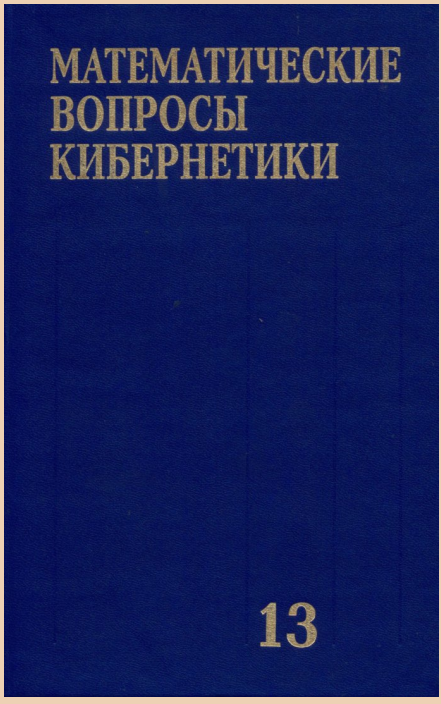
ИПМ им. М.В. Келдыша РАН • Электронная библиотека
Математические вопросы кибернетики • Выпуск 13



К. В. Воронцов

Комбинаторный подход к оценке качества обучаемых алгоритмов

Рекомендуемая форма библиографической ссылки: Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. Вып. 13. — М.: ФИЗМАТЛИТ, 2004. — С. 5–36. URL: http://library.keldysh.ru/mvk.asp?id=2004-5

КОМБИНАТОРНЫЙ ПОДХОД К ОЦЕНКЕ КАЧЕСТВА ОБУЧАЕМЫХ АЛГОРИТМОВ*)

К. В. ВОРОНЦОВ

(МОСКВА)

§ 1. Обзор современных исследований по проблеме качества обучения алгоритмов

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, т. е. 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 распознавания, восстановления регрессии и прогнозирования был накоплен уже к середине 60-х годов. Большую популярность приобрел 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. Этот метод получил название *минимизации эмпирического риска*.

На практике исследователи столкнулись с проблемой *переобучения* (overfitting). Чем больше у алгоритма свободных параметров, тем меньшего числа ошибок на обучении можно добиться путем 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).

---

*) Работа выполнена в рамках программы Отделения математических наук РАН «Алгебраические и комбинаторные методы математической кибернетики», поддержана Российским фондом фундаментальных исследований (проекты 02–01–00325, 01–07–90242) и Фондом содействия отечественной науке.

Из этого наблюдения был сделан вывод, что для всякой задачи существует оптимальная сложность модели, при которой достигается наилучшее качество обобщения. Первое формальное обоснование этого практического опыта было дано в статистической теории восстановления зависимостей по эмпирическим данным, разработанной В. Н. Вапником и А. Я. Червоненкисом в конце 60-х — начале 70-х [2, 3].

Статистическая теория Вапника–Червоненкиса. Обобщающая способность определяется как вероятность ошибки алгоритма, полученного в результате обучения, либо как частота его ошибок на некоторой независимой и, вообще говоря, неизвестной контрольной выборке. Далее вводится подходящая мера сложности семейств алгоритмов, называемая *емкостью* или размерностью Вапника–Червоненкиса (VC-dimension). Основным результатом теории являются количественные оценки, показывающие, что качество получаемых алгоритмов улучшается с ростом длины обучающей выборки и уменьшением частоты ошибок на обучении, но ухудшается при увеличении сложности семейства. Эти оценки позволяют обосновать метод *структурной минимизации риска* (СМР), непосредственно направленный на выбор модели оптимальной сложности. В СМР фиксируется некоторая 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. Более подробно основные положения статистической теории излагаются в § 3.

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 [62]. Некоторые семейства имеют бесконечную емкость и находятся за границами применимости теории, тем не менее с их помощью удается 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 [9–12], 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 справедливы при любом распределении вероятностей и любой восстанавливаемой зависимости, следовательно, ориентированы на «худший случай». Они не учитывают трех важных особенностей самой задачи и процесса поиска ее решения, существенно влияющих на качество обучения.

Во-первых, это особенности распределения объектов в пространстве — они могут лежать в подпространстве меньшей размерности. В задачах восстановления зависимостей этот «вырожденный» случай довольно распространен, поскольку 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, выделяя эффективное подсемейство алгоритмов, реально получаемых в результате обучения.

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

Эффективная сложность. При решении конкретной задачи далеко не каждый алгоритм из выбранного семейства имеет шансы быть полученным в результате обучения. Как правило, реально работает не все семейство, а лишь небольшая его часть. Этот факт был замечен еще В. Н. Вапником, предложившим понятие эффективной емкости вместе с алгоритмом ее практического измерения [38, 88]. Эффективная емкость не превосходит полной емкости 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, см. например, [32, 39, 40, 81, 90].

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 емкости [15–17]. При этом построение алгоритма проводится в алгебраическом расширении семейства алгоритмов вычисления оценок (АВО) [12]. В отличие от стандартного подхода, здесь существенно используются свойства метода обучения, но не учитываются особенности распределения объектов и восстанавливаемой зависимости.

Статья [89] содержит исторический обзор, отражающий процесс постепенного уточнения оценок Вапника–Червоненкиса. Отмечается, что наилучшая оценка, справедливая при самых общих предположениях, получена М. Талаграндом [85]. На ее 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) [54]. От исходного СМР они отличаются тем, что структура вложенных подсемейств не задается заранее, а формируется в процессе обучения. В этом случае оценки качества учитывают все три типа особенностей, упомянутых выше. Результатом обучения является не только сам алгоритм, но и оценка его обобщающей способности. Такая оценка уже не выписывается явно в виде формулы, а вычисляется по ходу построения алгоритма. Наличие оценки качества на каждом промежуточном шаге позволяет эффективно управлять процессом обучения.

Принцип самоограничения алгоритмов применяется также для обоснования стандартных методов построения решающих деревьев [76]. Эти методы основаны на аналогичной стратегии — в ходе построения алгоритма по обучающей выборке происходит последовательное сужение подсемейства алгоритмов, в котором ведется поиск решения [67].

Граничность объектов. 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 надежнее 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) [46, 84] и методов взвешенного голосования [34], о которых пойдет речь чуть ниже. Он также позволяет обосновать алгоритмы, использующие метрику (функцию расстояния) в пространстве объектов, если предположить, что разделяющая поверхность проходит на достаточном удалении от обучающих объектов [35].

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 [69].

С понятием отступа тесно связана еще одна мера сложности семейства алгоритмов, альтернативная функции роста — *fat-размерность* (fat-shattering dimension) [30, 32, 36, 64].

Композиционная структура алгоритмов. Третье направление исследований связано с понятием композиции алгоритмов. Во многих прикладных задачах удается 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 недостатки одних алгоритмов будут скомпенсированы достоинствами других, и качество композиции окажется лучше, чем качество каждого из базовых алгоритмов в отдельности.

Известно много 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 [6, 9–13, 25].

В методе Л. А. Растригина пространство объектов разбивается на *области компетентности*, и для каждой области строится свой алгоритм [18]. Затем эти алгоритмы «склеиваются». Качество такой конструкции определяется качеством отдельных алгоритмов.

Раньше других стали применяться простые процедуры голосования (по большинству, по старшинству, и другие), не имеющие собственных параметров для настройки. Ввиду существенной дискретности они используются, главным образом, при решении задач классификации. Более общая процедура *взвешенного голосования* (weighted voting) строит выпуклую комбинацию алгоритмов — линейную комбинацию с неотрицательными нормированными весами. Эта процедура обладает существенно большей гибкостью, и ее применяют как для классификации, так и для восстановления регрессии. Многие широко известные алгоритмические конструкции явно или неявно используют принцип взвешенного голосования, в частности, нейронные сети, потенциальные функции [1], алгоритмы вычисления оценок [12], и другие.

В работе П. Бартлетта [34] 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 все еще сильно завышены (требуемая длина обучения имеет порядок 10^4–10^5). Этот результат обосновывает ряд эвристических приемов, направленных на уменьшение весов при настройке нейронных сетей, таких как сокращение весов (weight decay) и ранний останов (early stopping).

Результаты, первоначально полученные для линейных комбинаций, оказались применимыми 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 [58]. Получены оценки обобщающей способности и для 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 [70]. Для всех этих случаев оценки обобщающей способности выражаются через долю обучающих объектов с малым отступом.

Для успешной коррекции необходимо, чтобы базовые алгоритмы достаточно сильно различались. Если алгоритмы строятся независимо друг от друга, возникает опасность, что некоторые из них окажутся одинаковыми или почти одинаковыми. Существуют специальные техники, направленные на увеличение различий между базовыми алгоритмами.

В методе *баггинга* (bagging — сокращение от «bootstrap aggregation»), предложенном Л. Брейманом [43–45], 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) Р. Френда и И. Шапира [53, 55, 79] 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 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 [5, 6, 24]. В отличие от бустинга, здесь используется не выпуклая комбинация, а более сложная корректирующая операция в виде нелинейной монотонной функции достаточно общего вида. Монотонность можно рассматривать как обобщение выпуклости: любая выпуклая корректирующая операция является монотонной, обратное в общем случае неверно. При выпуклой коррекции вес каждого базового алгоритма остается постоянным на всем пространстве объектов, что представляется не вполне обоснованной эвристикой. Монотонная коррекция обладает существенно более богатыми возможностями для настройки. С другой стороны, для монотонных корректирующих операций, как для более широкого семейства функций, существенно выше опасность переобучения.

Обобщающая способность бустинга исследована, пожалуй, наиболее хорошо. Во многих случаях экспериментально наблюдается почти неограниченное улучшение качества обучения при наращивании числа алгоритмов в композиции [56]. 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 корректности на обучении [80]. С другой стороны, бустинг строит выпуклую комбинацию вещественнозначных классификаторов, которая проявляет свойство стабильности (см. ниже) [52].

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) [57]. Эмпирические исследования [83] на четырех 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 надежного 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 эффективно сужать изначально широкое семейство алгоритмов, подстраивать его под конкретную задачу.

Стабильность. Следующее, четвертое, направление исследований связано с понятием *стабильности* (stability) [41, 42, 66]. Метод обучения называется стабильным, если небольшие вариации обучающей выборки, такие как вставка или удаление одного объекта, приводят к незначительным изменениям получаемого алгоритма. Существуют различные способы формального определения стабильности, например, в работе [66] 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 [49, 50, 78]. Эти методы широко используются благодаря своей простоте, однако порождают семейства алгоритмов бесконечной емкости. 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.

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 этого использовались классические неравенства Хёфдинга и Бернштейна. Более точные результаты удается получать с помощью неравенств Чернова [47], метода ограниченных разностей МакДиармида [72] и изопериметрических неравенств Талагранда [85, 86]. Вводное изложение этих математических техник можно найти в обзорах [29, 68].

Скользящий контроль. Еще одно направление исследований связано с использованием *скользящего контроля* (cross-validation) [51, 65].

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 [65]. 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, LOO). Если используются все разбиения с контрольной выборкой фиксированной, но не обязательно единичной, длины, говорят об оценке полного скользящего контроля (complete cross-validation) [74]. Если генерируется случайное подмножество разбиений с контрольной выборкой фиксированной длины, говорят о бутстреп-оценке (bootstrap estimate) [28]. Если множество разбиений образуется k непересекающимися контрольными выборками, говорят о k-кратном скользящем контроле (k-fold cross-validation).

Несмотря на известную громоздкость, в некоторых случаях техника скользящего контроля непосредственно приводит к простым изящным результатам. В частности, для машин опорных векторов доказано, что LOO не превосходит доли опорных векторов во всей выборке. В силу несмещенности LOO отсюда немедленно вытекает, что вероятность ошибки не превосходит математического ожидания доли опорных векторов [48]. На практике частота ошибок на контроле часто оказывается еще в несколько раз меньше.

При исследовании локальных методов обучения использование функционала LOO становится естественной «вынужденной мерой» из-за очевидной смещенности эмпирического риска. В частности, для метода одного ближайшего соседа частота ошибок на обучении всегда равна нулю. Начиная с работ Девроя, Роджерса, Вагнера [49, 50, 78] функционал LOO остается одним из основных инструментов исследования стабильности методов обучения.

На практике скользящий контроль чаще всего применяется либо для выбора одной модели алгоритмов из нескольких (model selection) [61], либо для оптимизации небольшого числа параметров, определяющих структуру алгоритма, таких, как степень полинома, параметр регуляризации 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 смещенную заниженную оценку обобщающей способности. Причиной является все та же переподгонка, которая приводит к заниженности эмпирического риска [75]. Известно, что скользящий контроль дает несмещенную 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 [59, 61, 63], 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.

Причина этих неудач анализируется в [37], 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 емкости семейства и требующим дополнительных предположений о стабильности метода обучения [63]. 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 емкости семейства, а только от длины обучения и контроля. Данный результат проясняет природу скользящего контроля и показывает, что завышенность предыдущих оценок связана с неудачным выбором функционала качества.

Отсюда вытекают два важных вывода.

Во-первых, теория качества обучения может оказаться весьма чувствительной к исходной аксиоматике, в частности, к формализации самого понятия качества обучения.

Во-вторых, 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, развиваемом в настоящей работе.

Комбинаторный подход возник как попытка более точного построения статистической теории Вапника–Червоненкиса, начиная с исходных ее постулатов. Для этого имелись две основные предпосылки.

Первой предпосылкой было понимание того, что обобщающую способность целесообразно определять как частоту ошибок на конечной контрольной выборке, но не как вероятность ошибки, которая является величиной ненаблюдаемой, и которую невозможно вычислить точно. На практике любая обучаемая система сталкивается только с конечными выборками, будь то обучающие, контрольные или рабочие совокупности объектов. Использование гипотетических вероятностей может приводить (и реально приводит — см. «основную лемму» [3, стр. 219]) к лишним промежуточным шагам при доказательстве оценок и понижению их точности.

Второй предпосылкой было понимание того, что принцип минимизации эмпирического риска в заранее заданном семействе алгоритмов не достаточно точно описывает процесс обучения. На практике всегда используется конкретная процедура обучения, которая по заданной обучающей выборке строит единственный алгоритм, совсем не обязательно минимизирующий эмпирический риск.

В комбинаторном подходе явным образом вводится понятие *метода обучения*, по отношению к которому семейство алгоритмов становится вторичной конструкцией. Это позволяет рассматривать любые методы, а не только минимизацию эмпирического риска. Качество обучения по прецедентам (обобщающая способность метода) характеризуется комбинаторными функционалами, основанными на принципе скользящего контроля и зависящими только от метода обучения и заданной конечной выборки. В данной работе изучается несколько разновидностей функционала полного скользящего контроля. Соответствующие определения вводятся в § 4.

Верхние оценки комбинаторных функционалов аналогичны по своей структуре статистическим, но вместо сложности всего семейства в них фигурирует сложность *локального подсемейства*, состоящего из алгоритмов, которые могут быть выданы методом обучения в данной конкретной зада-

че. Наибольший интерес представляет случай, когда локальная сложность оказывается существенно меньше сложности всего семейства.

Комбинаторные оценки, полученные в § 5, справедливы для любого метода обучения и любой конечной выборки, не обязательно случайной, независимой, одинаково распределенной. Эти оценки выводятся комбинаторными методами без использования теории вероятностей. Оценки стандартных вероятностных функционалов оказываются непосредственным следствием комбинаторных. Это означает, что при переходе от статистической теории к комбинаторной соблюдается «принцип соответствия», а проблема качества обучения имеет скорее комбинаторную, чем вероятностную природу. Комбинаторная перестройка аксиоматики приводит к пересмотру многих положений статистической теории, некоторые из них рассматриваются в §§ 7–10.

В § 11 указываются три основные причины завышенности сложностных оценок качества. Наиболее существенной из них представляется погрешность, неизбежно возникающая при оценивании качества через сложность. Она связана с самой структурой сложностных оценок и присуща как вероятностным, так и комбинаторным оценкам. Две другие причины завышенности удается устранить с помощью комбинаторного подхода.

Универсальные ограничения. Ввиду принципиальной завышенности сложностных оценок можно выдвинуть предположение, что получить приемлемые численные результаты возможно только путем явного привлечения априорной информации о восстанавливаемой зависимости. Основная идея этого направления состоит в том, что если метод обучения строит алгоритмы, в некотором смысле «согласованные» с имеющейся априорной информацией, то для такого метода может существовать оценка обобщающей способности, существенно лучшая, чем в общем случае.

Отметим, что соответствие обучающей выборки (локальной информации) и априорных ограничений (универсальной информации) подробно изучается в теории универсальных и локальных ограничений К. В. Рудакова [13, 19–23] 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 алгоритмов. Однако оценки обобщающей способности в данной теории не рассматриваются. Вообще, проблема влияния априорной информации на качество восстановления зависимости представляется наименее изученной. В настоящей работе получены два результата в этом направлении.

В § 12 рассматривается метод ближайшего соседа в сочетании с априорной информацией о компактности классов. Полученное выражение функционала качества является точным и не зависит от сложностных характеристик семейства, которое, как известно, имеет бесконечную емкость.

В § 13 рассматривается класс методов, строящих монотонные алгоритмы, в сочетании с априорной информацией о монотонности или почти-монотонности восстанавливаемой зависимости. Соответствующая оценка качества также не зависит от сложности семейства, которое также имеет бесконечную емкость. Она является существенно более точной на малых выборках, чем оценки, полученные ранее другими авторами [27, 82].

В § 14 перечисляются некоторые проблемы, остающиеся открытыми в комбинаторном подходе.

Дополнением к приведенному выше обзору является периодически пополняемая частично аннотированная библиографическая база MachLearn, размещенная по адресу www.ccas.ru/frc. Настоящая работа содержит все доказательства, опущенные в кратком сообщении [7].

§ 2. Задача обучения по прецедентам

Имеется множество объектов X, множество ответов Y и множество $\mathfrak{A}$ отображений из X в Y, элементы которого будем называть алгоритмами, имея в виду, что они являются эффективно вычислимыми функциями. Предполагается, что существует фиксированное отображение $y^*\colon X \to Y$, не обязательно принадлежащее $\mathfrak{A}$, значения которого $y_i = y^*(x_i)$ известны только на объектах конечной обучающей выборки $X^l = \{x_1, \ldots, x_l\}$.

Задача обучения по прецедентам заключается в том, чтобы построить алгоритм $a^* \in \mathfrak{A}$, удовлетворяющий трем требованиям.

Во-первых, он должен выдавать на объектах обучающей выборки заданные ответы: $a^*(x_i) = y_i$, $i = 1, \ldots, l$. Равенство здесь может пониматься как точное или приближенное в зависимости от конкретной задачи. Требования такого вида называют *локальными ограничениями* [13], подчеркивая, что они связаны с конечным числом обучающих объектов и допускают эффективную проверку за конечное число шагов.

Во-вторых, на алгоритм a^* могут накладываться дополнительные ограничения общего характера, которым он должен удовлетворять как отображение, действующее из X в Y. Например, это могут быть ограничения симметричности, непрерывности, гладкости, монотонности, и т. д., а также их сочетания. Требования такого вида называют *универсальными ограничениями* [13], подчеркивая, что они не зависят от конкретной обучающей выборки и относятся к отображению «в целом». Как правило, они не допускают эффективной конечной проверки и учитываются в самой конструкции алгоритма на этапе его разработки. В общем случае универсальные ограничения выражаются условием $a^* \in \mathfrak{A}_u$, где $\mathfrak{A}_u$ — заданное подмножество алгоритмов, определяемое спецификой задачи.

В-третьих, искомый алгоритм a^* должен обладать способностью к обобщению, т. е. приближать восстанавливаемую зависимость y^* не только на объектах обучающей выборки, но и на всем множестве X. Данное требование можно формализовать с помощью различных функционалов качества, некоторые из которых будут рассмотрены ниже.

Частота ошибок алгоритма $a \in \mathfrak{A}$ на произвольной выборке объектов $X^p = \{x_1, \ldots, x_p\}$ есть

$$\nu(a, X^p) = \frac{1}{p} \sum_{i=1}^{p} I(x_i, a(x_i)),$$

где $I(x, y)$ — *индикатор ошибки*, принимающий значение 1, если ответ y является ошибочным для объекта x, и 0 — в противном случае. Выбор индикатора существенно зависит от конкретной задачи, в первую очередь от природы множества Y. В задачах классификации с двумя непересекающимися классами, когда $Y = \{0, 1\}$, обычно полагают

$$I(x, y) = |y - y^*(x)|.$$

В задачах восстановления регрессии, когда $Y = \mathbb{R}$, можно задать *)

$$I(x, y) = [|y - y^*(x)| \geqslant \delta(x)],$$

где $\delta(x)$ — фиксированная функция. Использование бинарного индикатора ошибки позволяет единообразно описывать широкий класс задач, включая и классификацию, и восстановление регрессии.

*) Здесь и далее квадратные скобки обозначают отображение логического результата в число: [ложь] = 0, [истина] = 1.

§ 3. О статистической теории восстановления зависимостей

Напомним вкратце основные положения статистической теории Вапника–Червоненкиса [2, 3, 87].

Предполагается, что множество объектов X является вероятностным пространством с некоторым неизвестным распределением вероятностей, и все рассматриваемые конечные наборы объектов выбираются случайно и независимо согласно этому распределению.

Задано семейство алгоритмов $A \subset \mathfrak{A}$. Из него выбирается алгоритм a^*, допускающий наименьшее число ошибок на заданной обучающей выборке X^l:

$$a^* = \arg\min_{a \in A} \nu(a, X^l).$$

Этот метод называется *минимизацией эмпирического риска.* В семействе может существовать много алгоритмов, минимизирующих эмпирический риск. Однако способ построения конкретного алгоритма в статистической теории не рассматривается, и предполагается, что в качестве решения может быть выбран любой из этих алгоритмов.

Качество алгоритма a^* характеризуется либо вероятностью ошибки, либо частотой ошибок $\nu(a^*, X^k)$ на контрольной выборке X^k, также случайной и независимой. Вычислить эту величину не представляется возможным, поэтому ставится задача определить условия, при которых она не сильно отличается от эмпирического риска. Достаточным условием является малое значение функционала равномерного отклонения частоты ошибок в двух выборках:

$$P_\varepsilon^{lk}(A) = \mathsf{P}\left\{\sup_{a \in A}(\nu(a, X^k) - \nu(a, X^l)) > \varepsilon\right\}.$$

Супремум вводится для того, чтобы оценить максимальное возможное отклонение, поскольку в общем случае заранее неизвестно, какой именно алгоритм будет получен в результате обучения.

Если $P_\varepsilon^{lk}(A) \to 0$ при $l, k \to \infty$, то говорят, что имеет место равномерная сходимость частоты ошибок в двух выборках. Равномерная сходимость является достаточным условием *обучаемости* семейства алгоритмов.

При $l = k$ для любого распределения вероятностей на множестве X и любой восстанавливаемой зависимости y^* справедлива следующая оценка скорости равномерной сходимости [3]:

$$P_\varepsilon^{lk}(A) \leqslant \Delta^A(2l) \cdot 1.5\, e^{-\varepsilon^2 l}, \tag{1}$$

где $\Delta^A(\cdot)$ — функция роста семейства алгоритмов A.

Определение 1. *Функцией роста* $\Delta^A(L)$ семейства A называется максимальное количество различных бинарных векторов вида $[I(x_i, a(x_i))]_{i=1}^L$, порождаемых всевозможными алгоритмами $a \in A$ на произвольной выборке $\{x_1, \ldots, x_L\}$.

Минимальное число h, при котором $\Delta^A(h) < 2^h$, называется *емкостью* или *VC-размерностью* семейства алгоритмов A. Если такого числа h не существует, то говорят, что емкость семейства бесконечна.

Если семейство A имеет конечную емкость h, то функция роста зависит от L полиномиально:

$$\Delta^A(L) \leqslant C_L^0 + C_L^1 + \ldots + C_L^h \leqslant 1.5\,\frac{L^h}{h!}. \tag{2}$$

В этом случае имеет место равномерная сходимость частот, и семейство является обучаемым. Таким образом, в статистической теории для оценивания

качества обучения по прецедентам достаточно знать только длину выборки и емкость семейства алгоритмов.

Получение оценок емкости для конкретных семейств алгоритмов является отдельной, зачастую довольно трудной, задачей. Практически сразу было доказано, что емкость семейства линейных решающих правил равна числу свободных параметров или, что то же самое, размерности линейного пространства, в котором строится разделяющая гиперплоскость. Оценки емкости получены также для нейронных сетей [31, 33, 60, 73]; решающих деревьев [8], корректных алгебраических замыканий подмодели АВО [16], комитетных решающих правил [71], и других семейств.

Емкость семейств, основанных на хранении всей обучающей последовательности, как правило, бесконечна (например, алгоритмов ближайших соседей, алгоритмов вычисления оценок — АВО). Емкость семейств, гарантирующих корректность алгоритма на обучающей выборке, также, как правило, бесконечна.

Вернемся к формуле (1). Имея зависимость $P_\varepsilon^{lk}(A)$ от ε, легко выразить из нее ε как функцию от емкости h, длины обучения l и желаемого значения функционала $\eta = P_\varepsilon^{lk}(A)$. При $l = k$ для любого распределения на множестве X и любого алгоритма a из семейства A справедливо неравенство

$$\nu(a, X^k) < \nu(a, X^l) + \sqrt{\frac{h}{l}\left(\ln\frac{2l}{h} + 1\right) - \frac{\ln\eta}{l}}. \qquad (3)$$

Первое слагаемое в этой оценке представляет эмпирический риск, убывающий с ростом емкости h. Второе слагаемое возрастает с ростом емкости, и его можно рассматривать как штраф за сложность (complexity penalty). Сумма в общем случае достигает минимума при некотором h.

Для определения оптимальной сложности модели алгоритмов был предложен метод *структурной минимизации риска*. Предположим, что в семействе A выделена последовательность подсемейств возрастающей емкости $A_1 \subset A_2 \subset \ldots \subset A_h = A$. Тогда в ней можно выбрать оптимальное подсемейство, для которого достигается минимальное значение правой части (3), и гарантировать заданное качество обучения.

К сожалению, практическому применению описанного подхода препятствует чрезвычайная завышенность оценок (1) и (3). Чтобы в этом убедиться, достаточно выполнить численный расчет требуемой длины обучающей выборки l как функции от (h, η, ε). Результаты расчета приведены в табл. 1.

Таблица 1

Достаточная длина обучающей выборки l как функция от емкости h, точности ε и значения функционала качества P_ε^{lk}.

P_ε^{lk}	0.01				1			
$h \setminus \varepsilon$	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2
0	60106	2404	601	150	14054	562	140	35
2	295074	9012	1946	408	245330	6963	1423	273
5	673222	19884	4192	848	623320	17823	3664	711
10	1307418	38160	7974	1589	1257471	36095	7444	1452
20	2579359	74855	15572	3082	2529396	72789	15043	2944
50	6401335	185193	38433	7575	6351365	183127	37903	7437
100	12775769	369275	76581	15075	12725798	367208	76051	14937

Правая половина таблицы, соответствующая значению $\eta = 1$, показывает границу применимости оценок Вапника–Червоненкиса. При меньших l верхняя оценка вероятности становится больше 1.

Первая строка таблицы соответствует другому крайнему случаю, когда $h=0$ и семейство состоит из единственного алгоритма. При этом достигается наилучшая возможная оценка.

Требуемая длина выборки существенно превышает количество объектов, с которыми приходится иметь дело на практике. В методе структурной минимизации риска завышенность оценок приводит к чрезмерному упрощению алгоритмов [62].

§ 4. Комбинаторные функционалы качества обучения

Принцип минимизации эмпирического риска в заранее заданном семействе алгоритмов является не достаточно точной формализацией процесса обучения.

Во-первых, не вполне ясно, где проходит граница семейства. Может оказаться так, что формально выписано очень широкое семейство, но на практике процедура обучения выдает алгоритмы лишь из небольшой его части.

Во-вторых, доставлять минимум эмпирическому риску могут многие алгоритмы, но в качестве решения всегда выбирается только один. Конкретизация метода его построения, возможно, позволила бы учесть специфические особенности процесса обучения.

В-третьих, далеко не все методы обучения минимизируют эмпирический риск. Тем не менее, многие из них неплохо зарекомендовали себя на практике, например, алгоритмы, использующие технику скользящего контроля или внешних критериев МГУА [14], алгоритмы, основанные на регуляризации эмпирического риска, алгоритмы явной максимизации отступа, алгоритмы бустинга и баггинга, и другие.

Определение 2. *Методом обучения* называется отображение μ, которое произвольной конечной обучающей выборке X^l ставит в соответствие определенный алгоритм $a=\mu(X^l)$. Говорят также, что метод μ строит алгоритм a по обучающей выборке X^l.

Будем полагать, что метод μ строит алгоритмы, выбирая их из некоторого семейства алгоритмов $A\subseteq\mathfrak{A}_u$. Будем также считать, что метод μ симметричен, т. е. результат $\mu(X^l)$ не изменяется при произвольной перестановке элементов обучающей выборки.

Определение 3. Алгоритм a называется *корректным* на выборке X^l, если $\nu(a, X^l)=0$. Метод μ называется *корректным* на выборке X^l, если алгоритм $\mu(X^l)$ корректен на X^l.

Малая частота ошибок $\nu(\mu(X^l), X^l)$ на заданной обучающей выборке X^l в общем случае не гарантирует, что построенный алгоритм будет столь же хорошо работать на остальных выборках.

Более того, частота ошибок $\nu(\mu(X^l), X^k)$ на контрольной выборке X^k, заданной независимо от X^l, и в общем случае с ней не пересекающейся, также не вполне адекватно характеризует качество обучения. Недостаток заключается в том, что фиксируется некоторое, вообще говоря, произвольное разбиение выборки $X^l\cup X^k$ на обучающую и контрольную части. Если значение $\nu(\mu(X^l), X^k)$ достаточно мало́, то нет гарантии, что при другом разбиении $X_1^l\cup X_1^k$ той же выборки значение $\nu(\mu(X_1^l), X_1^k)$ будет также мало́.

Из этих соображений вытекает естественное требование, чтобы функционал, характеризующий качество обучения по конечной выборке, был инвариантен относительно произвольных перестановок выборки.

Пусть l и k — произвольные фиксированные числа, $L = l + k$, и задана выборка $X^L = \{x_1, \ldots, x_L\}$. Обозначим через (X_n^l, X_n^k), $n = 1, \ldots, N$, всевозможные разбиения выборки X^L на обучающую и контрольную подвыборки длиной l и k соответственно. Число разбиений $N = C_L^l$.

Следующие функционалы характеризуют обобщающую способность метода μ по конечной совокупности объектов X^L.

Классический функционал полного скользящего контроля [65, 74]:

$$Q_c^{lk}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} \nu(\mu(X_n^l), X_n^k).$$

Функционал среднего отклонения частоты ошибок на контроле от частоты ошибок на обучении:

$$Q_d^{lk}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} (\nu(\mu(X_n^l), X_n^k) - \nu(\mu(X_n^l), X_n^l)).$$

Функционал среднего положительного отклонения частоты ошибок на контроле от частоты ошибок на обучении:

$$Q_{d+}^{lk}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} (\nu(\mu(X_n^l), X_n^k) - \nu(\mu(X_n^l), X_n^l))_+,$$

где $(z)_+ = z\,[z > 0]$ для любого действительного z.

Функционал скользящего контроля, терпимый к незначительной доле ошибок ε на контрольной подвыборке, $0 \leqslant \varepsilon \leqslant 1$:

$$Q_\varepsilon^{lk}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} [\nu(\mu(X_n^l), X_n^k) > \varepsilon].$$

Функционал скользящего контроля, терпимый к незначительным отклонениям частоты ошибок на контрольной выборке от частоты ошибок на обучении:

$$Q_{\nu,\varepsilon}^{lk}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} [\nu(\mu(X_n^l), X_n^k) - \nu(\mu(X_n^l), X_n^l) > \varepsilon].$$

Введем также функцию средней частоты ошибок на обучении:

$$\bar{\nu}_L^l(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} \nu(\mu(X_n^l), X_n^l).$$

Далее условимся опускать аргументы (μ, X^L) у функционалов, а также верхние индексы lk, показывающие, что эти функционалы зависят от соотношения числа обучающих и контрольных объектов.

Функционалы Q_c, Q_d, Q_{d+}, Q_ε, $Q_{\nu,\varepsilon}$, основанные на принципе скользящего контроля, будем называть *комбинаторными*, поскольку они определяются через множество всех разбиений выборки.

Непосредственно из определений вытекают следующие соотношения:

$$Q_c = Q_d + \bar{\nu}_L^l, \qquad Q_d \leqslant Q_{d+}, \qquad Q_{\nu,\varepsilon} \leqslant Q_\varepsilon.$$

Если метод μ корректен на всех подвыборках длины l, то

$$\bar{\nu}_L^l = 0, \qquad Q_c = Q_d = Q_{d+}, \qquad Q_{\nu,\varepsilon} = Q_\varepsilon.$$

Чуть менее очевидны следующие двусторонние оценки.

Лемма 1. *Для произвольных* μ, X^L *и* $\varepsilon \in [0, 1]$

$$\begin{aligned} \varepsilon Q_\varepsilon &< Q_c \leqslant \varepsilon + (1-\varepsilon) Q_\varepsilon, \\ \varepsilon Q_{\mu,\varepsilon} &< Q_{d+} \leqslant \varepsilon + (1-\varepsilon) Q_{\mu,\varepsilon}, \\ \varepsilon Q_{\mu,\varepsilon} &< Q_c \leqslant \varepsilon + (1-\varepsilon) Q_{\mu,\varepsilon} + \bar{\nu}_L^l. \end{aligned}$$

Доказательство. Первые две оценки вытекают непосредственно из определений и следующего двустороннего неравенства, справедливого при любых x и ε из $[0, 1]$:

$$\varepsilon[x > \varepsilon] < x \leqslant \varepsilon + (1-\varepsilon)[x > \varepsilon].$$

Третья оценка получается из первой и приведенных выше соотношений:

$$\varepsilon Q_{\mu,\varepsilon} \leqslant \varepsilon Q_\varepsilon < Q_c = Q_d + \bar{\nu}_L^l \leqslant Q_{d+} + \bar{\nu}_L^l \leqslant \varepsilon + (1-\varepsilon) Q_{\mu,\varepsilon} + \bar{\nu}_L^l.$$

Лемма доказана.

Перечисленные неравенства позволяют говорить о взаимозаменяемости функционалов. Выбор конкретного функционала не столь принципиален и может определяться априорными предпочтениями или удобством вывода оценок.

§ 5. Локальная сложность и оценки качества обучения

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. Этот эффект будем называть локализацией семейства алгоритмов. Наибольший интерес представляют ситуации, когда сложность локального подсемейства алгоритмов оказывается существенно меньше сложности всего семейства A.

Существование эффекта локализации снимает искусственный запрет на использование сложных алгоритмов. Важно не столько ограничить емкость семейства, сколько разработать метод обучения, способный подстраиваться под конкретные задачи, всякий раз по-разному локализуя «рабочую область» семейства. При фиксации восстанавливаемой зависимости метод обучения должен строить алгоритмы, «похожие» на нее. Тогда не важно, сколько еще «не похожих» алгоритмов содержится в семействе. Это свойство предлагается называть *локализующей способностью* метода обучения, подчеркивая, что оно является важной составной частью его обобщающей способности.

Определение 4. *Локальным семейством алгоритмов*, порожденным методом μ на выборке X^L, называется множество алгоритмов

$$A_L^l(\mu, X^L) = \{\mu(X_n^l) \mid n = 1, \ldots, N\}, \quad \text{где } N = C_L^l.$$

Определение 5. *Локальной функцией роста* $\Delta_L^l(\mu, X^L)$ метода μ на выборке X^L называется число различных бинарных векторов вида $[I(x_i, a(x_i))]_{i=1}^L$, порождаемых всевозможными алгоритмами a из $A_L^l(\mu, X^L)$.

Определение 6. *Степенью некорректности* метода μ на выборке X^L называется максимальная частота ошибок на всевозможных обучающих подвыборках длины l:

$$\sigma_L^l(\mu, X^L) = \max_{n=1,\ldots,N} \nu(\mu(X_n^l), X_n^l).$$

В дальнейшем будем использовать сокращенные обозначения Δ_L^l, A_L^l и σ_L^l, опуская аргументы (μ, X^L).

Локальная функция роста существенно отличается от функции роста всего семейства $\Delta^A(L)$. Она зависит от конкретной выборки, метода обучения и соотношения чисел l и k. Для нее тривиальное ограничение сверху есть $A_L^l \leqslant C_L^l$, в то время как $\Delta^A(L) \leqslant 2^L$. Очевидно также, что $A_L^l \leqslant \Delta^A(L)$.

Теорема 1. *Пусть метод μ имеет на выборке X^L степень некорректности $\sigma = \sigma_L^l(\mu, X^L)$. Тогда для любого $\varepsilon \in [0, 1)$ справедлива оценка*

$$Q_{\nu,\varepsilon}^{lk}(\mu, X^L) < \Delta_L^l(\mu, X^L) \cdot \Gamma_L^l(\varepsilon, \sigma), \tag{4}$$

где функция $\Gamma_L^l(\varepsilon, \sigma)$ определяется следующим образом:

$$\Gamma_L^l(\varepsilon, \sigma) = \max_{m \in M(\varepsilon,\sigma)} \sum_{s \in S(\varepsilon,\sigma)} \frac{C_m^s C_{L-m}^{l-s}}{C_L^l},$$

$$M(\varepsilon, \sigma) = \{m \mid \varepsilon k < m \leqslant k + \sigma l\},$$

$$S(\varepsilon, \sigma) = \{s \mid \max(0, m-k) \leqslant s \leqslant \sigma l,\ s < (m - \varepsilon k) l/L\}.$$

Доказательство в целом следует Вапнику [3], за исключением того, что оценивается комбинаторный функционал $Q_{\nu,\varepsilon}^{lk}$ вместо вероятностного P_ε^{lk}, и учитывается степень некорректности метода.

Введем на множестве A_L^l отношение эквивалентности, положив для произвольных a и a' из A_L^l

$$a \sim a' \iff (\forall x \in X^L)\, I(x, a) = I(x, a'),$$

т. е. алгоритмы эквивалентны, если они допускают ошибки на одних и тех же объектах выборки X^L. Это отношение разбивает множество A_L^l на классы, обозначаемые далее A_{mi}, где m, $m = 0, \ldots, L$, — число ошибок, допускаемых на выборке X^L алгоритмами данного класса, i, $i = 1, \ldots, \Delta_m$, — порядковый номер класса среди всех классов, алгоритмы которых допускают m ошибок, Δ_m — число способов получить m ошибок на выборке X^L всевозможными алгоритмами из A_L^l. Число всех классов эквивалентности равно локальной функции роста метода μ на выборке X^L:

$$\Delta_L^l = \Delta_0 + \Delta_1 + \ldots + \Delta_L. \tag{5}$$

Эквивалентность на алгоритмах порождает эквивалентность на разбиениях, если для произвольных n и u из $\{1, \ldots, N\}$ положить $n \sim u \iff \mu(X_n^l) \sim \mu(X_u^l)$. При этом образуются классы эквивалентности $N_{mi} = \{n \mid \mu(X_n^l) \in A_{mi}\}$ на множестве разбиений, взаимно однозначно соответствующие классам A_{mi}.

Запишем функционал качества, суммируя разбиения отдельно по каждому классу эквивалентности:

$$Q_{\nu,\varepsilon}^{lk} = \frac{1}{N} \sum_{m=0}^{L} \sum_{i=1}^{\Delta_m} \sum_{n \in N_{mi}} [\nu(\mu(X_n^l), X_n^k) > \nu(\mu(X_n^l), X_n^l) + \varepsilon].$$

Значение функционала не изменится, если алгоритм $\mu(X_n^l)$, $n \in N_{mi}$, заменить на произвольный элемент a_{mi} из класса A_{mi}. Учтем также, что при $m \leqslant \varepsilon k$ и при $m > k + \sigma l$ под знаком суммы оказывается нуль, поэтому суммирование достаточно проводить только по $m \in M(\varepsilon, \sigma)$:

$$Q_{\nu,\varepsilon}^{lk} = \sum_{m \in M(\varepsilon,\sigma)} \sum_{i=1}^{\Delta_m} \underbrace{\frac{1}{N} \sum_{n \in N_{mi}} [\nu(a_{mi}, X_n^k) > \nu(a_{mi}, X_n^l) + \varepsilon]}_{\gamma_{mi}}. \tag{6}$$

Оценим сверху внутреннюю сумму γ_{mi}, заменив класс эквивалентности N_{mi} множеством всех разбиений. Обозначим через s число ошибок на обучающей подвыборке, $0 \leqslant s \leqslant \sigma l$, и просуммируем разбиения отдельно при каждом s:

$$\gamma_{mi} \leqslant \frac{1}{N}\sum_{n=1}^{N}[\nu(a_{mi}, X_n^k) > \nu(a_{mi}, X_n^l) + \varepsilon] = \sum_{s=0}^{\sigma l}\left[\frac{m-s}{k} > \frac{s}{l} + \varepsilon\right]\frac{1}{N}\sum_{n=1}^{N}\left[\nu(a_{mi}, X_n^l) = \frac{s}{l}\right].$$

Внутренняя сумма равна $C_m^s C_{L-m}^{l-s}$ — числу разбиений выборки длины L на две подвыборки таких, что из m ошибок в подвыборку длины l попадают s ошибок. Таким образом,

$$\gamma_{mi} \leqslant \sum_{s \in S(\varepsilon, \sigma)} \frac{C_m^s C_{L-m}^{l-s}}{C_L^l}.$$

Эта величина уже не зависит от i, поэтому в (6) ее можно вынести за знак суммирования по i. Используя (5), приходим к неравенству

$$Q_{\nu, \varepsilon}^{lk} \leqslant \sum_{m \in M(\varepsilon, \sigma)} \Delta_m \gamma_{mi} \leqslant \Delta_L^l \max_{m \in M(\varepsilon, \sigma)} \gamma_{mi}.$$

Подставляя сюда оценку γ_{mi}, получаем требуемое.

Теорема доказана.

Локальная функция роста Δ_L^l не превосходит функцию роста всего семейства $\Delta(A)$. Согласно [3], комбинаторный множитель $\Gamma_L^l(\varepsilon, 1)$ не превосходит $1.5\, e^{-\varepsilon^2 l}$ при $l = k$. Отметим, что в общем случае нет оснований приравнивать l и k, за исключением удобства вывода экспоненциальной верхней оценки комбинаторного множителя.

Следствие 1. *При $l = k$ для любых μ и X^L справедлива оценка функционала $Q_{\nu, \varepsilon}^{lk}$ по Вапнику—Червоненкису с точностью до замены функции роста всего семейства на локальную функцию роста:*

$$Q_{\nu, \varepsilon}^{lk}(\mu, X^L) < 1.5\, \Delta_L^l(\mu, X^L) e^{-\varepsilon^2 l} \leqslant 1.5\, \Delta^A(L) e^{-\varepsilon^2 l}. \tag{7}$$

Усиление оценки достигается, главным образом, благодаря модификации исходного функционала и отказу от избыточного требования равномерной сходимости. Данный результат впервые упоминался в работе [4].

В соответствии с леммой 1 оценки, аналогичные (4), справедливы и для других комбинаторных функционалов. Более аккуратная техника позволяет несколько уточнить верхние оценки функционалов Q_{d+}^{lk} и Q_c^{lk}.

Теорема 2. *Пусть метод μ имеет на выборке X^L степень некорректности $\sigma = \sigma_L^l(\mu, X^L)$. Тогда для любого $\varepsilon \in [0, 1)$ справедлива оценка*

$$Q_{d+}^{lk}(\mu, X^L) < \varepsilon + \Delta_L^l(\mu, X^L) \cdot \widetilde{\Gamma}_L^l(\varepsilon, \sigma), \tag{8}$$

где

$$\widetilde{\Gamma}_L^l(\varepsilon, \sigma) = \max_{m \in M(\varepsilon, \sigma)} \sum_{s \in S(\varepsilon, \sigma)} \left(\frac{ml - sL}{lk} - \varepsilon\right) \frac{C_m^s C_{L-m}^{l-s}}{C_L^l},$$

и множества $M(\varepsilon, \sigma)$ и $S(\varepsilon, \sigma)$ определяются так же, как в теореме 1.

Доказательство. При любых $x \in \mathbb{R}$ и $\varepsilon \geqslant 0$ имеет место неравенство $(x)_+ \leqslant \varepsilon + (x-\varepsilon)_+$. Следовательно, справедлива верхняя оценка

$$Q_{d+}^{lk} \leqslant \varepsilon + \frac{1}{N} \sum_{n=1}^{N} (\nu(\mu(X_n^l), X_n^k) - \nu(\mu(X_n^l), X_n^l) - \varepsilon)_+.$$

Аналогично предыдущей теореме введем классы эквивалентных разбиений N_{mi} и классы эквивалентных алгоритмов A_{mi}, выбрав по одному представителю $a_{mi} \in A_{mi}$, где $m = 0, \ldots, L$, $i = 1, \ldots, \Delta_m$. Запишем функционал Q_{d+}^{lk}, суммируя разбиения отдельно по классам эквивалентности:

$$Q_{d+}^{lk} \leqslant \varepsilon + \frac{1}{N} \sum_{m=0}^{L} \sum_{i=1}^{\Delta_m} \sum_{n \in N_{mi}} (\nu(a_{mi}, X_n^k) - \nu(a_{mi}, X_n^l) - \varepsilon)_+.$$

При $m \leqslant \varepsilon k$ и при $m > k + \sigma l$ под знаком суммы оказывается нуль, поэтому суммирование достаточно проводить только по $m \in M(\varepsilon, \sigma)$:

$$Q_{d+}^{lk} \leqslant \varepsilon + \sum_{m \in M(\varepsilon,\sigma)} \sum_{i=1}^{\Delta_m} \underbrace{\frac{1}{N} \sum_{n \in N_{mi}} (\nu(a_{mi}, X_n^k) - \nu(a_{mi}, X_n^l) - \varepsilon)_+}_{\widetilde{\gamma}_{mi}}. \qquad (9)$$

Оценим сверху внутреннюю сумму $\widetilde{\gamma}_{mi}$, заменив класс эквивалентности N_{mi} множеством всех разбиений. Обозначив через s число ошибок на обучающей подвыборке, $0 \leqslant s \leqslant \sigma l$, просуммируем разбиения отдельно при каждом s:

$$\widetilde{\gamma}_{mi} \leqslant \frac{1}{N} \sum_{n=1}^{N} (\nu(a_{mi}, X_n^k) - \nu(a_{mi}, X_n^l) - \varepsilon)_+ =$$
$$= \sum_{s=0}^{\sigma l} \left[\frac{m-s}{k} > \frac{s}{l} + \varepsilon\right] \left(\frac{m-s}{k} - \frac{s}{l} - \varepsilon\right) \frac{1}{N} \sum_{n=1}^{N} \left[\nu(a_{mi}, X_n^l) = \frac{s}{l}\right].$$

Внутренняя сумма равна $C_m^s C_{L-m}^{l-s}$ — числу разбиений выборки длины L на две подвыборки таких, что из m ошибок в подвыборку длины l попадают s ошибок. Таким образом,

$$\widetilde{\gamma}_{mi} \leqslant \sum_{s \in S(\varepsilon,\sigma)} \left(\frac{m-s}{k} - \frac{s}{l} - \varepsilon\right) \frac{C_m^s C_{L-m}^{l-s}}{C_L^l}.$$

Эта величина уже не зависит от i, поэтому в (9) ее можно вынести за знак суммирования по i. Используя (5), приходим к неравенству

$$Q_{d+}^{lk} \leqslant \varepsilon + \sum_{m \in M(\varepsilon,\sigma)} \Delta_m \widetilde{\gamma}_{mi} \leqslant \varepsilon + \Delta_L^l \max_{m \in M(\varepsilon,\sigma)} \widetilde{\gamma}_{mi}.$$

Подставляя сюда оценку $\widetilde{\gamma}_{mi}$, получаем требуемое неравенство.

Теорема доказана.

Из доказанной теоремы и неравенства $Q_c^{lk} \leqslant Q_{d+}^{lk} + \bar{\nu}_L^l$ немедленно вытекает верхняя оценка функционала скользящего контроля.

Следствие 2. *Пусть метод μ имеет на выборке X^L степень некорректности $\sigma = \sigma_L^l(\mu, X^L)$. Тогда для любого $\varepsilon \in [0, 1)$ справедлива оценка*

$$Q_c^{lk}(\mu, X^L) < \bar{\nu}_L^l + \varepsilon + \Delta_L^l(\mu, X^L) \cdot \widetilde{\Gamma}_L^l(\varepsilon, \sigma). \qquad (10)$$

Оценки (8) и (10) содержат искусственно введенный параметр ε. Чтобы избавиться от него, необходимо решить дополнительную задачу минимизации данных оценок по ε.

§ 6. О вероятностных функционалах и «принципе соответствия»

Полученные результаты свидетельствуют о возможности построения не-вероятностной теории качества обучения по прецедентам.

В отличие от функционала Вапника–Червоненкиса P_ε^{lk}, комбинаторные функционалы зависят от метода обучения и конкретной выборки, которая не обязана быть случайной. Если же снова предположить, что X — вероятностное пространство, X^L — случайная, независимая, одинаково распределенная выборка, то математическое ожидание комбинаторных функционалов принимает форму вероятностных функционалов качества:

$$\mathsf{E}Q_c^{lk}(\mu, X^L) = \mathsf{P}\{I(\mu(X^l), x) = 1\},$$
$$\mathsf{E}Q_\varepsilon^{lk}(\mu, X^L) = \mathsf{P}\{\nu(\mu(X^l), X^k) > \varepsilon\},$$
$$\mathsf{E}Q_{и,\varepsilon}^{lk}(\mu, X^L) = \mathsf{P}\{\nu(\mu(X^l), X^k) - \nu(\mu(X^l), X^l) > \varepsilon\}.$$

Первая строка выражает хорошо известный факт, что скользящий контроль Q_c дает несмещенную оценку вероятности ошибки [3]. Остальные комбинаторные функционалы также являются несмещенными оценками соответствующих вероятностных функционалов. Они не содержат избыточно сильного требования равномерной сходимости, поэтому характеризуют качество обучения даже более точно, чем функционал P_ε^{lk}. Непосредственно из определений следует, что P_ε^{lk} является завышенной верхней оценкой $\mathsf{E}Q_{и,\varepsilon}^{lk}$, который, собственно, и описывает качество обучения:

$$\mathsf{E}Q_{и,\varepsilon}^{lk}(\mu, X^L) \leqslant P_\varepsilon^{lk}(A).$$

Любая верхняя оценка комбинаторного функционала легко преобразуется в верхнюю оценку соответствующего вероятностного функционала путем применения операции матожидания к обеим частям неравенства. В частности, из (4) получается более точный вариант оценки Вапника–Червоненкиса:

$$\mathsf{P}\{\nu(\mu(X^l), X^k) - \nu(\mu(X^l), X^l) > \varepsilon\} \leqslant \mathsf{E}\Delta_L^l(\mu, X^L) \cdot \Gamma_L^l(\varepsilon, \sigma).$$

Перечисленные факты позволяют говорить о соблюдении «принципа соответствия» при переходе от статистической теории Вапника–Червоненкиса к более точной теории качества обучения, основанной на анализе комбинаторных функционалов.

Оценки статистической теории выводились при условии, что распределение вероятностей на множестве объектов существует, но неизвестно. Теперь оказывается, что они остаются верны, если просто полагать выборку произвольной, не требуя от нее случайности, независимости и одинаковой распределенности. Заметим, что использование вероятностных функционалов качества может приводить к лишним промежуточным шагам при выводе оценок и понижению их точности (типичный пример — «основная лемма» в статистической теории [3, стр. 219]).

Неожиданным на первый взгляд представляется отказ от требования независимости выборки. В теории вероятностей независимость означает инвариантность вероятностной меры относительно всевозможных перестановок выборки. При доказательстве теоремы 1 ту же роль играет инвариантность функционала качества относительно всевозможных перестановок выборки (свойство симметричности функционала). Это требование можно считать слабой формой гипотезы независимости, при которой ограничение переносится с исходных данных на функционал качества. Заметим, что все введенные выше комбинаторные функционалы симметричны.

Таким образом, природа оценок (4) и (8) является не вероятностной, а исключительно комбинаторной, и вытекает из дискретности индикатора ошибки $I(x, y)$ и симметричности функционала качества.

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.

§ 7. О важности требования корректности

Полученные выше комбинаторные оценки зависят от степени некорректности σ. В теории Вапника–Червоненкиса рассматривались только крайние случаи: $\sigma = 0$ (детерминистская постановка задачи) и $\sigma = 1$. Интересно исследовать промежуточные ситуации, когда $0 < \sigma < 1$.

Комбинаторный множитель $\Gamma_L^l(\varepsilon, \sigma)$ монотонно не убывает по σ. Наименьшее значение достигается при $\sigma = 0$, когда метод обучения является корректным. Наибольшее значение достигается при $\sigma = 1$, когда нет никакого априорного знания о количестве ошибок, допускаемых на обучении.

В случае корректности комбинаторный множитель принимает вид

$$\Gamma_L^l(\varepsilon, 0) = \frac{C_{L-\lceil \varepsilon k \rceil}^l}{C_L^l} \leqslant \left(\frac{k}{L}\right)^{\varepsilon k}.$$

Если h — емкость локального подсемейства, то из (2) и теоремы 1 следует оценка качества корректного метода обучения:

$$Q_{u,\varepsilon}^{lk} < (C_L^0 + C_L^1 + \ldots + C_L^h)\frac{C_{L-\lceil \varepsilon k \rceil}^l}{C_L^l}.$$

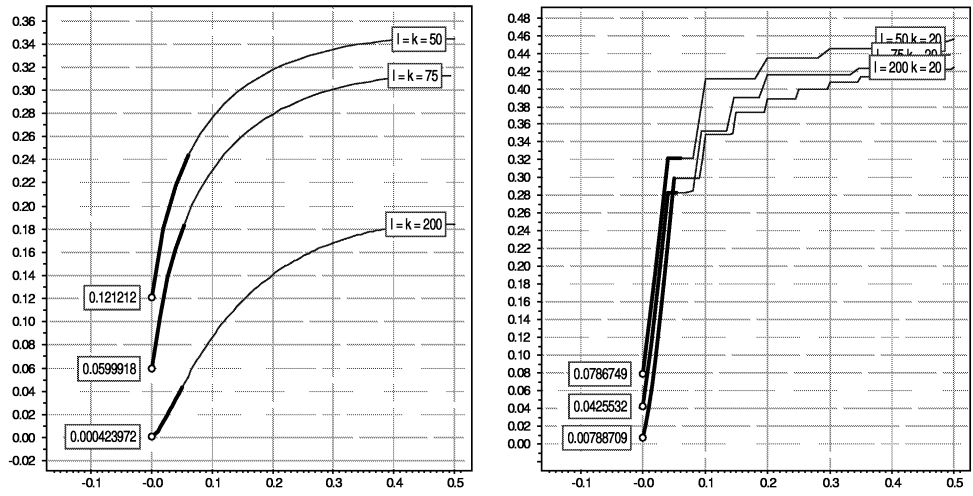
Ниже приводится таблица с результатами численного расчета требуемой длины обучающей выборки l по полученному соотношению. Эти значения существенно лучше приведенных в табл. 1.

Таблица 2

Достаточная длина обучающей выборки l как функция от емкости h, точности ε и значения функционала качества Q_ε^{lk} в случае корректного метода обучения.

Q_ε^{lk}	0.01				1			
$h \setminus \varepsilon$	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2
0	800	160	80	40	200	40	10	5
2	2900	460	200	85	2100	300	130	50
5	6300	980	420	170	5500	820	340	130
10	12000	1840	780	315	11200	1680	700	275
20	23500	3560	1510	600	22800	3420	1430	560
50	58200	8780	3710	1470	57400	8620	3630	1430
100	107000	17500	7380	2920	107000	17340	7300	2880

По мере увеличения некорректности σ комбинаторный множитель $\Gamma_L^l(\varepsilon, \sigma)$ возрастает настолько быстро, что достигает значительной величины, сравнимой с $\Gamma_L^l(\varepsilon, 1)$, уже при $\sigma \approx \varepsilon$ (см. рис. 1)



(а) при $l = k = 50, 75, 200$ (б) при $l = 50, 75, 200$; $k = 20$

Рис. 1. График зависимости $\Gamma_L^l(\varepsilon, \sigma)$ от степени некорректности σ при $\varepsilon = 0.05$. Жирной линией выделены участки $\sigma \in [0, \varepsilon]$.

Таким образом, получить приемлемые (хотя бы не превышающие 10^3) численные оценки достаточной длины обучения возможно только при условии корректности, и только для семейств небольшой локальной емкости.

Комбинаторный подход позволяет по-новому взглянуть на проблему построения корректных алгоритмов.

Очевидно, для обеспечения корректности необходимо усложнять конструкцию алгоритмов. Согласно статистической теории это приводит к значительному увеличению функции роста, на фоне которого эффект уменьшения комбинаторного множителя остается незаметным. Отсюда в статистической теории 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.

Остаются открытыми вопросы — как получить нетривиальные оценки локальной функции роста, и какие методы обучения обладают способностью локализовать семейства алгоритмов.

§ 8. О функционале равномерного относительного отклонения

Авторы статистической теории прекрасно понимали, что требование равномерной сходимости является чрезмерно сильным. Чтобы оценить частоту ошибок на контроле $\nu(a, X^k)$ по частоте ошибок на обучении $\nu(a, X^l)$, достаточно потребовать равномерной сходимости не по всему семейству, а только в области минимальных частот. Построить эту область в явном виде не представляется возможным, поэтому был введен функционал равномерного относительного отклонения частот в двух подвыборках, и для него была получена оценка [3]:

$$\mathsf{P}\left\{\sup_{a \in A} \frac{\nu(a, X^k) - \nu(a, X^l)}{\sqrt{\nu(a, X^L)}} > \varepsilon\right\} < \Delta^A(L) \max_{m_0 \leqslant m \leqslant m_1} \sum_{s = s_0(m)}^{s_1(m)} \frac{C_m^s C_{L-m}^{l-s}}{C_L^l},$$

где $m_0 = \left\lceil \frac{(\varepsilon k)^2}{L} \right\rceil$, $m_1 = L$, $s_0(m) = \max(0, m-k)$, $s_1(m) = \left\lfloor \frac{l}{L}\left(m - \varepsilon k \sqrt{\frac{m}{L}}\right) \right\rfloor$.

Нетрудно показать, что эта оценка выводится из той же комбинаторной формулы (4) путем замены переменной $\varepsilon\sqrt{\frac{m}{L}} \to \varepsilon$. Оценки обоих функционалов, абсолютного и относительного — это просто два разных способа оценить сверху комбинаторный множитель. По сути дела, оценка равномерного относительного отклонения является небольшим техническим усовершенствованием, и не описывает эффект сужения семейства. Для его описания необходимо вводить локальную функцию роста.

§ 9. О структурной минимизации риска

Комбинаторные функционалы качества имеют неоспоримое преимущество перед вероятностными — их можно измерять по выборке. Следовательно, при выборе оптимальной структуры алгоритма в методе структурной минимизации риска можно 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 емкости теряет смысл. Вместо этого достаточно брать конечный набор методов обучения $\mu_1, \ldots, \mu_T$ и выбирать из них лучший по критерию скользящего контроля. Эмпирические исследования показывают, что данная техника выбора модели во многих случаях предпочтительнее явной оптимизации сложности [62]. В частности, там показано, что принцип структурной минимизации риска склонен переупрощать, а принцип минимальной длины описания (minimum description length) [77] — переусложнять модель.

Возникает вопрос: нужны ли вообще завышенные оценки качества, если лучше обходиться без них?

§ 10. Об эффективной емкости

В работах [38, 88] было введено понятие эффективной емкости и показано, что статистические оценки остаются верны, если в них заменить емкость на эффективную емкость. Там же был предложен метод ее измерения по заданной выборке для задач классификации с двумя классами.

Целесообразность эмпирического измерения емкости связана с двумя обстоятельствами. Во-первых, не для всех семейств алгоритмов удается получить аналитические оценки емкости. Во-вторых, в конкретных задачах эффективная емкость может оказаться существенно меньше полной емкости семейства.

Идея метода заключается в том, чтобы при различных длинах выборки l измерить функционал в левой части неравенства

$$Q^{lk}_{\sup}(A) = \frac{1}{N}\sum_{n=1}^{N}\Big[\sup_{a\in A}(\nu(a, X^k_n) - \nu(a, X^l_n)) > \varepsilon\Big] < C\frac{L^h}{h!}\cdot e^{-\varepsilon^2 l},$$

где C — некоторая константа и $k = l$.

Далее делается предположение, что зависимость левой части от длины выборки имеет при некотором значении параметра h такое же алгебраическое выражение, что и правая часть. Это значение h и называется *эффективной емкостью*. Предположение хорошо подтверждается в случае линейных решающих правил [88].

Для измерения $Q^{lk}_{\sup}(A)$ был придуман изящный метод избежать вычисления супремума. В случае классификации на два класса максимизация разности $\nu(a, X^k_n) - \nu(a, X^l_n)$ эквивалентна минимизации суммы $\nu(a, \widetilde{X}^k_n) + \nu(a, X^l_n)$, где выборка $\widetilde{X}^k_n$ получается из X^k_n путем замены исходных классификаций на ошибочные. Если метод μ минимизирует эмпирический риск, то алгоритм $a_n = \mu(\widetilde{X}^k_n \cup X^l_n)$ доставляет разности частот максимальное значение. Тогда

$$Q^{lk}_{\sup}(A) = \frac{1}{N} \sum_{n=1}^{N} [\nu(a_n, X^k_n) - \nu(a_n, X^l_n) > \varepsilon]. \tag{11}$$

Собственно измерение заключается в том, чтобы оценить данную сумму по меньшему числу разбиений, выбранных случайным образом. Погрешность такого измерения легко оценивается в соответствии с законом больших чисел.

Эффективная емкость позволяет учесть особенности распределения объектов, но не учитывает особенностей восстанавливаемой зависимости и метода обучения, поскольку алгоритм специально обучают делать ошибки. Для случая линейных разделяющих правил эффективная емкость с высокой точностью равна размерности подпространства, в котором лежит выборка [88].

С позиций комбинаторного подхода очевидно, что функционал равномерного отклонения $Q^{lk}_{\sup}(A)$ должен быть заменен функционалом скользящего контроля $Q^{lk}_{\nu,\varepsilon}$. В этом случае процедура измерения (11) остается той же, с тем отличием, что теперь $a_n = \mu(X^l_n)$, т. е. искусственного привнесения ошибок в обучающую выборку уже не требуется.

При этом возникает новое понятие — *локальная эффективная емкость*. Это такое значение параметра h, при котором зависимость $Q^{lk}_{\nu,\varepsilon}$ от L наилучшим образом аппроксимируется формулой

$$Q^{lk}_{\nu,\varepsilon}(\mu, X^L) \approx C \frac{L^h}{h!} \cdot \Gamma^l_L(\varepsilon, \sigma).$$

В отличие от эффективной емкости Вапника, локальная эффективная емкость учитывает все три фактора: особенности распределения объектов, особенности восстанавливаемой зависимости и особенности метода обучения.

Сравнение емкости с измеренным значением эффективной емкости позволяет оценить, насколько хорошо метод «улавливает» эффективную размерность пространства объектов [88].

Сравнительное измерение эффективной емкости и локальной эффективной емкости позволяет оценить, насколько существенен эффект локализации, т. е. насколько хорошо данный метод обучения подстраивается под конкретную зависимость на конкретной выборке.

§ 11. О причинах завышенности сложностных оценок

Для анализа причин завышенности сложностных оценок качества представим отношение правой и левой частей неравенства (7) в следующем виде:

$$\frac{\Delta(A) \cdot 1.5\, e^{-\varepsilon^2 l}}{Q^{lk}_{\nu,\varepsilon}} = \frac{\Delta(A)}{\Delta^l_L} \cdot \frac{1.5\, e^{-\varepsilon^2 l}}{\Gamma^l_L} \cdot \frac{\Delta^l_L \Gamma^l_L}{Q^{lk}_{\nu,\varepsilon}}.$$

В каждой из дробей числитель является верхней оценкой знаменателя. Три сомножителя в правой части равенства описывают соответственно три основные причины завышенности сложностных оценок качества.

Первая причина — пренебрежение эффектом локализации.

Вторая причина — относительная погрешность экспоненциальной оценки комбинаторного множителя, которая, в отличие от абсолютной погрешности, увеличивается с ростом длины выборки. Если ставить целью получение оценок, непосредственно применимых на практике, то придется смириться с необходимостью вычисления или табулирования достаточно сложных комбинаторных выражений.

Третья причина — погрешность разложения функционала скользящего контроля в произведение локальной функции роста Δ_L^l и комбинаторного множителя Γ_L^l. Эта причина представляется наиболее существенной, поскольку она вызвана переходом от анализа качества к анализу сложности и связана с самой природой сложностных оценок. Она в одинаковой степени относится к вероятностным и комбинаторным оценкам, основанным на функции роста.

Перспективным подходом к получению существенно более точных оценок представляется полный отказ от использования сложностных характеристик. Оценки такого вида известны для стабильных алгоритмов [42] и выпуклых комбинаций классификаторов [35]. В этих работах учитывались специфические особенности метода обучения. Выборка, как обычно, считалась случайной и независимой при произвольном распределении вероятностей и произвольной восстанавливаемой зависимости. Полученные оценки все еще сильно завышены, а достаточная длина обучения составляет порядка 10^4 объектов.

По всей видимости, получение оценок, непосредственно применимых на практике, невозможно без априорных предположений о свойствах выборки и восстанавливаемой зависимости. Ниже рассматриваются два частных случая: компактность и монотонность. В первом случае удается получить точное выражение функционала качества вместо верхней оценки.

§ 12. Априорные ограничения компактности

Пусть на множестве X определена полуметрика $\rho(x, x')$. Рассмотрим метод обучения μ, который строит алгоритмы одного ближайшего соседа $a = \mu(X^l)$, работающие следующим образом:

$$a(x) = y^*(\arg\min_{x' \in X^L} \rho(x, x')) \quad \text{для всех } x \in X.$$

Точное выражение функционала скользящего контроля Q_c для алгоритма k ближайших соседей и некоторых его модификаций найдено в [74]. Авторы этой работы ставили целью вывод эффективных вычислительных формул. Воспользуемся этим результатом, чтобы ввести характеристику выборки, выражающую априорную информацию о компактности классов.

Для каждого объекта x_i, $i = 1, \ldots, L$ выборки X^L расположим остальные $L-1$ объектов в порядке возрастания расстояния до x_i, пронумеровав их двойными индексами:

$$x_i = x_{i0}, \quad x_{i1}, x_{i2}, \ldots x_{i,L-1}.$$

Таким образом,

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

Обозначим через $r_m(x_i)$ ошибку, возникающую, если правильный ответ $y^*(x_i)$ на объекте x_i заменить ответом на его m-м соседе:

$$r_m(x_i) = I(x_i, y^*(x_{im})), \qquad i = 1, \ldots, L, \quad m = 1, \ldots, L-1.$$

Определение 7. *Профилем компактности* выборки X^L называется функция $K(m, X^L)$, выражающая долю объектов выборки, для которых правильный ответ не совпадает с правильным ответом на m-м соседе:

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

Теорема 3. *Для задачи классификации методом ближайшего соседа справедливо следующее точное выражение функционала* Q_c:

$$Q_c^{lk}(\mu, X^L) = \sum_{m=1}^{k} K(m, X^L) \frac{C_{L-1-m}^{l-1}}{C_{L-1}^{l}}. \tag{12}$$

Доказательство [74]. Запишем в функционале скользящего контроля частоту ошибок через сумму индикаторов ошибки и переставим знаки суммирования:

$$Q_c^{lk} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{k} \sum_{x \in X_n^k} I(x, \mu(X_n^l)) = \frac{1}{k} \sum_{i=1}^{L} \frac{1}{N} \underbrace{\sum_{n=1}^{N} [x_i \in X_n^k] I(x_i, \mu(X_n^l))}_{N_i}. \tag{13}$$

Внутренняя сумма, обозначенная через N_i, выражает число разбиений выборки X^L, при которых объект x_i оказывается в контрольной подвыборке и алгоритм $\mu(X_n^l)$ допускает на нем ошибку. Данная ситуация реализуется для таких разбиений, при которых m первых объектов из последовательности $x_{i0}, x_{i1}, \ldots, x_{i,L-1}$ попадают в контрольную подвыборку, m-й сосед x_{im} находится в обучающей подвыборке и принадлежит другому классу, т. е. $r_m(x_i) = 1$. Число таких разбиений в точности равно

$$N_i = \sum_{m=1}^{k} r_m(x_i) C_{L-1-m}^{l-1},$$

поскольку C_{L-1-m}^{l-1} есть число способов выбрать $(l-1)$ обучающих объектов из оставшихся $(L-1-m)$. Подставляя N_i в (13), и используя определение профиля компактности, получаем требуемое выражение функционала Q_c^{lk}.

Теорема доказана.

Комбинаторный множитель $C_{L-1-m}^{l-1} / C_{L-1}^{l}$ убывает с ростом m. Чтобы обеспечить малое значение функционала Q_c^{lk}, достаточно потребовать, чтобы функция $K(m, X^L)$ принимала малые значения при малых m, т. е. чтобы близкие объекты лежали преимущественно в одном классе. При больших m ее рост компенсируется комбинаторным множителем, поэтому далекие объекты могут располагаться как угодно. Таким образом, форма профиля компактности может рассматриваться как разновидность априорной информации о компактности классов.

Заметим, что емкость семейства алгоритмов, индуцируемого методом ближайшего соседа, бесконечна, поэтому классическая теория Вапника–Червоненкиса вообще не дает оценок качества для данного случая.

§ 13. Априорные ограничения монотонности

Рассмотрим еще один случай не-сложностных оценок качества — когда имеется априорная информация о монотонности или почти-монотонности восстанавливаемой зависимости.

Данный тип ограничений часто возникает в прикладных задачах как результат формализации экспертных знаний вида «чем больше значение признака $f(x)$ на объекте x, тем больше значение отклика $y(x)$» или, наоборот, «чем меньше $f(x)$, тем больше $y(x)$». Например, в медицинских задачах возможны суждения типа «чем тяжелее состояние пациента, тем интенсивнее должна быть терапия». В задаче кредитного скоринга возможны суждения типа «чем больше заработная плата клиента, тем более надежным заемщиком он является». Поскольку такие суждения часто носят не обязательный, а рекомендательный характер, целесообразно рассматривать не только строго монотонные, но и почти-монотонные зависимости.

Некоторые методы построения монотонных алгоритмов по конечным выборкам рассматриваются в [6, 24] для задач классификации и восстановления регрессии.

Рассмотрим задачу классификации, в которой множество X частично упорядочено, $Y=\{0, 1\}$, индикатор ошибки имеет вид $I(x, y)=|y^*(x)-y|$, метод обучения μ выбирает алгоритмы из множества A всех монотонных отображений из X в Y.

Определение 8. *Степенью немонотонности* выборки X^L называется наименьшая частота ошибок, допускаемых на ней монотонными алгоритмами:

$$\delta(X^L)=\min_{a\in A}\nu(a, X^L).$$

Выборка X^L называется монотонной, если из $x_i \leqslant x_j$ следует $y_i \leqslant y_j$ для всех $i, j=1, \ldots, L$. Выборка монотонна тогда и только тогда, когда $\delta(X^L)=0$. Если метод μ минимизирует эмпирический риск, т. е. строит алгоритмы с минимальной частотой ошибок на обучающей выборке в классе всех монотонных функций A, то метод μ будет корректным на любой монотонной выборке [6].

Определение 9. *Верхним и нижним клином* объекта $x_i \in X^L$ называются соответственно множества

$$W_0(x_i)=\{x_k \in X^L \mid x_i < x_k \text{ и } y_k=0\},$$
$$W_1(x_i)=\{x_k \in X^L \mid x_k < x_i \text{ и } y_k=1\}.$$

Введем сокращенное обозначение $W_i=W_{y_i}(x_i)$. Мощность клина $w_i=|W_i|$ характеризует глубину погружения объекта x_i в тот класс, которому он принадлежит. Чем меньше w_i, тем ближе объект к границе класса. Для граничных объектов $w_i=0$.

Определение 10. *Профилем монотонности* выборки X^L называется функция $M(m, X^L)$, выражающая долю объектов выборки с клином мощности m:

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

Теорема 4. *Если метод μ минимизирует эмпирический риск в классе всех монотонных функций, и степень немонотонности выборки X^L равна δ, то*

$$Q_c^{lk}(\mu, X^L) \leqslant \sum_{m=0}^{\delta L+k-1} M(m, X^L) \sum_{s=\max\{0,\, w_i-k+1\}}^{\min\{\delta L,\, l,\, m\}} \frac{C_m^s C_{L-1-m}^{l-s}}{C_{L-1}^l}. \qquad (14)$$

Доказательство. Запишем в функционале скользящего контроля частоту ошибок через сумму индикаторов ошибки и поменяем местами знаки суммирования:

$$Q_c^{lk} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{k}\sum_{x\in X_n^k} I(x, \mu(X_n^l)) = \frac{1}{k}\sum_{i=1}^{L}\frac{1}{N}\underbrace{\sum_{n=1}^{N}[x_i \in X_n^k] I(x_i, \mu(X_n^l))}_{N_i}. \quad (15)$$

Внутренняя сумма, обозначенная через N_i, выражает число разбиений выборки X^L, при которых объект x_i оказывается в контрольной подвыборке, и построенный по обучающей подвыборке алгоритм допускает на нем ошибку.

Оценим N_i, воспользовавшись следующим свойством клиньев, вытекающим непосредственно из определения. Если алгоритм a монотонный и допускает ошибку на объекте x_i, то он допускает ошибку и на всех объектах из клина W_i.

В зависимости от соотношения мощности клина w_i и степени немонотонности выборки возможны два случая.

Если $w_i \geqslant \delta L + k$, то ни при каком разбиении монотонный алгоритм не будет ошибаться на x_i, поскольку $\delta L + k$ есть максимальное число ошибок, которое может допустить монотонная функция на всей выборке X^L. Это вытекает из допущения, что метод μ строит алгоритм с минимимальным числом ошибок на обучающей выборке в классе всех монотонных функций. Минимальное число ошибок на любой подвыборке X_n^l не превосходит минимального числа ошибок на всей выборке X^L. Следовательно число ошибок на обучении не превышает δL. Таким образом, в этом случае $N_i = 0$.

Рассмотрим второй случай, когда $w_i < \delta L + k$. Пусть s — число объектов из W_i, находящихся в обучающей подвыборке,

$$\max\{0, w_i - k + 1\} \leqslant s \leqslant \min\{\delta L, l, w_i\}.$$

Имеется $C_{w_i}^s$ способов выбрать s обучающих объектов из клина W_i. Для каждого из этих способов имеется $C_{L-1-w_i}^{l-s}$ вариантов выбрать $l - s$ обучающих объектов из множества $X^L \setminus (W_i \cup \{x_i\})$. В итоге получаем оценку числа разбиений:

$$N_i \leqslant \sum_{s=\max\{0, w_i - k + 1\}}^{\min\{\delta L, l, w_i\}} C_{w_i}^s C_{L-1-w_i}^{l-s}. \quad (16)$$

Представим N в виде $N = C_L^l = \frac{L}{k} C_{L-1}^l$ и подставим оценку (16) в (15), учитывая, что $N_i = 0$ при $w_i \geqslant \delta L + k$:

$$Q_c^{lk} \leqslant \frac{1}{k}\sum_{\substack{i=1 \\ w_i < \delta L + k}}^{L}\frac{k}{L}\sum_{s=\max\{0, w_i - k + 1\}}^{\min\{\delta L, l, w_i\}}\frac{C_{w_i}^s C_{L-1-w_i}^{l-s}}{C_{L-1}^l}.$$

Применяя определение профиля монотонности, получаем требуемое (14).

Теорема доказана.

Следствие 3. *Оценка* (14) *монотонно не убывает по* δ, *достигая наименьшего значения при* $\delta = 0$, *когда выборка монотонна и метод* μ *является корректным*:

$$Q_c^{lk}(\mu, X^L) \leqslant \sum_{m=0}^{k-1} M(m, X^L)\frac{C_{L-1-m}^l}{C_{L-1}^l}. \quad (17)$$

Оценка (14), в отличие от сложностных оценок, всегда не превышает 1. Наибольшее значение 1 достигается, если $w_i = 0$ для всех $i = 1, \ldots, L$. Это тот случай, когда оба класса состоят из попарно несравнимых объектов, и вся выборка распадается на две антицепи. Наименьшее значение достигается, когда выборка монотонна и линейно упорядочена. В этом случае число клиньев мощности w не превышает 2 для всех $w = 1, \ldots, k$, откуда вытекает $Q_c^{lk} \leqslant 2/l$.

Комбинаторный множитель в (14) убывает с ростом m. Чтобы обеспечить малое значение функционала Q_c, достаточно потребовать, чтобы функция $M(m, X^L)$ принимала малые значения при малых m. При больших m ее рост компенсируется комбинаторным множителем. Таким образом, качество монотонного классификатора тем выше, чем меньше объектов имеют клинья небольшой мощности. Для этого отношение порядка на множестве объектов X должно быть близко к линейному вблизи границы классов. Форма профиля монотонности может рассматриваться как формальное выражение априорной информации о плотности отношения порядка [26, 27] вблизи границы классов.

Емкость класса монотонных классификаторов бесконечна, поскольку на выборке длины L, состоящей из попарно несравнимых элементов, реализуется ровно 2^L дихотомий. Таким образом, классическая теория Вапника–Червоненкиса вообще не дает оценок качества для данного случая. Известно [82], что эффективная емкость класса монотонных функций не превосходит длины максимальной антицепи в выборке X^L. Оценка (14) существенно более точная, особенно при малых выборках.

Интересно отметить большое структурное сходство оценок (12) и (17), полученных для таких различных, на первый взгляд, априорных ограничений, как компактность и монотонность.

§ 14. Открытые проблемы

Получение оценок качества лишено смысла, если они не способствуют совершенствованию применяемых на практике алгоритмов.

Ожидается, что оценка (12) позволит обосновать критерии оптимизации метрик для методов ближайших соседей, потенциальных функций и других, основанных на анализе сходства объектов. В частности, представляется возможной разработка методов комбинирования некорректных эвристических метрик путем явной оптимизации профиля компактности.

Ожидается, что оценка (14) позволит обосновать проблемно-ориентированные методы монотонной коррекции [6]. Открытой проблемой остается теоретическое и/или эмпирическое сравнение монотонной и выпуклой коррекции (в частности, бустинга) с точки зрения их обобщающей способности.

Ограничения компактности и монотонности — далеко не единственные «разумные» виды априорной информации. Остается открытым вопрос о получении не-сложностных оценок при априорных ограничениях иного вида.

Пока не ясно, существуют ли нетривиальные оценки локальной функции роста для конкретных методов обучения. Различные методы, работающие в одном и том же семействе, могут порождать различные локальные подсемейства на различных классах задач. Отсюда вытекает целесообразность введения и исследования нового понятия — локализирующей способности метода обучения, как важной составляющей его обобщающей способности.

Возможен пересмотр не только теории равномерной сходимости, но и других техник, используемых для анализа качества обучения, с позиций комбинаторного частотного подхода. Теоретико-вероятностные предположения представляются избыточными не только в рассмотренном случае,

но также при анализе стабильности, отступа и структуры алгоритмических композиций.

Важной проблемой представляется исследование границ применимости скользящего контроля на стадии построения алгоритма, в частности, при выборе наилучшего метода обучения из заданного конечного набора методов.

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.

СПИСОК ЛИТЕРАТУРЫ

1. Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970.
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 // Тез. докл. 7-ой Всеросс. конф. «Математические методы распознавания образов». — Пущино, 1995. — С. 24–26.
5. Воронцов К. В. О проблемно-ориентированной оптимизации базисов задач распознавания // Журн. выч. матем. и матем. физики. — 1998. — V. 38, № 5. — P. 870–880.
6. Воронцов К. В. Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания // Журн. выч. матем. и матем. физики. — 2000. — V. 40, № 1. — P. 166–176.
7. Воронцов К. В. Комбинаторные оценки качества обучения по прецедентам // Докл. РАН. — 2004. — V. 394, № 2. — P. 175–178.
http://www.ccas.ru/frc/papers/voron04qualdan.pdf
8. Дюличева Ю. Ю. Оценка VCD r-редуцированного эмпирического леса // Таврический вестник информатики и математики. — 2003. — № 1. — P. 31–42.
9. Журавлёв Ю. И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть I // Кибернетика. — 1977. — № 4. — P. 5–17.
10. Журавлёв Ю. И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть II // Кибернетика. — 1977. — № 6. — P. 21–27.
11. Журавлёв Ю. И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть III // Кибернетика. — 1978. — № 2. — P. 35–43.
12. Журавлёв Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. Вып. 33. — М.: Наука, 1978. — С. 5–68.
13. Журавлёв Ю. И., Рудаков К. В. Об алгебраической коррекции процедур обработки (преобразования) информации // Проблемы прикл. математики и информатики. — 1987. — P. 187–198.
http://www.ccas.ru/frc/papers/zhurrud87correct.pdf
14. Ивахненко А. Г., Юрачковский Ю. П. Моделирование сложных систем по экспериментальным данным. — М.: Радио и связь, 1987.
15. Матросов В. Л. Корректные алгебры ограниченной емкости над множествами некорректных алгоритмов // Докл. АН СССР. — 1980. — V. 253, № 1. — P. 25–30.
16. Матросов В. Л. Емкость алгебраических расширений модели алгоритмов вычисления оценок // Журн. выч. матем. и матем. физики. — 1984. — V. 24, № 11. — P. 1719–1730.
17. Матросов В. Л. Емкость алгоритмических многочленов над множеством алгоритмов вычисления оценок // Журн. выч. матем. и матем. физики. — 1985. — V. 25, № 1. — P. 122–133.
18. Растригин Л. А., Эренштейн Р. Х. Коллективные правила распознавания. — М.: Энергия, 1981.
19. Рудаков К. В. О симметрических и функциональных ограничениях для алгоритмов классификации // Докл. АН СССР. — 1987. — V. 297, № 1. — P. 43–46.
http://www.ccas.ru/frc/papers/rudakov87dan.pdf
20. Рудаков К. В. Универсальные и локальные ограничения в проблеме коррекции эвристических алгоритмов // Кибернетика. — 1987. — № 2. — P. 30–35.
http://www.ccas.ru/frc/papers/rudakov87universal.pdf
21. Рудаков К. В. Полнота и универсальные ограничения в проблеме коррекции эвристических алгоритмов классификации // Кибернетика. — 1987. — № 3. — P. 106–109.
22. Рудаков К. В. Симметрические и функциональные ограничения в проблеме коррекции эвристических алгоритмов классификации // Кибернетика. — 1987. — № 4. — P. 73–77.
http://www.ccas.ru/frc/papers/rudakov87symmetr.pdf

23. Рудаков К. В. О применении универсальных ограничений при исследовании алгоритмов классификации // Кибернетика. — 1988. — № 1. — P. 1–5.
http://www.ccas.ru/frc/papers/rudakov88universal.pdf
24. Рудаков К. В., Воронцов К. В. О методах оптимизации и монотонной коррекции в алгебраическом подходе к проблеме распознавания // Докл. РАН. — 1999. — V. 367, № 3. — P. 314–317.
25. Рязанов В. В., Сенько О. В. О некоторых моделях голосования и методах их оптимизации // Распознавание, классификация, прогноз. — 1990. — V. 3. — P. 106–145.
26. Сёмочкин А. Н. Линейные достроения частичного порядка на конечных множествах // Деп. ВИНИТИ РАН, № 2964–В98. — М., 1998.
27. Сёмочкин А. Н. Оценки функционала качества для класса алгоритмов с универсальными ограничениями монотонности // Деп. ВИНИТИ РАН, № 2965–В98. — М., 1998.
28. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. — М.: Финансы и статистика, 1988.
29. Anthony M. Uniform Glivenko-Cantelli theorems and concentration of measure in the mathematical modelling of learning // Techn. Rep. LSE-CDAM-2002-07. — 2002.
http://www.maths.lse.ac.uk/Personal/martin/mresearch.html
30. Anthony M., Bartlett P. L. Neural network learning: theoretical foundations. — Cambridge Univ. Press, 1999.
31. Anthony M., Shawe-Taylor J. A result of Vapnik with applications // Discrete Appl. Math. — 1993. — V. 47, № 2. — P. 207–217.
http://citeseer.ist.psu.edu/anthony91result.html
32. Antos A., Kegl B., Linder T., Lugosi G. Data-dependent margin-based generalization bounds for classification // J. of Machine Learning Research. — 2002. — P. 73–98.
http://citeseer.ist.psu.edu/article/antos02datadependent.html
33. Bartlett P. Lower bounds on the Vapnik–Chervonenkis dimension of multi-layer threshold networks // Proc. of the Sixth Annual ACM Conference on Computational Learning Theory. — New York: ACM Press, 1993. — P. 144–150.
http://citeseer.ist.psu.edu/bartlett93lower.html
34. Bartlett P. L. For valid generalization the size of the weights is more important than the size of the network // Adv. in Neural Information Processing Systems. — The MIT Press, 1997. — V. 9. — P. 134.
http://citeseer.ist.psu.edu/bartlett97for.html
35. Bartlett P. L. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network // IEEE Trans. Inform. Theory. — 1998. — V. 44, № 2. — P. 525–536.
http://discus.anu.edu.au/~bartlett
36. Bartlett P. L., Long P. M., Williamson R. C. Fat-shattering and the learnability of real-valued functions // J. Computer and System Sci. — 1996. — V. 52, № 3. — P. 434–452.
http://citeseer.ist.psu.edu/bartlett95fatshattering.html
37. Bontempi G., Birattari M. A bound on the cross-validation estimate for algorithm assessment // Eleventh Belgium/Netherlands Conference on Artificial Intelligence (BNAIC). — 1999. — P. 115–122.
http://citeseer.ist.psu.edu/225930.html
38. Bottou L., Cortes C., Vapnik V. On the effective VC dimension // 1994.
http://citeseer.ist.psu.edu/bottou94effective.html
39. Boucheron S., Lugosi G., Massart P. A sharp concentration inequality with applications // Random Structures and Algorithms. — 2000. — V. 16, № 3. — P. 277–292.
http://citeseer.ist.psu.edu/article/boucheron99sharp.html
40. Boucheron S., Lugosi G., Massart P. Concentration inequalities using the entropy method // Annals of Probability. — 2003. — V. 31, № 3.
http://citeseer.ist.psu.edu/boucheron02concentration.html
41. Bousquet O., Elisseeff A. Algorithmic Stability and Generalization Performance // Adv. in Neural Information Processing Systems. — The MIT Press, 2001. — V. 13. — P. 196–202.
http://citeseer.ist.psu.edu/article/bousquet00algorithmic.html
42. Bousquet O., Elisseeff A. Stability and generalization // J. of Machine Learning Research. — 2002. — № 2. — P. 499–526.
http://citeseer.ist.psu.edu/article/bousquet00stability.html
43. Breiman L. Bagging predictors // Machine Learning. — 1996. — V. 24, № 2. — P. 123–140.
http://citeseer.ist.psu.edu/breiman96bagging.html
44. Breiman L. Bias, variance, and arcing classifiers // Statistics Dept., Univ. of California / Techn. Rep. № 460. — 1996.
http://citeseer.ist.psu.edu/breiman96bias.html
45. Breiman L. Arcing classifiers // Annals of Statistics. — 1998. — V. 26, № 3. — P. 801–849.
http://citeseer.ist.psu.edu/breiman98arcing.html

46. Burges C. J. C. A tutorial on support vector machines for pattern recognition // Data Mining and Knowledge Discovery. — 1998. — V. 2, № 2. — P. 121–167.
http://citeseer.ist.psu.edu/burges98tutorial.html
47. Chernoff H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations // Annals of Math. Stat. — 1952. — V. 23. — P. 493–509.
48. Cortes C., Vapnik V. Support-vector networks // Machine Learning. — 1995. — V. 20, № 3. — P. 237–297.
http://citeseer.ist.psu.edu/cortes95supportvector.html
49. Devroye L. P., Wagner T. J. Distribution-free inequalities for the deleted and holdout error estimates // IEEE Trans. Inform. Theory. — 1979. — V. 25, № 2. — P. 202–207.
50. Devroye L. P., Wagner T. J. Distribution-free performance bounds for potential function rules // IEEE Trans. Inform. Theory. — 1979. — V. 25, № 5. — P. 601–604.
51. Efron B. The jackknife, the bootstrap, and other resampling plans. — SIAM, Philadelphia, 1982.
52. Evgeniou T., Pontil M., Elisseeff A. Leave one out error, stability, and generalization of voting combinations of classifiers // Techn. Rep. INSEAD 2001-21-TM. — 2001.
http://citeseer.ist.psu.edu/445768.html
53. Freund Y. Boosting a weak learning algorithm by majority // Proc. of the Workshop on Computational Learning Theory. — Morgan Kaufmann Publ., 1990.
http://citeseer.ist.psu.edu/freund95boosting.html
54. Freund Y. Self Bounding Learning Algorithms // Proc. of the Workshop on Computational Learning Theory. — Morgan Kaufmann Publ., 1998.
http://citeseer.ist.psu.edu/freund98self.html
55. Freund Y., Schapire R. E. A decision-theoretic generalization of on-line learning and an application to boosting // European Conf. on Computational Learning Theory. — 1995. — P. 23–37.
http://citeseer.ist.psu.edu/article/freund95decisiontheoretic.html
56. Freund Y., Schapire R. E. Experiments with a new boosting algorithm // Int. Conf. on Machine Learning. — 1996. — P. 148–156.
http://citeseer.ist.psu.edu/freund96experiments.html
57. Freund Y., Schapire R. E. Discussion of the paper «Arcing classifiers» by Leo Breiman // Annals of Statistics. — 1998. — V. 26, № 3. — P. 824–832.
http://citeseer.ist.psu.edu/freund97discussion.html
58. Golea M., Bartlett P., Lee W. S., Mason L. Generalization in decision trees and DNF: Does size matter? // Adv. in Neural Information Processing Systems. — The MIT Press, 1998. — V. 10.
http://citeseer.ist.psu.edu/golea97generalization.html
59. Holden S. B. Cross-validation and the PAC learning model // Dept. of CS, Univ. College, London / Techn. Rep. RN/96/64. — 1996.
60. Karpinski M., Macintyre A. Polynomial bounds for VC dimension of sigmoidal neural networks // 27th ACM Symp. Theory Comput. — 1995. — P. 200–208.
http://citeseer.ist.psu.edu/karpinski95polynomial.html
61. Kearns M. A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split // Adv. in Neural Information Processing Systems. — The MIT Press, 1996. — V. 8. — P. 183–189.
http://citeseer.ist.psu.edu/kearns96bound.html
62. Kearns M. J., Mansour Y., Ng A. Y., Ron D. An experimental and theoretical comparison of model selection methods // Computational Learning Theory. — 1995. — P. 21–30.
http://citeseer.ist.psu.edu/kearns95experimental.html
63. Kearns M. J., Ron D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation // Computational Learning Theory. — 1997. — P. 152–162.
http://citeseer.ist.psu.edu/kearns97algorithmic.html
64. Kearns M. J., Schapire R. E. Efficient distribution-free learning of probabilistic concepts // Computational Learning Theory and Natural Learning Systems, Vol. I: Constraints and Prospect. — Bradford/MIT Press, 1994.
http://citeseer.ist.psu.edu/article/kearns93efficient.html
65. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection // IJCAI. — 1995. — P. 1137–1145.
http://citeseer.ist.psu.edu/kohavi95study.html
66. Kutin S., Niyogi P. Almost-everywhere algorithmic stability and generalization error // Univ. of Chicago / Techn. Rep. TR-2002-03. — 2002.
http://citeseer.ist.psu.edu/kutin02almosteverywhere.html
67. Langford J., Blum A. Microchoice bounds and self bounding learning algorithms // Computational Learning Theory. — 1999. — P. 209–214.
http://citeseer.ist.psu.edu/langford01microchoice.html

68. Lugosi G. On concentration-of-measure inequalities // Machine Learning Summer School, Australian National Univ., Canberra — 2003.
http://citeseer.ist.psu.edu/lugosi98concentrationmeasure.html
69. Mason L., Bartlett P., Baxter J. Direct optimization of margins improves generalization in combined classifiers // Dept. of Systems Engineering, Australian National Univ. / Techn. Rep. — 1998.
http://citeseer.ist.psu.edu/mason98direct.html
70. Mason L., Bartlett P., Golea M. Generalization error of combined classifiers // Dept. of Systems Engineering, Australian National Univ. / Techn. Rep. — 1997.
http://citeseer.ist.psu.edu/mason97generalization.html
71. Mazurov V. D., Khachai M. Yu., Rybin, A. I. Committee constructions for solving problems of selection, diagnostics and prediction // Proc. Steklov Inst. Math. — 2002. — V. 1. — P. 67–101.
http://tom.imm.uran.ru/khachay/publications/mine/psis67.pdf
72. McDiarmid C. On the method of bounded differences // In Surveys in Combinatorics, London Math. Soc. Lect. Notes Ser. — 1989. — V. 141. — P. 148–188.
73. Mertens S., Engel A. Vapnik–Chervonenkis dimension of neural networks with binary weights // Phys. Rev. E. — 1997. — V. 55, № 4. — P. 4478–4488.
74. Mullin M., Sukthankar R. Complete cross-validation for nearest neighbor classifiers // Proc. of Int. Conf. on Machine Learning. — 2000.
http://citeseer.ist.psu.edu/309025.html
75. Ng A. Y. Preventing overfitting of cross-validation data // Proc. 14th Int. Conf. on Machine Learning. — Morgan Kaufmann, 1997. — P. 245–253.
http://citeseer.ist.psu.edu/ng97preventing.html
76. Quinlan J. R. Induction of decision trees // Machine Learning. — 1986. — V. 1, № 1. — P. 81–106.
77. Rissanen J. Modeling by shortest data description // Automatica. — 1978. — V. 14. — P. 465–471.
78. Rogers W., Wagner T. A finite sample distribution-free performance bound for local discrimination rules // Annals of Statistics. — 1978. — V. 6, № 3. — P. 506–514.
79. Schapire R. The boosting approach to machine learning: An overview // MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA. — 2001.
http://citeseer.ist.psu.edu/schapire02boosting.html
80. 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. — V. 26, № 5. — P. 1651–1686.
http://citeseer.ist.psu.edu/article/schapire98boosting.html
81. Shawe-Taylor J., Bartlett P. L. Structural risk minimization over data-dependent hierarchies // IEEE Trans. Inform. Theory. — 1998. — V. 44, № 5. — P. 1926–1940.
http://citeseer.ist.psu.edu/article/shawe-taylor98structural.html
82. Sill J. The capacity of monotonic functions // Discrete Appl. Math. — 1998. — V. 86. — P. 95–107.
http://citeseer.ist.psu.edu/49191.html
83. Skurichina M., Kuncheva L. I., Duin R. P. W. Bagging and boosting for the nearest mean classifier: Effects of sample size on diversity and accuracy // Multiple classifier systems (Proc. Third Int. Workshop MCS, Cagliari, Italy). — V. 2364. — Berlin: Springer, 1994. — P. 62–71.
http://citeseer.ist.psu.edu/539135.html
84. Smola A., Bartlett P., Scholkopf B., Schuurmans D. Advances in large margin classifiers // MIT Press, Cambridge, MA. — 2000.
http://citeseer.ist.psu.edu/article/smola00advances.html
85. Talagrand M. Sharper bounds for gaussian and empirical processes // Annals of Probability. — 1994. — № 22. — P. 28–76.
86. Talagrand M. Concentration of measure and isoperimetric inequalities in product space // Publ. Math. I.H.E.S. — 1995. — № 81. — P. 73–205.
http://citeseer.ist.psu.edu/talagrand95concentration.html
87. Vapnik V. Statistical learning theory. — New York: Wiley, 1998.
88. Vapnik V., Levin E., Cun Y. L. Measuring the VC-dimension of a learning machine", Neural Computation. — 1994. — V. 6, № 5. — P. 851–876.
http://citeseer.ist.psu.edu/vapnik94measuring.html
89. Vayatis N., Azencott R. Distribution-dependent Vapnik–Chervonenkis bounds // Lect. Notes Comp. Sci. — 1999. — V. 1572. — P. 230–240.
http://citeseer.ist.psu.edu/vayatis99distributiondependent.html
90. Williamson R., Shawe-Taylor J., Scholkopf B., Smola A. Sample based generalization bounds // NeuroCOLT Techn. Rep. NC-TR-99-055. — 1999.
http://citeseer.ist.psu.edu/williamson99sample.html

Поступило в редакцию 13 IV 2004