комбинатоной теории переобучения предложено учитывать орбиты действия группы симметрий множества векторов ошибок на подвыборки генеральной совокупности. На основе этой идеи получена новая формула разложения вероятности переобучения по орбитам разбиений. По сравнению с известной до этого формулой разложения по орбитам векторов ошибок новая формула позволяет существенно упростить вычисление вероятности переобуения для ряда модельных множеств векторов ошибок. В частности, на ее основе получены новые точные (не заышенные) оценки вероятности переобучения для трех модельных семейств, являющихся различными подмножествами шара в Булевом кубе. Новые оценки еще раз

подтверждают хорошо известный в комбинаторной теории переобучения факт о том, что получение не сильно завышенных оценок невозможно без тщательного учета геометрических и структурных свойств множества бинарных векторов ошибок.

4. Получено новое PAC-Байесовское эмпирическое неравенство Бернштейна, полностью вычислимое на основе наблюдаемых данных и во многих случаях ведущее к существенно более точным оценкам по сравнению с известными ранее результатами. Доказательство неравенства основано на новом PAC-Байесовском неравенстве для дисперсии, которое позволяет оценить сверху неизвестное усредненное значение дисперсии потерь с помощью усредненного значения выборочной дисперсии равномерно по классу всех усредняющих распределений. Результаты серии экспериментов, использующих модельные выборки и реальные выборки из репозитория UCI, согласуются с теоретически обоснованным выводом о превосходстве нового неравенства над известными аналогами.

Публикации автора по теме диссертации

Публикации из списка ВАК

1. Tolstikhin I., Seldin Y. PAC-Bayes-Empirical-Bernstein Inequality // Advances in Neural Information Processing Systems (NIPS). 2013.

2. Фрей А. И., Толстихин И. О. Комбинаторные оценки вероятности переобучения на основе покрытий множества алгоритмов // Доклады РАН. 2014. Т. 455, № 3. С. 265–268.

3. Tolstikhin I., Blanchard G., Kloft M. Localized Complexities for Transductive Learning // Proceedings of the 27th Annual Conference on Learning Theory (COLT 2014), JMLR W&CP. 2014. P. 857–884.

Прочие публикации

4. Толстихин И. О. Точная оценка вероятности переобучения для одного специального семейства алгоритмов // Межд. науч. конф. студентов, аспирантов и молодых ученых «Ломоносов 2010»: Докл. 2010. С. 54–57.

5. Толстихин И. О. Вероятность переобучения плотных и разреженных семейств алгоритмов // Межд. конф. Интеллектуализация обработки информации ИОИ-8: Докл. 2010. С. 83–86.

6. Толстихин И. О. Локализация оценок избыточного риска в комбинаторной теории переобучения // Межд. конф. Интеллектуализация обработки информации ИОИ-9: Докл. 2012. С. 54–57.

7. Frey A., Tolstikhin I. Combinatorial bounds on probability of overfitting based on clustering and coverage of classifiers // Machine Learning and Data Analysis (JMLDA). 2013. Vol. 1, no. 6. P. 761–778.