УДК 519.7:004.855.5

ОБЗОР СОВРЕМЕННЫХ ИССЛЕДОВАНИЙ ПО ПРОБЛЕМЕ КАЧЕСТВА ОБУЧЕНИЯ АЛГОРИТМОВ

К. В. Воронцов

Вычислительный центр имени А. А. Дородницына РАН
РФ, г.Москва, ГСП-1, ул. Вавилова, 40, 119991
e-mail: *voron@ccas.ru*

Abstract

The review considers basic ideas of machine learning theory concerning generalization bounds and learning algorithms grounds. Among them are: classical VC theory and structural risk minimization, effective VC-dimension and data-dependent bounds, margin, ensembles of algorithms (weighted voting, boosting and bagging), stability, cross-validation. A new combinatorial approach to proving non-probabilistic generalization bounds is considered a little more detailed.

Вопрос о качестве алгоритмов, синтезированных по конечным выборкам прецедентов, является фундаментальной проблемой теории обучаемых систем (machine learning theory).

В общем случае задача обучения по прецедентам заключается в том, чтобы по заданной выборке пар «объект–ответ» восстановить функциональную зависимость между объектами и ответами, то есть построить алгоритм, способный выдавать адекватные ответы на предъявляемые объекты. Когда множество допустимых ответов конечно, говорят о задачах *классификации* или *распознавания образов*. Когда множество допустимых ответов бесконечно, например, является множеством действительных чисел или векторов, говорят о задачах *восстановления регрессии*. Когда объекты соответствуют моментам времени, а ответы характеризуют будущее поведение процесса или явления, говорят о задачах *прогнозирования*.

Значительный опыт решения прикладных задач такого типа был накоплен уже к середине 60-х годов XX века. Большую популярность приобрёл подход, основанный на построении модели восстанавливаемой зависимости в виде параметрического семейства алгоритмов. С помощью численной оптимизации в семействе выбирался алгоритм, допускающий наименьшее число ошибок на заданной обучающей выборке. Проще говоря, осуществлялась подгонка (fitting) модели под выборку. Функционал частоты ошибок или средней ошибки алгоритма на обучающей выборке принято называть *эмпирическим риском*, а сам подход — *минимизацией эмпирического риска*.

На практике исследователи столкнулись с эффектом, называемым *переобучением или переподгонкой* (overtraining, overfitting). Чем больше у алгоритма

свободных параметров, тем меньшего числа ошибок на обучении можно добиться путём оптимизации. Однако по мере нарастания сложности модели «оптимальные» алгоритмы начинают слишком хорошо подстраиваться под конкретные данные, улавливая не только черты восстанавливаемой зависимости, но и ошибки измерения обучающей выборки, и погрешность самой модели. В результате ухудшается качество работы алгоритма вне обучающей выборки, или, как говорят, его *способность к обобщению* (generalization performance).

Из этого наблюдения был сделан вывод, что для всякой задачи существует оптимальная сложность модели, при которой достигается наилучшее качество обобщения. Первым формальным обоснованием этого практического опыта стала статистическая теория восстановления зависимостей по эмпирическим данным, разработанная В. Н. Вапником и А. Я. Червоненкисом в конце 60-х — начале 70-х годов [1, 2, 3].

1. Статистическая теория Вапника-Червоненкиса

В статистической теории предполагается, что на множестве объектов существует некоторое (неизвестное) распределение вероятностей, и обучающая совокупность объектов выбирается случайно и независимо в соответствии с данным распределением. Предполагается также, что алгоритм, минимизирующий эмпирический риск, ищется в некотором заранее фиксированном семействе алгоритмов. Оно может содержать множество алгоритмов, доставляющих минимум эмпирическому риску, однако в статистической теории способ построения алгоритма (метод обучения) не рассматривается и предполагается, что в качестве решения может быть выдан любой алгоритм из этого множества.

Обобщающая способность определяется как вероятность ошибки найденного алгоритма, либо как частота его ошибок на неизвестной контрольной выборке, также случайной, независимой и одинаково распределённой.

Далее постулируется принцип *равномерной сходимости* (uniform convergence) частоты ошибок. Чтобы по частоте ошибок найденного алгоритма на обучающей выборке можно было судить о частоте его ошибок на любой другой выборке, эти частоты должны стремиться друг к другу с ростом длины выборки, причём одновременно (равномерно) по всему семейству алгоритмов. Оценки качества обучения в статистической теории являются, по сути дела, оценками скорости этой сходимости. Именно принцип равномерной сходимости и приводит к введению специальной меры сложности семейства алгоритмов, называемой *ёмкостью* или размерностью Вапника-Червоненкиса (VC-dimension).

Получение оценок ёмкости для конкретных семейств алгоритмов является отдельной, зачастую довольно трудной, задачей. Практически сразу было доказано, что ёмкость семейства линейных решающих правил равна числу свободных параметров или, что то же самое, размерности линейного пространства, в котором строится разделяющая гиперплоскость. Оценки ёмкости получены также для нейронных сетей [30, 26, 54, 67], решающих деревьев [10], корректных алгебраических замыканий подмодели ABO [15], комитетных решающих правил [65], и других семейств.

Основным результатом статистической теории являются количественные оценки, связывающие надёжность алгоритмов с длиной обучающей выборки и сложностью семейства. Эти оценки позволяют обосновать метод *структурной минимизации риска* (СМР), непосредственно направленный на выбор модели оптимальной сложности. В СМР фиксируется определённая структура вложенных подсемейств различной сложности, затем в каждом подсемействе решается задача обучения по прецедентам, и из полученных алгоритмов выбирается тот, для которого оценка качества принимает наилучшее значение.

К сожалению, оценки Вапника-Червоненкиса сильно завышены, что приводит к требованию слишком длинных обучающих выборок (10^5–10^6 объектов), а в методе структурной минимизации риска — к чрезмерному упрощению алгоритмов [55]. Некоторые семейства имеют бесконечную ёмкость и находятся за границами применимости теории, тем не менее с их помощью удаётся решать прикладные задачи, и довольно успешно. В частности, это относится к метрическим методам, основанным на явном хранении обучающей выборки, таким как метод ближайших соседей, а также к методам алгебраического подхода [12, 6], гарантирующим безошибочное распознавание заданной выборки. На практике качество обучения почти всегда оказывается существенно лучше, чем предсказывает статистическая теория.

Причина завышенности статистических оценок кроется в их слишком большой общности. Они ориентированы на «худший случай» и не учитывают трёх важных особенностей самой задачи и процесса поиска её решения, которые могут оказывать решающее влияние на качество обучения.

Во-первых, это особенности распределения объектов в пространстве. В частности, они могут лежать в подпространстве меньшей размерности. Этот «вырожденный» случай довольно распространён, поскольку в прикладных задачах наличие зависимых или почти зависимых признаков является скорее правилом, чем исключением.

Во-вторых, это особенности самой восстанавливаемой зависимости. Она может быть гладкой, симметричной, монотонной или обладать другими специальными свойствами, что резко сужает пространство поиска решения.

В-третьих, это особенности метода обучения. Он может подстраиваться под задачу, образуя эффективное подсемейство алгоритмов, реально получаемых в результате обучения.

Появление статистической теории вызвало большое количество исследований, направленных на уточнение оценок качества. Однако проблема получения численных оценок, непосредственно применимых на практике, оказалась вызывающе трудной, и до сих пор остаётся открытой.

Далее будут перечислены некоторые направления современных исследований по проблемам обоснования обучаемых алгоритмов и получения оценок качества обучения. Разумеется, предлагаемая классификация весьма условна и не претендует на полноту.

2. Эффективная сложность

Первое направление связано с понятием *эффективной сложности*. При решении конкретной задачи далеко не каждый алгоритм из выбранного семейства имеет шансы быть полученным в результате обучения. Как правило, реально работает не всё семейство, а лишь небольшая его часть. Этот факт был замечен ещё В. Н. Вапником, предложившим понятие эффективной ёмкости вместе с алгоритмом её практического измерения [80, 33]. Эффективная ёмкость не превосходит полной ёмкости семейства и зависит от выборки. Она учитывает особенности исходного распределения объектов, но не принимает во внимание особенностей восстанавливаемой зависимости и метода обучения. В дальнейшем концепция оценок, зависящих от данных (data dependent bounds), получила развитие во многих работах [74, 82, 34, 35, 28].

К этому направлению примыкают также работы В. Л. Матросова, который впервые показал, что при специальном выборе метода обучения возможно обеспечить корректное распознавание любой заданной обучающей выборки, пользуясь подмножеством алгоритмов ограниченной ёмкости [14, 15, 16]. При этом построение алгоритма проводится в алгебраическом расширении семейства АВО (алгоритмов вычисления оценок) [12]. В отличие от стандартного подхода, здесь существенно используются свойства метода обучения, но не учитываются особенности распределения объектов и восстанавливаемой зависимости.

Статья [81] содержит исторический обзор, отражающий процесс постепенного уточнения оценок Вапника-Червоненкиса. Отмечается, что наилучшая оценка, справедливая при самых общих предположениях, получена М. Талаграндом [78]. На её основе выводится новая, несколько более точная, оценка, справедливая при некотором «разумном» ограничении класса вероятностных распределений на множестве исходных объектов.

При использовании оценок, зависящих от данных, метод структурной минимизации риска трансформируется и приводит к построению *самоограничивающих алгоритмов* (self bounding learning algorithms) [51]. От исходного СМР они отличаются тем, что структура вложенных подсемейств не задаётся заранее, а формируется в процессе обучения. В этом случае оценки качества учитывают все три типа особенностей: распределение объектов, свойства восстанавливаемой зависимости и метода обучения. Результатом обучения является не только сам алгоритм, но и достаточно точная оценка его обобщающей способности.

Принцип самоограничения алгоритмов применяется также для обоснования стандартных методов построения решающих деревьев [70]. Эти методы основаны на аналогичной стратегии — в ходе построения алгоритма по обучающей выборке происходит последовательное сужение подсемейства алгоритмов, в котором ведётся поиск решения [61].

3. Отступ (margin)

Второе направление связано с понятием *отступа* или *маржи* (margin) в задачах классификации с пороговым решающим правилом. Несколько упрощая, можно сказать, что отступ — это расстояние от объекта до границы классов. Если объект относится алгоритмом к чужому классу, то его отступ отрицателен. Чем больше в обучающей выборке объектов с большим отступом, тем лучше разделяются классы, тем надёжнее может быть классификация. Идея уточнения оценок качества заключается в том, чтобы сравнивать вероятность ошибки не с частотой ошибок на обучении, а с долей обучающих объектов, имеющих отрицательный или малый положительный отступ. При этом величина эмпирического риска искусственно завышается, зато вероятность ошибки существенно более точно оценивается по объектам, далеко отстоящим от границы классов.

Подход, основанный на понятии отступа, оказался особенно плодотворным при исследовании линейных пороговых классификаторов, в частности, машин опорных векторов (support vectors machines, SVM) [41, 77] и методов взвешенного голосования.

В работе П. Бартлетта [29] впервые было показано, что эффективная сложность выпуклой комбинации классификаторов равна не суммарной, и даже не максимальной (как ранее предполагалось), а средней взвешенной сложности отдельных классификаторов, взятых с теми же весами, с которыми они входят в комбинацию. Иными словами, взвешенное голосование не увеличивает сложность алгоритма, а лишь сглаживает прогнозы базовых классификаторов. Вытекающие отсюда оценки обобщающей способности существенно точнее классических сложностных оценок Вапника-Червоненкиса, хотя и они всё ещё сильно завышены (требуемая длина обучения имеет порядок 10^4–10^5). Этот результат

обосновывает ряд эвристических приёмов, направленных на уменьшение весов при настройке нейронных сетей, таких как «weight decay» и «early stopping». Он также позволяет обосновать алгоритмы, использующие метрику (функцию расстояния) в пространстве объектов, если предположить, что разделяющая поверхность проходит на достаточном удалении от обучающих объектов [31].

Результаты, первоначально полученные для линейных комбинаций, оказались применимы и к более широкому классу алгоритмов. В частности, бинарные решающие деревья и дизъюнктивные нормальные формы допускают представление в виде выпуклой комбинации булевых функций с пороговым решающим правилом [52]. Техника отступа позволяет оценивать обобщающую способность и более сложных алгоритмических композиций, представимых в виде пороговых выпуклых комбинаций над пороговыми выпуклыми комбинациями. Примерами таких конструкций являются сигмоидальные нейросети с одним скрытым уровнем и взвешенное голосование решающих деревьев [64]. Для всех этих случаев оценки обобщающей способности выражаются через долю обучающих объектов с малым отступом.

Наиболее ярким конструктивным результатом данного подхода являются методы обучения, направленные на явную максимизацию отступа. Они позволяют строить алгоритмы с лучшей обобщающей способностью, что подтверждается теоретически и экспериментально [63].

С понятием отступа тесно связана ещё одна мера сложности семейства алгоритмов, альтернативная функции роста — *fat-размерность* (fat-shattering dimension) [57, 25, 28].

4. Композиции алгоритмов

Третье направление исследований связано с понятием композиции алгоритмов. Во многих прикладных задачах удаётся построить несколько различных алгоритмов, ни один из которых не восстанавливает искомую зависимость достаточно хорошо. Тогда имеет смысл объединить эти алгоритмы с помощью корректирующей операции, в надежде на то, что ошибки одних алгоритмов будут скомпенсированы другими, и качество композиции окажется лучше, чем каждого из базовых алгоритмов в отдельности.

Известно несколько альтернативных способов конструирования алгоритмических композиций.

Наиболее общая теория алгоритмических композиций разработана в *алгебраическом подходе к построению корректных алгоритмов*, предложенном академиком РАН Ю. И. Журавлёвым и активно развиваемом его учениками [12, 11].

В методе Л. А. Растригина пространство объектов разбивается на *области компетентности*, и для каждой области строится свой алгоритм [17].

В методе *баггинга* (bagging — сокращение от «bootstrap aggregation»), предложенном Л. Брейманом [38, 39, 40], производится взвешенное голосование базовых алгоритмов, обученных на различных подвыборках данных, либо на различных частях признакового описания объектов. Выделение подмножеств объектов и/или признаков производится, как правило, случайным образом.

Метод *бустинга* (boosting), предложенный Р. Френдом и И. Шапиром [50, 47, 73] также является разновидностью взвешенного голосования, но базовые алгоритмы строятся последовательно, и процесс увеличения различий между ними управляется детерминированным образом. А именно, для каждого базового алгоритма, начиная со второго, веса обучающих объектов пересчитываются так, чтобы он точнее настраивался на тех объектах, на которых чаще ошибались все предыдущие базовые алгоритмы. Веса алгоритмов также вычисляются исходя из числа допущенных ими ошибок.

Идея последовательной компенсации ошибок предыдущих алгоритмов реализована также в оптимизационных (проблемно-ориентированных) методах алгебраического подхода [18, 5, 6]. В отличие от бустинга, здесь используется не выпуклая комбинация, а более сложная корректирующая операция в виде нелинейной монотонной функции достаточно общего вида.

Обобщающая способность бустинга исследована, пожалуй, наиболее хорошо. Во многих случаях экспериментально наблюдается почти неограниченное улучшение качества обучения при наращивании числа алгоритмов в композиции [48]. Более того, качество на тестовой выборке может продолжать улучшаться даже после достижения безошибочного распознавания обучающей выборки. Эти наблюдения противоречат непосредственным выводам статистической теории, основанным на анализе сложности.

Существует несколько объяснений феноменов бустинга. С одной стороны, бустинг активно максимизирует отступы обучающих объектов, и продолжает «раздвигать классы» даже после достижения безошибочного распознавания обучающей выборки [72]. С другой стороны, бустинг строит выпуклую комбинацию вещественнозначных классификаторов, которая проявляет свойство стабильности [46] (см. ниже).

Имеется много работ по сравнительному анализу обобщающей способности бустинга и баггинга. Баггинг направлен исключительно на уменьшение *вариации* (variance) модели, в то время как бустинг способствует уменьшению и вариации, и *смещения* (bias) [49]. Эмпирические исследования [76] на 4 реальных задачах показывают, что бустинг работает лучше на больших обучающих выборках, баггинг — на малых. При увеличении длины выборки бустинг повышает разнообразие классификаторов активнее, чем баггинг. Наконец, бустинг лучше воспроизводит границы классов сложной формы.

Работы Бартлетта, Френда, Шапира и др. решительным образом изменили представления о соотношении качества и сложности. Если ранее считалось, что для надёжного восстановления зависимости необходимо ограничивать сложность используемого семейства алгоритмов, то теперь исследователи приходят к выводу, что семейство может быть сколь угодно сложным, однако первостепенную роль играет *метод обучения* — тот способ, с помощью которого по обучающей выборке строится алгоритм из выбранного семейства. По всей видимости, некоторые разновидности взвешенного голосования, такие как бустинг, являются «удачными» методами, способными подстраиваться под конкретную задачу.

5. Стабильность метода обучения

Следующее, четвёртое, направление исследований связано с понятием *стабильности* (stability) [36, 37, 60]. Метод обучения называется стабильным, если небольшие вариации обучающей выборки приводят к незначительным изменениям получаемого алгоритма. Существуют различные способы формального определения стабильности, например, в работе [60] вводится 12 различных определений и устанавливаются взаимосвязи между ними. Как правило, оценки качества стабильных методов не зависят от сложностных характеристик семейства. В частности, получены оценки стабильности и обобщающей способности локальных методов типа ближайших соседей и потенциальных функций [71, 43, 44]. Эти методы широко используются благодаря своей простоте, однако порождают семейства алгоритмов бесконечной ёмкости. Доказана стабильность бустинга, машин опорных векторов, методов минимизации эмпирического риска с регуляризирующей штрафной функцией, и некоторых других. К сожалению, численные оценки требуемой длины обучения для стабильных методов также сильно завышены, как сложностные, и дают только качественное обоснование соответствующих алгоритмов.

6. Концентрация вероятности

Современные исследования таких свойств обучаемых алгоритмов, как эффективная сложность, отступ, композиционная структура и стабильность, существенно опираются на современный математический аппарат, описывающий явление концентрации вероятностной меры (measure concentration). В первых работах Вапника и Червоненкиса для этой цели использовались классические неравенства Хёфдинга и Бернштейна. Более точные результаты удаётся получать с помощью неравенств Чернова [42], метода ограниченных разностей Мак-Диармида [66] и изопериметрических неравенств Талагранда [78, 79]. Вводное изложение этих математических техник можно найти в обзорах [27, 62].

7. Скользящий контроль

Ещё одно направление исследований связано с использованием *скользящего контроля* (cross-validation) [45, 59].

Процедура скользящего контроля заключается в следующем. Фиксируется некоторое множество разбиений исходной выборки на две части: обучающую и контрольную. Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке и вычисляется частота его ошибок на контрольной подвыборке. Оценка скользящего контроля определяется как средняя по всем разбиениям частота ошибок на контроле. Фактически, скользящий контроль непосредственно измеряет обобщающую способность метода обучения на заданной конечной выборке.

В зависимости от способа формирования множества разбиений различают несколько разновидностей скользящего контроля [59]:

если множество разбиений одноэлементно, говорят об оценке качества на отдельной тестовой выборке (hold-out estimate);

если используются все разбиения с контрольной выборкой единичной длины, говорят об оценке с одним отделяемым объектом (leave-one-out estimate);

если используются все разбиения с контрольной выборкой фиксированной, но не обязательно единичной, длины, говорят об оценке полного скользящего контроля (complete cross-validation) [68];

если генерируется случайное подмножество разбиений с контрольной выборкой фиксированной длины, говорят о бутстреп-оценке (bootstrap estimate);

если множество разбиений образуется k непересекающимися контрольными выборками, говорят о k-кратном скользящем контроле (k-fold cross-validation).

На практике скользящий контроль применяется либо для выбора модели алгоритмов (model selection) из нескольких моделей-претендентов [58], либо для оптимизации небольшого числа параметров, определяющих структуру алгоритма, таких, как степень полинома или количество нейронов на скрытом уровне нейронной сети. Считается, что настройка значительной доли параметров по скользящему контролю лишена смысла. Когда контрольная выборка существенно вовлекается в процесс обучения, скользящий контроль начинает выдавать смещённую заниженную оценку обобщающей способности. Причиной является всё то же переобучение, которое приводит к заниженности эмпирического риска [69]. Известно, что скользящий контроль даёт несмещённую оценку вероятности ошибки в том случае, когда он используется для проверки качества по окончании обучения. Однако до сих пор нет исчерпывающих исследований, показывающих, в какой степени скользящий контроль может использоваться на стадии обучения.

Интуиция подсказывает, что скользящий контроль должен характеризовать обобщающую способность алгоритма лучше, чем частота ошибок на обучении. Тем не менее, этот факт долгое время не удавалось доказать. Попытки предпринимались неоднократно [58, 56, 53], но были получены лишь «разумные» верхние границы (sanity-check bounds) для отклонения скользящего контроля от вероятности ошибок алгоритма. Указанные оценки даже несколько хуже, чем оценки Вапника-Червоненкиса для отклонения эмпирического риска и требуют дополнительных предположений о стабильности метода обучения [56].

Причина этих неудач анализируется в [32], где вводятся и сравниваются два альтернативных способа формализации понятия обобщающей способности. При первом способе, близком к подходу Вапника-Червоненкиса, оценивается качество *отдельного алгоритма*, полученного в результате обучения. Это приводит к завышенным оценкам, зависящим от ёмкости семейства и требующим дополнительных предположений о стабильности метода обучения [56]. При втором способе оценивается качество *метода обучения* в целом. Оказывается, в этом случае оценка отклонения скользящего контроля от вероятности ошибки алгоритма, обученного на случайной выборке, не зависит от ёмкости семейства, а только от длины обучения и контроля. С ростом длины обеих выборок указанное отклонение стремится к нулю. Данный результат проясняет природу скользящего контроля и показывает, что завышенность предыдущих оценок связана с неудачным выбором исходного функционала качества.

Отсюда вытекает важный вывод: теория качества обучения может оказаться весьма чувствительной к исходной аксиоматике, в частности, к формализации самого понятия качества обучения. Второй важный вывод заключается в том, что скользящий контроль характеризует обобщающую способность метода не намного хуже, чем вероятность ошибки. Наиболее точное выражение эти идеи нашли в комбинаторном подходе к обоснованию обучаемых алгоритмов.

8. Комбинаторный подход

Комбинаторный подход [4, 9, 8] возник как попытка более точного построения статистической теории Вапника-Червоненкиса, начиная с исходных её постулатов. Для этого имелось две основные предпосылки.

Во-первых, сложилось понимание того, что принцип минимизации эмпирического риска в заранее заданном семействе алгоритмов не достаточно точно описывает процесс обучения. Во-первых, не вполне ясно, где проходит граница семейства. Может оказаться так, что формально выписано очень широкое семейство, но на практике процедура обучения выдаёт алгоритмы лишь из небольшой его части. Во-вторых, доставлять минимум эмпирическому риску могут многие

алгоритмы, однако в качестве решения всегда выбирается только один. Конкретизация метода его построения, возможно, позволит учесть специфические особенности процесса обучения. В-третьих, далеко не все методы обучения, хорошо зарекомендовавшие себя на практике, минимизируют эмпирический риск. В качестве примеров можно привести методы выбора модели по скользящему контролю или другим внешним критериям [13], методы регуляризации эмпирического риска, методы явной максимизации отступа, бустинг, баггинг, и т. д.

В комбинаторном подходе явным образом вводится понятие *метода обучения* как отображения, которое конечной обучающей выборке ставит в соответствие некоторый вполне определённый алгоритм. Семейство алгоритмов становится вторичной конструкцией — это все алгоритмы, которые могут быть получены в результате применения данного метода обучения ко всевозможным конечным выборкам. Одновременно появляется возможность единообразно рассматривать любые методы, а не только минимизацию эмпирического риска.

Второй предпосылкой было понимание того, что вероятность ошибки является гипотетической величиной, которую невозможно вычислить, а иногда даже и оценить, например, в случае малых выборок. В то же время, на практике любая обучаемая система сталкивается только с конечными выборками, будь то обучающие, контрольные или рабочие совокупности объектов. Поэтому обобщающую способность алгоритмов целесообразно характеризовать именно относительно конечных выборок. Желательно также, чтобы функционал качества можно было с контролируемой точностью измерять по имеющимся эмпирическим данным. Наконец, использование гипотетических вероятностей может приводить к лишним промежуточным шагам при доказательстве оценок и понижать их точность.

В комбинаторном подходе качество обучения по прецедентам (обобщающая способность метода) характеризуется функционалами полного скользящего контроля, зависящими только от метода обучения и заданной конечной выборки. Такие функционалы предлагается называть *комбинаторными*, поскольку они определяются через множество всех разбиений выборки.

Получены верхние оценки комбинаторных функционалов, аналогичные по своей структуре статистическим [9]. Они оказываются даже более точными, поскольку вместо сложности всего семейства в них фигурирует сложность *локального подсемейства*, состоящего из алгоритмов, выдаваемых методом обучения в данной конкретной задаче.

Комбинаторные оценки, в отличие от статистических, справедливы для любого метода обучения и любой конечной выборки, не обязательно случайной,

независимой, одинаково распределённой. Их доказательство проводится исключительно комбинаторными методами и вообще не опирается на теорию вероятностей. Данный факт представляется весьма неожиданным. До сих пор вероятностная природа проблемы качества обучения оставалась, пожалуй, единственным постулатом статистической теории, никогда не подвергавшимся сомнению. Но возможна и другая точка зрения: само понятие вероятности содержит «встроенный» предельный переход, поэтому его применение не вполне уместно в дискретных задачах с конечными, зачастую малыми, выборками.

Комбинаторный подход не отвергает, а уточняет статистическую теорию. Любая комбинаторная оценка легко «превращается» в вероятностную, если снова принять стандартный набор вероятностных гипотез и применить операцию математического ожидания одновременно к функционалу и его оценке. Таким образом, при переходе от статистической теории к комбинаторной соблюдается «принцип соответствия».

В то же время, комбинаторная перестройка аксиоматики приводит к существенному пересмотру многих положений статистической теории.

1. Становится полностью очевидной избыточность требования равномерной сходимости. На практике восстанавливаемая зависимость и метод обучения всегда фиксированы, а обучающая выборка — конечна. Поэтому лишь конечная часть семейства может быть получена в результате обучения, остальные алгоритмы остаются незадействованными. Разумеется, наибольший интерес представляют ситуации, когда сложность локального подсемейства оказывается существенно меньше сложности всего семейства. Этот эффект предлагается называть *локализацией* семейства алгоритмов. Существование эффекта локализации снимает искусственный запрет на использование сложных алгоритмов. Важно не столько ограничить ёмкость семейства, сколько разработать метод обучения, способный подстраиваться под конкретные задачи, всякий раз поразному локализуя «рабочую область» семейства. При фиксации восстанавливаемой зависимости метод обучения должен строить алгоритмы, «похожие» на неё. Тогда не важно, сколько ещё «не похожих» алгоритмов содержится в семействе. Это свойство предлагается называть *локализующей способностью* метода обучения. Оно является важной компонентой его обобщающей способности.

2. Комбинаторный подход позволяет по-новому взглянуть на проблему построения корректных алгоритмов (не допускающих ошибок на обучающей выборке). Комбинаторные оценки представляются в виде произведения локальной функции роста, которая может быть много меньше функции роста всего семейства, и комбинаторного множителя, который быстро возрастает по мере

увеличения числа ошибок на обучении. Очевидно, для обеспечения корректности необходимо усложнять конструкцию алгоритмов. Согласно статистической теории это приводит к значительному увеличению функции роста всего семейства, на фоне которого эффект уменьшения комбинаторного множителя остаётся незаметным. Отсюда делается вывод, что не следует добиваться безошибочной работы алгоритма на обучающем материале. С точки зрения комбинаторного подхода усложнение конструкции алгоритма не обязательно приводит к существенному увеличению локальной функции роста. В этом случае требование корректности становится крайне желательным, поскольку оно резко уменьшает комбинаторный множитель. Отметим, что идея построения корректных алгоритмических композиций является центральной в алгебраическом подходе к распознаванию [12].

3. Отличительной чертой комбинаторного подхода является сохранение комбинаторного множителя в исходном, достаточно громоздком, виде. Элементарные расчёты показывают, что его экспоненциальные приближения, принятые в статистической теории, приводят к ослаблению оценки в несколько раз. Современные вычислительные средства позволяют достаточно эффективно работать с исходной формулой.

4. Существенно трансформируется метод структурной минимизации риска. Поскольку комбинаторные функционалы можно измерять по выборке, появляется возможность вообще отказаться от завышенных верхних оценок, и перейти к непосредственному использованию скользящего контроля. Но это именно то, что предлагали делать Вапник и Червоненкис на практике, правда, без видимой связи с основными теоретическими результатами [3]. В комбинаторном подходе построение структуры вложенных подсемейств различной ёмкости теряет смысл. Вместо этого достаточно брать конечный набор методов обучения и выбирать из них лучший по критерию скользящего контроля. Некоторые эмпирические исследования показывают, что данная техника выбора модели алгоритмов во многих случаях предпочтительнее принципов структурной минимизации риска и минимальной длины описания (minimum description length), направленных на явную оптимизацию сложности [55].

5. Предложенное в работах [80, 33] понятие эффективной ёмкости основано на эмпирическом измерении функционала равномерного отклонения частоты ошибок в двух выборках для задач классификации. В комбинаторном подходе этот функционал очевидным образом заменяется на функционал скользящего контроля, что приводит к возникновению нового понятия *локальной эффективной ёмкости*. В отличие от эффективной ёмкости по Вапнику, локальная эффективная ёмкость учитывает все особенности распределения объектов, восстанавливаемой зависимости и метода обучения.

6. Анализ комбинаторных оценок позволяет назвать три основные причины завышенности сложностных оценок качества: пренебрежение эффектом локализации, погрешность экспоненциального приближения комбинаторного множителя и погрешность, связанная с самим переходом от качества к сложности. Комбинаторный аналог оценок Вапника-Червоненкиса позволяет устранить только первые две причины. В силу третьей причины любые сложностные оценки качества обучения являются принципиально завышенными.

Данный факт позволяет выдвинуть предположение, что получить приемлемые численные оценки качества возможно только при явном учёте априорной информации о выборке и восстанавливаемой зависимости.

9. Универсальные ограничения

Основная идея этого направления состоит в том, что если метод обучения строит алгоритмы, в некотором смысле «согласованные» с имеющейся априорной информацией, то обобщающая способность такого метода может оказаться существенно лучше, чем в общем случае.

Соответствие обучающей выборки (локальной информации) и априорных ограничений (универсальной информации) подробно изучается в теории универсальных и локальных ограничений К. В. Рудакова [19, 22, 20, 21, 23, 11] с позиций теории категорий и алгебраического подхода к проблеме распознавания. Алгебраическая теория позволяет проверять непротиворечивость этих двух типов информации и конструктивно описывать неизбыточные классы моделей алгоритмов, допускающие построение корректных (не ошибающихся на обучающей выборке) алгоритмов. Однако оценки обобщающей способности в данной теории не рассматриваются. Вообще, проблема влияния априорной информации на качество восстановления зависимости представляется наиболее сложной и наименее изученной.

Комбинаторный подход существенно облегчает развитие данного направления, поскольку отпадает необходимость согласовывать априорную информацию со свойствами вероятностной меры.

В частности, получена не-вероятностная оценка функционала скользящего контроля для случая, когда искомая зависимость монотонна или почти-монотонна, и метод обучения строит только монотонные отображения [7, 9]. Априорная информация выражается в форме «профиля монотонности» выборки, который характеризует плотность отношения порядка вблизи границы классов. Данная оценка никогда не превышает единицы, не зависит от сложности семейства (имеющего, как известно, бесконечную ёмкость), и является существенно более точной на малых выборках, чем оценки, полученные ранее [24, 75].

Ещё одна не-вероятностная оценка получена для метода ближайшего соседа при наличии априорной информации о компактности классов, выраженной в форме «профиля компактности» выборки. Данная оценка является точной и вытекает непосредственно из формул эффективного вычисления полного скользящего контроля для метода ближайших соседей [68]. Она также не зависит от сложностных характеристик семейства, имеющего бесконечную ёмкость.

В заключение отметим, что дополнением к данному обзору является периодически пополняемая частично аннотированная библиографическая база `MachLearn`, размещённая по адресу `www.ccas.ru/frc`.

Автор выражает глубокую признательность академику РАН Ю. И. Журавлёву за оказываемую поддержку и своему Учителю чл.-корр. РАН К. В. Рудакову за постоянное внимание к работе и ценные замечания.

Работа поддержана Российским фондом фундаментальных исследований (проекты № 02-01-00325, № 01-07-90242) и Фондом содействия отечественной науке.

Список литературы

[1] *Вапник В. Н., Червоненкис А. Я.* О равномерной сходимости частот появления событий к их вероятностям // *ДАН СССР.* — 1968. — Т. 181, № 4. — С. 781–784.

[2] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[3] *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

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

[5] *Воронцов К. В.* О проблемно-ориентированной оптимизации базисов задач распознавания // *ЖВМ и МФ.* — 1998. — Т. 38, № 5. — С. 870–880.
`http://www.ccas.ru/frc/papers/voron98jvm.pdf`.

[6] *Воронцов К. В.* Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания // *ЖВМ и МФ.* — 2000. — Т. 40, № 1. — С. 166–176.
`http://www.ccas.ru/frc/papers/voron00jvm.pdf`.

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

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

[9] *Воронцов К. В.* Комбинаторные оценки качества обучения по прецедентам // *Доклады РАН.* — 2004. — Т. 394, № 2.
`http://www.ccas.ru/frc/papers/voron04qualdan.pdf`.

[10] *Дюличева Ю. Ю.* Оценка VCD r-редуцированного эмпирического леса // *Таврический вестник информатики и математики.* — 2003. — № 1. — С. 31–42.

[11] *Журавлёв Ю. И., Рудаков К. В.* Об алгебраической коррекции процедур обработки (преобразования) информации // *Проблемы прикладной математики и информатики.* — 1987. — С. 187–198.
http://www.ccas.ru/frc/papers/zhurrud87correct.pdf.

[12] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // *Проблемы кибернетики.* — 1979. — Т. 33. — С. 5–68.

[13] *Ивахненко А. Г., Юрачковский Ю. П.* Моделирование сложных систем по экспериментальным данным. — М.: Радио и связь, 1987.

[14] *Матросов В. Л.* Корректные алгебры ограниченной ёмкости над множествами некорректных алгоритмов // *ДАН СССР.* — 1980. — Т. 253, № 1. — С. 25–30.

[15] *Матросов В. Л.* Ёмкость алгебраических расширений модели алгоритмов вычисления оценок // *ЖВМиМФ.* — 1984. — Т. 24, № 11. — С. 1719–1730.

[16] *Матросов В. Л.* Ёмкость алгоритмических многочленов над множеством алгоритмов вычисления оценок // *ЖВМиМФ.* — 1985. — Т. 25, № 1. — С. 122–133.

[17] *Растригин Л., Эренштейн Р.* Коллективные правила распознавания. — М.: Энергия, 1981. — Р. 244.

[18] *Рудаков К. В., Воронцов К. В.* О методах оптимизации и монотонной коррекции в алгебраическом подходе к проблеме распознавания // *Доклады РАН.* — 1999. — Т. 367, № 3. — С. 314–317.
http://www.ccas.ru/frc/papers/rudvoron99dan.pdf.

[19] *Рудаков К. В.* О симметрических и функциональных ограничениях для алгоритмов классификации // *ДАН СССР.* — 1987. — Т. 297, № 1. — С. 43–46.
http://www.ccas.ru/frc/papers/rudakov87dan.pdf.

[20] *Рудаков К. В.* Полнота и универсальные ограничения в проблеме коррекции эвристических алгоритмов классификации // *Кибернетика.* — 1987. — № 3. — С. 106–109.

[21] *Рудаков К. В.* Симметрические и функциональные ограничения в проблеме коррекции эвристических алгоритмов классификации // *Кибернетика.* — 1987. — № 4. — С. 73–77.
http://www.ccas.ru/frc/papers/rudakov87symmetr.pdf.

[22] *Рудаков К. В.* Универсальные и локальные ограничения в проблеме коррекции эвристических алгоритмов // *Кибернетика.* — 1987. — № 2. — С. 30–35.
http://www.ccas.ru/frc/papers/rudakov87universal.pdf.

[23] *Рудаков К. В.* О применении универсальных ограничений при исследовании алгоритмов классификации // *Кибернетика.* — 1988. — № 1. — С. 1–5.
http://www.ccas.ru/frc/papers/rudakov88universal.pdf.

[24] *Сёмочкин А. Н.* Оценки функционала качества для класса алгоритмов с универсальными ограничениями монотонности // *Депонир. в ВИНИТИ РАН.* — 1998. — № 2965–В98. — С. 20.

[25] *Anthony M., Bartlett P. L.* Neural Network Learning: Theoretical Foundations. — Cambridge University Press, Cambridge, 1999.

[26] *Anthony M., Shawe-Taylor J.* A result of Vapnik with applications // *Discrete Applied Mathematics.* — 1993. — Vol. 47, no. 2. — Pp. 207–217.
http://citeseer.nj.nec.com/anthony91result.html.

[27] *Anthony M.* Uniform glivenko-cantelli theorems and concentration of measure in the mathematical modelling of learning: Tech. Rep. LSE-CDAM-2002-07: 2002.
http://www.maths.lse.ac.uk/Personal/martin/mresearch.html.

[28] *Antos A., Kegl B., Linder T., Lugosi G.* Data-dependent margin-based generalization bounds for classification // *Journal of Machine Learning Research.* — 2002. — Pp. 73–98.
http://citeseer.nj.nec.com/article/antos02datadependent.html.

[29] *Bartlett P. L.* For valid generalization the size of the weights is more important than the size of the network // Advances in Neural Information Processing Systems / Ed. by M. C. Mozer, M. I. Jordan, T. Petsche. — Vol. 9. — The MIT Press, 1997. — P. 134.
http://citeseer.nj.nec.com/bartlett97for.html.

[30] *Bartlett P.* Lower bounds on the Vapnik-Chervonenkis dimension of multi-layer threshold networks // Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory. — ACM Press, New York, NY, 1993. — Pp. 144–150.
http://citeseer.nj.nec.com/bartlett93lower.html.

[31] *Bartlett P.* The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network // *IEEE Transactions on Information Theory.* — 1998. — Vol. 44, no. 2. — Pp. 525–536.
http://discus.anu.edu.au/~bartlett.

[32] *Bontempi G., Birattari M.* A bound on the cross-validation estimate for algorithm assessment // Eleventh Belgium/Netherlands Conference on Artificial Intelligence (BNAIC). — 1999. — Pp. 115–122.
http://citeseer.nj.nec.com/225930.html.

[33] *Bottou L., Cortes C., Vapnik V.* On the effective VC dimension. — 1994.
http://citeseer.nj.nec.com/bottou94effective.html.

[34] *Boucheron S., Lugosi G., Massart P.* A sharp concentration inequality with applications // *Random Structures and Algorithms.* — 2000. — Vol. 16, no. 3. — Pp. 277–292.
http://citeseer.nj.nec.com/article/boucheron99sharp.html.

[35] *Boucheron S., Lugosi G., Massart P.* Concentration inequalities using the entropy method. — 2003.
http://citeseer.nj.nec.com/boucheron02concentration.html.

[36] *Bousquet O., Elisseeff A.* Algorithmic stability and generalization performance // Advances in Neural Information Processing Systems 13. — 2001. — Pp. 196–202.
http://citeseer.nj.nec.com/article/bousquet00algorithmic.html.

[37] *Bousquet O., Elisseeff A.* Stability and generalization // *Journal of Machine Learning Research.* — 2002. — no. 2. — Pp. 499–526.
http://citeseer.nj.nec.com/article/bousquet00stability.html.

[38] *Breiman L.* Bagging predictors // *Machine Learning.* — 1996. — Vol. 24, no. 2. — Pp. 123–140.
http://citeseer.nj.nec.com/breiman96bagging.html.

[39] *Breiman L.* Bias, variance, and arcing classifiers: Tech. Rep. 460: Statistics Department, University of California, 1996.
http://citeseer.nj.nec.com/breiman96bias.html.

[40] *Breiman L.* Arcing classifiers. — 1998.
http://citeseer.nj.nec.com/breiman98arcing.html.

[41] *Burges C. J. C.* A tutorial on support vector machines for pattern recognition // *Data Mining and Knowledge Discovery.* — 1998. — Vol. 2, no. 2. — Pp. 121–167.
http://citeseer.nj.nec.com/burges98tutorial.html.

[42] *Chernoff H.* A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations // *Annals of Math. Stat.* — 1952. — Vol. 23. — Pp. 493–509.

[43] *Devroye L. P., Wagner T. J.* Distribution-free inequalities for the deleted and holdout error estimates // *IEEE Transactions on Information Theory.* — 1979. — Vol. 25, no. 2. — Pp. 202–207.

[44] *Devroye L. P., Wagner T. J.* Distribution-free performance bounds for potential function rules // *IEEE Transactions on Information Theory.* — 1979. — Vol. 25, no. 5. — Pp. 601–604.

[45] *Efron B.* The Jackknife, the Bootstrap, and Other Resampling Plans. — SIAM, Philadelphia, 1982.

[46] *Evgeniou T., Pontil M., Elisseeff A.* Leave one out error, stability, and generalization of voting combinations of classifiers: Tech. Rep. INSEAD 2001-21-TM: 2001.
`http://citeseer.nj.nec.com/445768.html`.

[47] *Freund Y., Schapire R. E.* A decision-theoretic generalization of on-line learning and an application to boosting // European Conference on Computational Learning Theory. — 1995. — Pp. 23–37.
`http://citeseer.nj.nec.com/article/freund95decisiontheoretic.html`.

[48] *Freund Y., Schapire R. E.* Experiments with a new boosting algorithm // International Conference on Machine Learning. — 1996. — Pp. 148–156.
`http://citeseer.nj.nec.com/freund96experiments.html`.

[49] *Freund Y., Schapire R. E.* Discussion of the paper "Arcing classifiers" by Leo Breiman // *The Annals of Statistics.* — 1998. — Vol. 26, no. 3. — Pp. 824–832.
`http://citeseer.nj.nec.com/freund97discussion.html`.

[50] *Freund Y.* Boosting a weak learning algorithm by majority // COLT: Proceedings of the Workshop on Computational Learning Theory. — Morgan Kaufmann Publishers, 1990.
`http://citeseer.nj.nec.com/freund95boosting.html`.

[51] *Freund Y.* Self bounding learning algorithms // COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers. — 1998.
`http://citeseer.nj.nec.com/freund98self.html`.

[52] *Golea M., Bartlett P., Lee W. S., Mason L.* Generalization in decision trees and DNF: Does size matter? // Advances in Neural Information Processing Systems / Ed. by M. I. Jordan, M. J. Kearns, S. A. Solla. — Vol. 10. — The MIT Press, 1998.
`http://citeseer.nj.nec.com/golea97generalization.html`.

[53] *Holden S. B.* Cross-validation and the pac learning model: Tech. Rep. RN/96/64: Dept. of CS, Univ. College, London, 1996.

[54] *Karpinski M., Macintyre A.* Polynomial bounds for VC dimension of sigmoidal neural networks // 27th ACM Symp. Theory Comput. — 1995. — Pp. 200–208.
`http://citeseer.nj.nec.com/karpinski95polynomial.html`.

[55] *Kearns M. J., Mansour Y., Ng A. Y., Ron D.* An experimental and theoretical comparison of model selection methods // Computational Learning Theory. — 1995. — Pp. 21–30.
`http://citeseer.nj.nec.com/kearns95experimental.html`.

[56] *Kearns M. J., Ron D.* Algorithmic stability and sanity-check bounds for leave-one-out cross-validation // Computational Learning Theory. — 1997. — Pp. 152–162.
`http://citeseer.nj.nec.com/kearns97algorithmic.html`.

[57] *Kearns M. J., Schapire R. E.* Efficient distribution-free learning of probabilistic concepts // Computational Learning Theory and Natural Learning Systems, Volume I: Constraints and Prospect, edited by Stephen Jose Hanson, George A. Drastal, and Ronald L. Rivest, Bradford/MIT Press. — 1994. — Vol. 1.
`http://citeseer.nj.nec.com/article/kearns93efficient.html`.

[58] *Kearns M.* A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split // Advances in Neural Information Processing Systems / Ed. by D. S. Touretzky, M. C. Mozer, M. E. Hasselmo. — Vol. 8. — The MIT Press, 1996. — Pp. 183–189.
http://citeseer.nj.nec.com/kearns96bound.html.

[59] *Kohavi R.* A study of cross-validation and bootstrap for accuracy estimation and model selection // IJCAI. — 1995. — Pp. 1137–1145.
http://citeseer.nj.nec.com/kohavi95study.html.

[60] *Kutin S., Niyogi P.* Almost-everywhere algorithmic stability and generalization error: Tech. Rep. TR-2002-03: University of Chicago, 2002.
http://citeseer.nj.nec.com/kutin02almosteverywhere.html.

[61] *Langford J., Blum A.* Microchoice bounds and self bounding learning algorithms // Computational Learing Theory. — 1999. — Pp. 209–214.
http://citeseer.nj.nec.com/langford01microchoice.html.

[62] *Lugosi G.* On concentration-of-measure inequalities. — Machine Learning Summer School 2003, Australian National University, Canberra. — 2003.
http://citeseer.nj.nec.com/lugosi98concentrationmeasure.html.

[63] *Mason L., Bartlett P., Baxter J.* Direct optimization of margins improves generalization in combined classifiers: Tech. rep.: Deparment of Systems Engineering, Australian National University, 1998.
http://citeseer.nj.nec.com/mason98direct.html.

[64] *Mason L., Bartlett P., Golea M.* Generalization error of combined classifiers: Tech. rep.: Department of Systems Engineering, Australian National University, 1997.
http://citeseer.nj.nec.com/mason97generalization.html.

[65] *Mazurov V., Khachai M., Rybin A.* Committee constructions for solving problems of selection, diagnostics and prediction // *Proceedings of the Steklov Institute of mathematics.* — 2002. — Vol. 1. — Pp. 67–101.
http://tom.imm.uran.ru/khachay/publications/mine/psis67.pdf.

[66] *McDiarmid C.* On the method of bounded differences // *In Surveys in Combinatorics, London Math. Soc. Lecture Notes Series.* — 1989. — Vol. 141. — Pp. 148–188.

[67] *Mertens S., Engel A.* Vapnik-Chervonenkis dimension of neural networks with binary weights // *Phys. Rev. E.* — 1997. — Vol. 55, no. 4. — Pp. 4478–4488.

[68] *Mullin M., Sukthankar R.* Complete cross-validation for nearest neighbor classifiers // Proceedings of International Conference on Machine Learning. — 2000.
http://citeseer.nj.nec.com/309025.html.

[69] *Ng A. Y.* Preventing overfitting of cross-validation data // Proc. 14th International Conference on Machine Learning. — Morgan Kaufmann, 1997. — Pp. 245–253.
http://citeseer.nj.nec.com/ng97preventing.html.

[70] *Quinlan J.* Induction of decision trees // *Machine Learning.* — 1986. — Vol. 1, no. 1. — Pp. 81–106.

[71] *Rogers W., Wagner T.* A finite sample distribution-free performance bound for local discrimination rules // *Annals of Statistics.* — 1978. — Vol. 6, no. 3. — Pp. 506–514.

[72] *Schapire R. E., Freund Y., Lee W. S., Bartlett P.* Boosting the margin: a new explanation for the effectiveness of voting methods // *Annals of Statistics.* — 1998. — Vol. 26, no. 5. — Pp. 1651–1686.
http://citeseer.nj.nec.com/article/schapire98boosting.html.

[73] *Schapire R.* The boosting approach to machine learning: An overview. — 2001.
http://citeseer.nj.nec.com/schapire02boosting.html.

[74] *Shawe-Taylor J., Bartlett P. L.* Structural risk minimization over data-dependent hierarchies // *IEEE Trans. on Information Theory.* — 1998. — Vol. 44, no. 5. — Pp. 1926–1940.
http://citeseer.nj.nec.com/article/shawe-taylor98structural.html.

[75] *Sill J.* The capacity of monotonic functions // *Discrete Applied Mathematics (special issue on VC dimension).* — 1998. — Vol. 86. — Pp. 95–107.
http://citeseer.nj.nec.com/49191.html.

[76] *Skurichina M., Kuncheva L., Duin R.* Bagging and boosting for the nearest mean classifier: Effects of sample size on diversity and accuracy // Multiple Classifier Systems (Proc. Third International Workshop MCS, Cagliari, Italy) / Ed. by J. K. F. Roli. — Vol. 2364. — Springer, Berlin, 2002. — Pp. 62–71.
http://citeseer.nj.nec.com/539135.html.

[77] *Smola A., Bartlett P., Scholkopf B., Schuurmans D.* Advances in large margin classifiers. — 2000.
http://citeseer.nj.nec.com/article/smola00advances.html.

[78] *Talagrand. M.* Sharper bounds for gaussian and empirical processes // *Annals of Probability.* — 1994. — no. 22. — Pp. 28–76.

[79] *Talagrand M.* Concentration of measure and isoperimetric inequalities in product space. — 1995.
http://citeseer.nj.nec.com/talagrand95concentration.html.

[80] *Vapnik V., Levin E., Cun Y. L.* Measuring the VC-dimension of a learning machine // *Neural Computation.* — 1994. — Vol. 6, no. 5. — Pp. 851–876.
http://citeseer.nj.nec.com/vapnik94measuring.html.

[81] *Vayatis N., Azencott R.* Distribution-dependent Vapnik-Chervonenkis bounds // *Lecture Notes in Computer Science.* — 1999. — Vol. 1572. — Pp. 230–240.
http://citeseer.nj.nec.com/vayatis99distributiondependent.html.

[82] *Williamson R., Shawe-Taylor J., Scholkopf B., Smola A.* Sample based generalization bounds: Tech. Rep. NeuroCOLT Technical Report NC-TR-99-055: 1999.
http://citeseer.nj.nec.com/williamson99sample.html.