УДК 519.7:004.855.5

Комбинаторные обоснования обучаемых алгоритмов[1]

© 2004 г. К. В. Воронцов

(*119991, Москва, ул. Вавилова, 40, ВЦ РАН*)
e-mail: voron@ccas.ru

Поступила в редакцию
30 января 2004 г.

Рассматриваются комбинаторные функционалы качества обучения по прецедентам, основанные на принципе скользящего контроля. Выводятся их верхние оценки, более точные, чем оценки статистической теории Вапника-Червоненкиса, и при этом не предполагающие случайности и независимости исходных данных. Описывается эффект локализации семейства алгоритмов и вводится понятие локальной функции роста. С позиций комбинаторного подхода пересматриваются основные положения статистической теории. Анализируются основные причины завышенности сложностных оценок качества. Библ. 24. Табл. 2.

Ключевые слова: обучение по прецедентам, теория Вапника-Червоненкиса, локальная функция роста, локальная эффективная емкость, степень некорректности.

В теории обучаемых систем качество обучения или обобщающую способность алгоритмов принято характеризовать вероятностью ошибки. К сожалению, это гипотетическая величина, которую невозможно вычислить, а иногда даже и адекватно оценить, например, в случае малых выборок. В то же время на практике любая обучаемая система сталкивается только с конечными выборками, будь то обучающие или тестовые совокупности объектов. Поэтому обобщающую способность алгоритмов целесообразно характеризовать именно относительно конечных выборок. Для эмпирического измерения качества обучения принято использовать независимые контрольные выборки, бутстреп или скользящий контроль [1]. В данной работе показано, что верхние оценки функционалов качества, основанных на принципе скользящего контроля, можно получать, вообще не прибегая к аппарату теории вероятностей. При этом они оказываются даже более точными, чем традиционные вероятностные.

В разд. 1 вводятся основные понятия, связанные с задачей обучения по прецедентам. В разд. 2 кратко перечисляются основные положения статистической теории Вапника-Червоненкиса [2], необходимые для дальнейшего изложения. В разд. 3 определяются комбинаторные функционалы, характеризующие качество заданного метода обучения на заданной конечной выборке.

---

[1]Работа выполнена в рамках программы Отделения математических наук РАН «Алгебраические и комбинаторные методы математической кибернетики», при финансовой поддержке РФФИ (коды проектов 02-01-00325, 01-07-90242) и Фонда содействия отечественной науке.

В разд. 4 выводятся верхние оценки комбинаторных функционалов, аналогичные оценкам вероятности равномерного отклонения частоты ошибок в двух подвыборках. Сходство этих оценок позволяет утверждать, что функционалы скользящего контроля характеризуют обобщающую способность по меньшей мере не хуже вероятностных. Но имеются и концептуальные отличия. Во-первых, комбинаторные оценки справедливы для любой выборки, не обязательно случайной, независимой и одинаково распределенной. Во-вторых, они зависят от сложности не всего семейства алгоритмов, а лишь его локального подмножества, состоящего из алгоритмов, реально получаемых в результате обучения. Таким образом, принцип равномерной сходимости оказывается избыточным, а семейство алгоритмов становится вторичной конструкцией по отношению к методу обучения.

Связь комбинаторных и вероятностных функционалов качества рассматривается в разд. 5. Верхние оценки комбинаторных функционалов легко переносятся на соответствующие вероятностные функционалы. Это позволяет говорить о соблюдении «принципа соответствия» при переходе от статистической теории к более точной комбинаторной.

В разд. 6 основные положения статистической теории пересматриваются с позиций комбинаторного подхода, в том числе: свойство корректности обучаемых алгоритмов, функционал равномерного относительного отклонения частот, метод структурной минимизации риска и понятие эффективной емкости. В разд. 7 анализируются основные причины завышенности сложностных оценок качества.

Данная работа содержит доказательства теорем, опущенные в кратком сообщении [3].

1. 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 $y^* : X \to Y$, 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: $a^*(x_i) = y_i$, $i = 1, 2, \ldots, l$. Равенство здесь может пониматься как точное или приближенное в зависимости от конкретной задачи. Требования такого вида называют *локальными ограничениями* [4], 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* [4], 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 конструк-

ции 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 при $Y = \{0, 1\}$ обычно полагают

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

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

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

где $\delta(x)$ — фиксированная функция. Квадратные скобки здесь и далее обозначают «естественное» отображение логической величины в число: [ложь] = 0, [истина] = 1.

Использование бинарного индикатора ошибки позволяет единообразно исследовать широкий класс задач, включающий как классификацию, так и восстановление регрессии.

2. О статистической теории восстановления зависимостей

В статистической теории Вапника-Червоненкиса [2], [5] предполагается, что множество объектов X является вероятностным пространством, и все рассматриваемые выборки являются случайными, независимыми, одинаково распределенными. Процесс обучения состоит в построении алгоритма a^* из заданного семейства алгоритмов $A \subset \mathfrak{A}$ путем *минимизации эмпирического риска*

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

В семействе может существовать много алгоритмов, минимизирующих эмпирический риск. Однако способ получения конкретного алгоритма никак не фиксируется, и предполагается, что может быть выбран любой из них.

Качество алгоритма a^* характеризуется либо вероятностью ошибки, либо частотой ошибок $\nu(a^*, X^k)$ на неизвестной контрольной выборке X^k. Вычислить эту

величину не представляется возможным, поэтому ставится задача определить условия, при которых она не сильно отличается от эмпирического риска $\nu(a^*, X^l)$. Достаточным условием является малое значение функционала равномерного отклонения частоты ошибок в двух выборках:

$$P_\varepsilon^{lk}(A) = \mathbf{P}\Big\{\sup_{a\in A}\big(\nu(a, X^k) - \nu(a, X^l)\big) > \varepsilon\Big\}. \tag{2.1}$$

В общем случае заранее неизвестно, какой именно алгоритм будет получен в результате обучения. Поэтому оценивается максимальное отклонение, достигаемое на наихудшем алгоритме. Если $P_\varepsilon^{lk}(A) \to 0$ при $l \to \infty$, то говорят, что имеет место равномерная сходимость частот ошибок в двух выборках. Это и есть достаточное условие *обучаемости* семейства алгоритмов.

При $l = k$ и любом распределении вероятностей на множестве объектов справедлива оценка [5]

$$P_\varepsilon^{lk}(A) \leqslant \Delta^A(2l) 1.5\, e^{-\varepsilon^2 l}, \tag{2.2}$$

где $\Delta^A(2l)$ — функция роста семейства алгоритмов A.

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

Очевидно, $\Delta^A(L)$ не превосходит 2^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:

$$\Delta^A(L) \leqslant C_L^0 + \cdots + C_L^h \leqslant 1.5\frac{L^h}{h!}. \tag{2.3}$$

В этом случае имеет место равномерная сходимость частот. Таким образом, для получения оценок качества обучения в статистической теории достаточно знать только длину выборки и емкость семейства алгоритмов.

Вычисление или оценивание емкости является отдельной, зачастую достаточно сложной, задачей для каждого конкретного семейства. Доказано, что емкость семейства линейных решающих правил равна числу свободных параметров или размерности линейного пространства, в котором строится разделяющая гиперплоскость. Оценки емкости получены также для нейронных сетей [6], решающих деревьев и решающего леса [7], корректных алгебраических замыканий подмодели алгоритмов вычисления оценок (АВО) [8], комитетов линейных неравенств [9] и других моделей алгоритмов.

Статистические оценки позволяют обосновать метод *структурной минимизации риска*, направленный на выбор подмодели алгоритмов оптимальной сложности. В этом методе фиксируется структура вложенных подсемейств возрастающей емкости $A_1 \subset \ldots \subset A_h = A$, и в каждом подсемействе решается задача обучения по прецедентам. Из полученных алгоритмов выбирается тот, для которого верхняя оценка $\nu(a, X^k)$ принимает наименьшее значение.

К сожалению, оценка (2.2) сильно завышена. Рассчитанные по ней значения достаточной длины обучения l существенно превышают количество объектов, с которыми приходится иметь дело на практике. В табл. 1 приведена зависимость достаточной длины обучающей выборки l как функции от емкости h, точности ε и значения функционала качества P_ε^{lk}. Правая половина таблицы, соответствующая значению $P_\varepsilon^{lk} = 1$, показывает границу применимости оценки (2.2). При меньших l оценка превышает единицу, т. е. становится тривиальной. В методе структурной минимизации риска завышенность оценки приводит к чрезмерному упрощению алгоритмов [10].

Табл. 1.

	$P_\varepsilon^{lk} = 0.01$				$P_\varepsilon^{lk} = 1$			
h	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
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

Завышенность статистических оценок является следствием их чрезмерной общности. Они ориентированы на «худший случай» и не учитывают трех важных особенностей самой задачи и процесса поиска ее решения. Во-первых, это особенности распределения объектов в пространстве — они могут лежать в подпространстве меньшей размерности, причем в задачах восстановления зависимостей этот «вырожденный» случай становится типичным из-за наличия зависимых или почти зависимых признаков. Во-вторых, это особенности самой восстанавливаемой зависимости — она может быть гладкой, симметричной, монотонной или обладать другими специальными свойствами, что резко сужает пространство допустимых решений. В-третьих, это особенности метода обучения — он может обладать способностью подстраиваться под задачу, выделяя эффективное подсемейство алгоритмов, реально получаемых в результате обучения.

3. Комбинаторные функционалы качества обучения

Принцип минимизации эмпирического риска в заранее заданном семействе алгоритмов является недостаточно 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, хорошо зарекомендовавшие себя на практике, минимизируют эмпирический риск. К ним относятся методы, использующие технику скользяще-

го контроля или внешних критериев, в частности, метод группового учета аргументов (МГУА) [11], методы явной максимизации отступа [12], бустинг [13] баггинг [14], и др.

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, то есть 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.

Частота ошибок $\nu(\mu(X^l), X^k)$ на заданной контрольной выборке X^k, в общем случае не пересекающейся с X^l, также не вполне адекватно характеризует качество обучения. Недостаток этого функционала 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), $n = 1, 2, \ldots, N$ всевозможные разбиения выборки X^L на обучающую и контрольную подвыборки длиной l и k соответственно. Число разбиений N равно C_L^l.

Следующие функционалы характеризуют качество обучения методом μ на конечном наборе объектов X^L и обладают требуемым свойством инвариантности.

1. Функционал полного скользящего контроля [1] имеет вид

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

2. Функционал среднего отклонения частоты ошибок на контроле от частоты ошибок на обучении есть

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

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

3. Функционал скользящего контроля, терпимый к незначительной доле ошибок ε на контрольной подвыборке имеет вид

$$Q_\varepsilon^{lk}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^{N} \left[\nu(\mu(X_n^l), X_n^k) > \varepsilon\right], \quad 0 \leqslant \varepsilon \leqslant 1.$$

4. Функционал скользящего контроля, терпимый к незначительным отклонениям частоты ошибок на контрольной выборке от частоты ошибок на обучении есть

$$Q^{lk}_{\nu,\varepsilon}(\mu, X^L) = \frac{1}{N}\sum_{n=1}^{N}\left[\nu(\mu(X^l_n), X^k_n) - \nu(\mu(X^l_n), X^l_n) > \varepsilon\right], \quad 0 \leqslant \varepsilon \leqslant 1.$$

Введем также функцию средней частоты ошибок на обучении:

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

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

Непосредственно из определений вытекают следующие соотношения между введенными функционалами: $Q_c \leqslant Q_d + \bar{\nu}^l_L$ и $Q_{\nu,\varepsilon} \leqslant Q_\varepsilon$. Если метод μ корректен на всех подвыборках длины l, то $\bar{\nu}^l_L = 0$ и $Q_c = Q_d$ и $Q_{\nu,\varepsilon} = Q_\varepsilon$. Чуть менее очевидны следующие двусторонние оценки.

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

$$\begin{aligned}
&\varepsilon Q_\varepsilon < Q_c \leqslant \varepsilon + (1-\varepsilon)Q_\varepsilon;\\
&\varepsilon Q_{\nu,\varepsilon} < Q_d \leqslant \varepsilon + (1-\varepsilon)Q_{\nu,\varepsilon};\\
&\varepsilon Q_{\nu,\varepsilon} < Q_c \leqslant \varepsilon + (1-\varepsilon)Q_{\nu,\varepsilon} + \bar{\nu}^l_L.
\end{aligned}$$

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 следует из первых двух и приведенных выше соотношений:

$$\varepsilon Q_{\nu,\varepsilon} \;\leqslant\; \varepsilon Q_\varepsilon \;<\; Q_c \;\leqslant\; Q_d + \bar{\nu}^l_L \;\leqslant\; \varepsilon + (1-\varepsilon)Q_{\nu,\varepsilon} + \bar{\nu}^l_L.$$

Лемма доказана.

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

4. 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 «непохожих» алгоритмов содержится в семействе. Это свойство предлагается называть *локализующей способностью* метода обучения.

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

$$A_L^l(\mu, X^L) = \left\{ \mu(X_n^l) \;\middle|\; n = 1, 2, \ldots, N \right\}, \quad N = C_L^l.$$

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

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 C_L^l, в то время как $\Delta^A(L) \leqslant 2^L$. Локальная функция роста не превосходит $\Delta^A(L)$.

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

$$\sigma_L^l(\mu, X^L) = \max_{n=1,2,\ldots,N} \nu(\mu(X_n^l), X_n^l).$$

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

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*

$$Q_{\nu,\varepsilon}^{lk}(\mu, X^L) < \Delta_L^l(\mu, X^L) \Gamma_L^l(\varepsilon, \sigma), \tag{4.1}$$

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

$$\Gamma_L^l(\varepsilon, \sigma) = \max_{m \in M(\varepsilon,\sigma)} \sum_{s \in S(\varepsilon,\sigma)} \frac{C_m^s C_{L-m}^{l-s}}{C_L^l},$$
$$M(\varepsilon, \sigma) = \left\{ m \mid \varepsilon k < m \leqslant k + \sigma l \right\},$$
$$S(\varepsilon, \sigma) = \left\{ s \mid \max(0, m-k) \leqslant s \leqslant \sigma l,\ s < (m - \varepsilon k) l / L \right\}.$$

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

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

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 через A_{mi}, где $m = 0, 1, \ldots, L$ — число ошибок, допускаемых на выборке X^L алгоритмами данного класса, $i = 1, 2, \ldots, \Delta_m$ — 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:

$$\Delta_L^l = \Delta_0 + \cdots + \Delta_L. \tag{4.2}$$

Эквивалентность на алгоритмах порождает эквивалентность на разбиениях, если для произвольных n и u из $\{1, 2, \ldots, N\}$ положить $n \sim u \Leftrightarrow \mu(X_n^l) \sim \mu(X_u^l)$. При этом образуются классы эквивалентности $N_{mi} = \left\{ n \;\middle|\; \mu(X_n^l) \in A_{mi} \right\}$ на множестве разбиений, взаимно однозначно соответствующие классам A_{mi}.

Запишем функционал качества, просуммировав 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 $m \in M(\varepsilon, \sigma)$:

$$Q_{\nu,\varepsilon}^{lk} = \sum_{m \in M(\varepsilon,\sigma)} \sum_{i=1}^{\Delta_m} \underbrace{\frac{1}{N} \sum_{n \in N_{mi}} [\nu(a_{mi}, X_n^k) > \nu(a_{mi}, X_n^l) + \varepsilon]}_{\gamma_{mi}}. \tag{4.3}$$

Оценим сверху внутреннюю сумму γ_{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:

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

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 ровно 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 (4.3) ее можно вынести за знак суммирования по i. Используя (4.2), 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.

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

$$Q^{lk}_{\nu,\varepsilon}(\mu, X^L) \leqslant \Delta^l_L(\mu, X^L) 1.5\, e^{-\varepsilon^2 l} \leqslant \Delta^A(L) 1.5\, e^{-\varepsilon^2 l}. \tag{4.4}$$

Доказательство непосредственно вытекает из следующих фактов: локальная функция роста $\Delta^l_L(\mu, X^L)$ не превосходит функции роста всего семейства $\Delta^A(L)$; комбинаторный множитель $\Gamma^l_L(\varepsilon, \sigma)$ не убывает по σ; согласно [2] при $l = k$ справедлива оценка $\Gamma^l_L(\varepsilon, 1) \leqslant 1.5\, e^{-\varepsilon^2 l}$.

Усиление оценки достигается, главным образом, благодаря модификации функционала качества, связанной с отказом от избыточного требования равномерной сходимости. Данный результат впервые упоминался в [15].

Отметим, что в общем случае нет оснований приравнивать l и k, за исключением удобства оценивания комбинаторного множителя.

В соответствии с леммой 1 аналогичные оценки справедливы и для других комбинаторных функционалов. Более аккуратная техника позволяет несколько уточнить верхние оценки функционалов Q^{lk}_d и Q^{lk}_c.

Теорема 3. *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*

$$Q^{lk}_d(\mu, X^L) < \varepsilon + \Delta^l_L(\mu, X^L) \tilde{\Gamma}^l_L(\varepsilon, \sigma), \tag{4.5}$$

где

$$\tilde{\Gamma}^l_L(\varepsilon, \sigma) = \max_{m \in M(\varepsilon,\sigma)} \sum_{s \in S(\varepsilon,\sigma)} \left(\frac{ml - sL}{lk} - \varepsilon \right) \frac{C^s_m C^{l-s}_{L-m}}{C^l_L},$$

множества $M(\varepsilon, \sigma)$ и $S(\varepsilon, \sigma)$ определяются так же, как в теореме 2.

Доказательство. Из неравенства $(x)_+ \leqslant \varepsilon + (x - \varepsilon)_+$, справедливого при любых $x \in \mathbb{R}$ и $\varepsilon \geqslant 0$, следует

$$Q^{lk}_d \leqslant \varepsilon + \frac{1}{N} \sum_{n=1}^{N} \left(\nu(\mu(X^l_n), X^k_n) - \nu(\mu(X^l_n), X^l_n) - \varepsilon \right)_+,$$

Аналогично доказательству предыдущей теоремы введем классы эквивалентных алгоритмов A_{mi}, классы эквивалентных разбиений N_{mi}, затем выберем по одному представителю a_{mi} из каждого класса A_{mi}, где m — число ошибок, допускаемых алгоритмом a_{mi} на всей выборке X^L, $m = 0, 1, \ldots, L$, $i = 1, 2, \ldots, \Delta_m$. Просуммируем разбиения отдельно по классам эквивалентности:

$$Q^{lk}_d \leqslant \varepsilon + \frac{1}{N} \sum_{m=0}^{L} \sum_{i=1}^{\Delta_m} \sum_{n \in N_{mi}} \left(\nu(a_{mi}, X^k_n) - \nu(a_{mi}, X^l_n) - \varepsilon \right)_+.$$

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

$$Q_d^{lk} \leqslant \varepsilon + \sum_{m \in M(\varepsilon,\sigma)} \sum_{i=1}^{\Delta_m} \underbrace{\frac{1}{N} \sum_{n \in N_{mi}} \big(\nu(a_{mi}, X_n^k) - \nu(a_{mi}, X_n^l) - \varepsilon\big)_+}_{\tilde{\gamma}_{mi}}. \tag{4.6}$$

Оценим сверху внутреннюю сумму $\tilde{\gamma}_{mi}$, 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:

$$\begin{aligned}\tilde{\gamma}_{mi} &\leqslant \frac{1}{N} \sum_{n=1}^{N} \big(\nu(a_{mi}, X_n^k) - \nu(a_{mi}, X_n^l) - \varepsilon\big)_+ = \\ &= \sum_{s=0}^{\sigma l} \Big[\frac{m-s}{k} > \frac{s}{l} + \varepsilon\Big] \Big(\frac{m-s}{k} - \frac{s}{l} - \varepsilon\Big) \frac{1}{N} \sum_{n=1}^{N} \Big[\nu(a_{mi}, X_n^l) = \frac{s}{l}\Big].\end{aligned}$$

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,

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

Эта величина уже не зависит от i, поэтому в (4.6) ее можно вынести за знак суммирования по i. Используя (4.2), приходим к неравенству:

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

Подставляя сюда оценку $\tilde{\gamma}_{mi}$, 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

$$Q_c^{lk}(\mu, X^L) < \bar{\nu}_L^l + \varepsilon + \Delta_L^l(\mu, X^L) \tilde{\Gamma}_L^l(\varepsilon, \sigma). \tag{4.7}$$

Оценки (4.5) и (4.7) содержат искусственно введенный параметр ε. Чтобы избавиться от него, необходимо решить дополнительную задачу минимизации по ε.

5. Вероятностные функционалы и «принцип соответствия»

Полученные результаты свидетельствуют о возможности построения невероятностной теории качества обучения по прецедентам.

В отличие от функционала 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:

$$\begin{aligned}
\mathbf{E}Q_c^{lk}(\mu, X^L) &= \mathbf{P}\{I(\mu(X^l), x) = 1\}, \\
\mathbf{E}Q_\varepsilon^{lk}(\mu, X^L) &= \mathbf{P}\big\{\nu(\mu(X^l), X^k) > \varepsilon\big\}, \\
\mathbf{E}Q_{\nu,\varepsilon}^{lk}(\mu, X^L) &= \mathbf{P}\big\{\nu(\mu(X^l), X^k) - \nu(\mu(X^l), X^l) > \varepsilon\big\}.
\end{aligned}$$

Первая строка выражает хорошо известный факт, что полный скользящий контроль Q_c дает несмещенную оценку вероятности ошибки [2]. Другие функционалы также являются несмещенными оценками соответствующих вероятностных функционалов, имеющих вполне понятную интерпретацию.

Любая верхняя оценка комбинаторного функционала легко преобразуется в верхнюю оценку соответствующего вероятностного функционала путем применения операции матожидания к обеим частям неравенства.

Введенные только что вероятностные функционалы более точно характеризуют качество обучения, чем функционал Вапника-Червоненкиса P_ε^{lk}, поскольку они не содержат избыточно сильного требования равномерной сходимости. Непосредственно из определений следует, что P_ε^{lk} является завышенной верхней оценкой функционала $\mathbf{E}Q_{\nu,\varepsilon}^{lk}$, который и описывает качество обучения:

$$\mathbf{E}Q_{\nu,\varepsilon}^{lk}(\mu, X^L) \leqslant P_\varepsilon^{lk}(A). \tag{5.1}$$

Таким образом, соблюдается «принцип 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. Более того, использование вероятностных функционалов качества может приводить к лишним промежуточным шагам при выводе оценок и понижению их точности (типичный пример — «основная лемма» в статистической теории [2, с. 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 2 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 (4.1) и (4.5) является не вероятностной, а исключительно комбинаторной, и вытекает из дискретности индикатора ошибки $I(x, y)$ и симметричности функционала качества.

6. Пересмотр некоторых положений статистической теории

1. 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 несколько упрощается:

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

Если h — емкость локального подсемейства, то из (2.3) и теоремы 2 следует оценка качества корректного метода обучения:

$$Q_\varepsilon^{lk} = Q_{\nu,\varepsilon}^{lk} < \left(C_L^0 + \cdots + C_L^h\right) \frac{C_{L-\lceil \varepsilon k \rceil}^l}{C_L^l}.$$

В табл. 2 приводятся результаты численного расчета требуемой длины обучающей выборки l по полученному соотношению. Эти значения существенно лучше приведенных в табл. 1.

Табл. 2.

	$Q_\varepsilon^{lk} = 0.01$				$Q_\varepsilon^{lk} = 1$			
h	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
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 по порядку величины с $\Gamma_L^l(\varepsilon, 1)$, уже при $\sigma \approx \varepsilon$. 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. Отметим, что идея построения корректных алгоритмических композиций является центральной в алгебраическом подходе к распознаванию [16].

2. В статистической теории признается, 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. Утверждается, что описать эту область в явном виде затруднительно, в связи с чем предлагается частичное решение проблемы. Вводится функционал равномерного относительного отклонения частот в двух подвыборках, для которого справедлива оценка [2]:

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

где

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

Нетрудно показать, что данная оценка выводится из комбинаторной формулы (4.1) путем замены переменной $\varepsilon\sqrt{m/L}\to\varepsilon$. Она всего лишь по-другому оценивает комбинаторный множитель, но не описывает эффекта сужения семейства. При непосредственном использовании (4.1) необходимость в относительных оценках отпадает.

3. Комбинаторные функционалы качества имеют неоспоримое преимущество перед вероятностными — их можно измерять по выборке. Это позволяет отказаться от использования завышенных верхних оценок в методе структурной минимизации риска. Более того, отпадает необходимость в построении структуры вложенных подсемейств различной емкости. Более общий подход заключается в том, чтобы из фиксированного набора методов обучения $\mu_1,\ldots,\mu_T$ выбрать лучший по критерию скользящего контроля на заданной конечной выборке прецедентов. Заметим, что в работе [2] именно этот подход рекомендуется применять на практике, правда, без видимой связи с основными теоретическими результатами. Эмпирические исследования [10] показывают, что данная техника выбора модели во многих случаях предпочтительнее принципов структурной минимизации риска и минимальной длины описания [17], основанных на различных формализациях понятия сложности алгоритма.

4. В работе [18] 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 значения функционала в левой части неравенства

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

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

Далее делается предположение, что зависимость левой части от длины выборки имеет при некотором значении параметров C и h такое же алгебраическое выражение, что и правая часть. Соответствующее значение параметра h и называется *эффективной емкостью*. Предположение хорошо подтверждается в случае линейных решающих правил [18].

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

$$Q^{lk}_{\sup}(A) = \frac{1}{N}\sum_{n=1}^{N}\left[\nu(a_n, X^k_n) - \nu(a_n, X^l_n) > \varepsilon\right]. \tag{6.1}$$

Собственно измерение заключается в том, чтобы оценить данную сумму по меньшему числу разбиений, выбранных случайным образом. Погрешность такого измерения легко оценивается по закону 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 [18].

В комбинаторном подходе функционал равномерного отклонения $Q^{lk}_{\sup}(A)$ заменяется функционалом скользящего контроля $Q^{lk}_{\nu,\varepsilon}$. В этом случае процедура измерения (6.1) 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

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

В отличие от понятия эффективной емкости, предложенного в [18], 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 [18].

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

7. О причинах завышенности сложностных оценок

Для анализа причин завышенности сложностных оценок качества представим отношение правой и левой частей неравенства (4.4) в следующем виде:

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

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. Сложность конечного подсемейства алгоритмов A^l_L, реально получаемых в результате обучения, может оказаться существенно меньше сложности всего семейства A.

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 [19] и выпуклых комбинаций классификаторов [20]. В этих работах учитывались только специфические особенности метода обучения, но не принимались во внимание особенности конкретной выборки и восстанавливаемой зависимости. Полученные оценки все еще сильно завышены, а достаточная длина обучения составляет порядка 10^4 объектов.

Есть основания полагать, что приемлемые численные оценки возможно получить только при явном привлечении априорной информации о свойствах выборки и восстанавливаемой зависимости. Отметим, что соответствие обучающей выборки (локальной информации) и априорных ограничений (универсальной информации) подробно изучается в теории универсальных и локальных ограничений [4], [21] с позиций теории категорий и алгебраического подхода к проблеме распознавания [16]. 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 наиболее сложной и наименее изученной. Комбинаторный подход существенно облегчает развитие данного направления. В частности, уже получена невероятностная оценка функционала Q_c для случая, когда искомая зависимость монотонна или почти монотонна, и метод обучения строит только монотонные отображения [3], [22]. Данная оценка никогда не превышает единицы, не зависит от сложности семейства (имеющего, как известно, бесконечную емкость), и является существенно более точной на малых выборках, чем оценки, полученные ранее в [23], [24].

Автор выражает глубокую признательность Ю. И. Журавлеву за оказываемую поддержку и своему Учителю К. В. Рудакову за постоянное внимание к работе и ценные замечания.

Список литературы

1. *Kohavi R.* A study of cross-validation and bootstrap for accuracy estimation and model selection // IJCAI. 1995. P. 1137–1145.

2. *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979.

3. *Воронцов К. В.* Комбинаторные оценки качества обучения по прецедентам // Докл. РАН. 2004. Т. 394. № 2. С. 175–178.

4. *Журавлев Ю. И., Рудаков К. В.* Об алгебраической коррекции процедур обработки (преобразования) информации // В сб. «Пробл. прикл. матем. и информатики» под ред. О. М. Белоцерковского и др. 1987. С. 187–198.

5. *Vapnik V.* Statistical Learning Theory. New York: Wiley, 1998.

6. *Karpinski M., Macintyre A.* Polynomial bounds for VC dimension of sigmoidal neural networks // 27th ACM Symp. Theor. Comput. 1995. P. 200–208.

7. *Дюличева Ю. Ю.* Оценка VCD r-редуцированного эмпирического леса // Таврич. вестник информатики и матем. 2003. № 1. С. 31–42.

8. *Матросов В. Л.* Емкость алгебраических расширений модели алгоритмов вычисления оценок // Ж. вычисл. матем. и матем. физ. 1984. Т. 24. № 11. С. 1719–1730.

9. *Mazurov V., Khachai M., Rybin A.* Committee constructions for solving problems of selection, diagnostics and prediction // M.: Proc. Inst. math. 2002. V. 1. P. 67–101.

10. *Kearns M. J., Mansour Y., Ng A. Y., Ron D.* An experimental and theoretical comparison of model selection methods // Comput. Learning Theor. 1995. P. 21–30.

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

12. *Mason L., Bartlett P., Baxter J.* Direct optimization of margins improves generalization in combined classifiers: Tech. rept. Dep. Systems Engng, Australian Nat. Univ., 1998.

13. *Schapire R. E., Freund Y., Lee W. S., Bartlett P.* Boosting the margin: a new explanation for the effectiveness of voting methods // Ann. Stat. 1998. V. 26. № 5. P. 1651–1686.

14. *Breiman L.* Bagging predictors // Mach. Learning. 1996. V. 24. № 2. P. 123–140.

15. *Воронцов К. В.* Качество восстановления зависимостей по эмпирическим данным // Матем. методы распознавания образов. 7 Всерос. конф. Тезисы докл. Пущино, 1995. С. 24–26.

16. *Журавлев Ю. И.* Корректные алгебры над множествами некорректных (эвристических) алгоритмов. I–III // Кибернетика. 1977. № 4. С. 5–17. 1977. № 6. С. 21–27. 1978. № 2. С. 35–43.

17. *Rissanen J.* Modeling by shortest data description // Automatica. 1978. V. 14. P. 465–471.

18. *Vapnik V., Levin E., Cun Y. L.* Measuring the VC-dimension of a learning machine // Neural Comput. 1994. V. 6. № 5. P. 851–876.

19. *Bousquet O., Elisseeff A.* Stability and generalization // J. Mach. Learning Res. 2002. № 2. P. 499–526.

20. *Bartlett P.* The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network // IEEE Trans. Informat. Theor. 1998. V. 44. № 2. P. 525–536.

21. *Рудаков К. В.* Универсальные и локальные ограничения в проблеме коррекции эвристических алгоритмов // Кибернетика. 1987. № 2. С. 30–35.

22. *Воронцов К. В.* Оценка качества монотонного решающего правила вне обучающей выборки // Интеллектуализация обработки инф. Тезисы докл. Симферополь, 2002. С. 24–26.

23. *Семочкин А. Н.* Оценки функционала качества для класса алгоритмов с универсальными ограничениями монотонности: — Деп. в ВИНИТИ. 1998. № 2965–В98. С. 20.

24. *Sill J.* The capacity of monotonic functions // Discrete Appl. Math. (special issue on VC dimension). 1998. V. 86. P. 95–107.