Комбинаторная вероятность и точность оценок обобщающей способности (Combinatorial Probability and Generalization Bounds Tightness)

К. В. Воронцов (K. V. Vorontsov), voron@ccas.ru

Computing Centre RAS, Vavilov st. 40, 119333 Moscow, Russia

Аннотация

Accurate prediction of a learning algorithm generalization ability is a most important problem in computational learning theory. The classical Vapnik–Chervonenkis (VC) generalization bounds are too general and therefore extremely overestimate the expected error. Most recent data-dependent bounds are still overestimated. To understand the causes of the bounds looseness we refuse the uniform convergence principle and adhere to a purely combinatorial approach that avoids any probabilistic assumptions, makes no approximations, and gives an empirical control of looseness. We introduce new data-dependent complexity measures: a *local shatter coefficient* and a non-scalar *local shatter profile*, which can give much tighter bounds than the classical *VC shatter coefficient*. An experiment on real datasets shows that the effective local measures may have a very small value, e. g. the *effective local* VC-dimension takes values within $[0, 1]$ and so it is irrelevant to the space dimension.

В задачах анализа данных число наблюдений всегда конечно. В то же время, повсеместно употребляемое понятие *вероятности* определяется либо путём мысленного перехода к бесконечной выборке, либо через вероятностную меру, которая в практических ситуациях, как правило, неизвестна и плохо поддаётся непосредственному измерению в эксперименте. Как указывал А. Н. Колмогоров, «представляется важной задача освобождения всюду, где это возможно, от излишних вероятностных допущений. На независимой ценности чисто комбинаторного подхода к теории информации я неоднократно настаивал в своих лекциях» [1]. Ю. К. Беляев в предисловии к книге [2] пишет: «возникло глубокое убеждение, что в теории выборочных методов можно получить содержательные аналоги большинства основных утверждений теории вероятностей и математической статистики, которые к настоящему времени найдены в предположении взаимной независимости результатов измерений». Таким образом, давно зреет понимание, что можно построить содержательную теорию, которая обращалась бы только с конечными выборками и не опиралась бы на гипотетическое множество «всех мыслимых объектов», из которых почти все никогда не наблюдались и не будут наблюдаться в эксперименте.

В разделе 1 предлагается слабая вероятностная аксиоматика, не опирающаяся на теорию меры, в которой все вероятности непосредственно измеримы в экспери-

менте. Она полностью согласуется с сильной (колмогоровской) аксиоматикой, но область её применимости ограничена задачами анализа данных. В разделе 2 вводится общая постановка задач эмпирического предсказания. В разделе 3 рассматривается задача предсказания частоты события, тесно связанная с законом больших чисел. В разделе 4 рассматривается задача обучения по прецедентам и теория Вапника–Червоненкиса (ТВЧ) в слабой аксиоматике. В разделе 5 анализируются основные причины завышенности оценок ТВЧ и предлагается эмпирическая методика для измерения степени завышенности, обусловленной каждой из причин. В разделе 6 теория обобщающей способности строится для логических закономерностей, рассматривается алгоритм поиска закономерностей в виде информативных конъюнкций. В разделе 7 приводятся результаты эмпирического измерения завышенности оценок ТВЧ для набора реальных задач классификации из репозитория UCI.

1. Слабая вероятностная аксиоматика

Пусть задано множество объектов $\mathbb{X}$. Конечные последовательности объектов будем называть *выборками из* $\mathbb{X}$. Множество всех выборок из $\mathbb{X}$ обозначим через $\mathbb{X}^*$. В любом эксперименте может наблюдаться лишь конечное множество объектов, будь то прошлые или будущие измерения. Поэтому будем рассматривать выборку $X^L = (x_1, \ldots, x_L)$, называемую *генеральной* или *полной* выборкой длины L. Обозначим через S_L группу всех $L!$ перестановок L элементов.

Аксиома 1.1 (о независимости элементов выборки). *Все перестановки генеральной выборки τX^L, $\tau \in S_L$ имеют одинаковые шансы реализоваться.*

Опр. 1.1. *Пусть на множестве выборок задан предикат $\psi\colon \mathbb{X}^* \to \{0,1\}$. Вероятностью события ψ назовём долю перестановок τX^L, при которых предикат истинен:*

$$\mathsf{P}_\tau\, \psi(\tau X^L) = \frac{1}{L!} \sum_{\tau \in S_L} \psi(\tau X^L). \tag{1.1}$$

Эта вероятность зависит от выборки X^L. Мы полагаем, что случайными являются не сами объекты, а только последовательность их появления. При этом знак вероятности P_τ надо понимать как сокращённое обозначение среднего арифметического по всем перестановкам τ. В слабой аксиоматике термин *вероятность* употребляется только в таком смысле — как синоним «доли перестановок выборки».

Опр. 1.2. *Пусть $\xi\colon \mathbb{X}^* \to \mathbb{R}$ — произвольная вещественная функция выборки. Распределением величины ξ на выборке X^L будем называть функцию $F_\xi\colon \mathbb{R} \to [0,1]$ вида*

$$F_\xi(z) = \mathsf{P}_\tau\big[\xi(\tau X^L) < z\big] \tag{1.2}$$

Опр. 1.3. *Математическим ожидаемым величины $\xi\colon \mathbb{X}^* \to \mathbb{R}$ на выборке X^L будем называть её среднее арифметическое по всем перестановкам τ:*

$$\mathsf{E}_\tau \xi(\tau X^L) = \frac{1}{L!} \sum_{\tau \in S_L} \xi(\tau X^L). \tag{1.3}$$

Заметим, что вероятность и матожидание формально определяются одинаково, как среднее арифметическое: $\mathsf{P}_\tau \equiv \mathsf{E}_\tau \equiv \frac{1}{L!}\sum_{\tau\in S_L}$.

Рассмотрим важный частный случай, когда предикат ψ является функцией двух подвыборок: $\psi(X^L) = \varphi(X^\ell, X^k)$, где $X^\ell \cup X^k = X^L$, $\ell + k = L$, причём значение предиката φ не зависит от порядка элементов в подвыборках X^ℓ и X^k. Рассмотрим множество всех $N = C_L^\ell$ разбиений генеральной выборки X^L на две подвыборки X_n^ℓ и X_n^k, где нижний индекс $n = 1, \ldots, N$ обозначает номер разбиения. Тогда из основной аксиомы 1.1 следует, что все разбиения имеют равные шансы реализоваться, и вероятность можно определять как долю разбиений выборки X^L:

$$\mathsf{P}_\tau\, \psi(\tau X^L) = \mathsf{P}_n\, \varphi(X_n^\ell, X_n^k) = \frac{1}{N}\sum_{n=1}^{N} \varphi(X_n^\ell, X_n^k).$$

Сопоставление с сильной вероятностной аксиоматикой. В классической (колмогоровской) аксиоматике на множестве объектов $\mathbb{X}$ вводится вероятностное пространство $\langle \mathbb{X}, \Omega, \mathsf{P}\rangle$, где Ω — аддитивная σ-алгебра событий на $\mathbb{X}$, а P — вероятностная мера, определённая на элементах из Ω, как правило, неизвестная. Рассматриваются случайные выборки независимых наблюдений, полученных согласно мере P, и исследуются некоторые измеримые функции этих выборок.

В слабой аксиоматике вероятностная мера вводится на конечном множестве перестановок (или разбиений), причём вероятностное распределение равномерно. Столь слабых вероятностных допущений уже достаточно для получения многих фундаментальных фактов теории вероятностей и математической статистики.

Если вероятность (1.1) вычислена в слабой аксиоматике, $\mathsf{P}_\tau\, \psi(\tau X^L) = p(X^L)$, то результат может быть легко перенесён и в сильную. Действительно, если принять гипотезу о независимости наблюдений в выборке X^L, то для любой перестановки τ выполняется $\mathsf{P}_{X^L}\, \psi(X^L) = \mathsf{P}_{X^L}\, \psi(\tau X^L)$, следовательно,

$$\mathsf{P}_{X^L}\, \psi(X^L) = \mathsf{E}_{X^L}\, \mathsf{P}_\tau\, \psi(\tau X^L) = \mathsf{E}_{X^L} p(X^L).$$

Перенос результата в сильную аксиоматику сводится к применению операции матожидания E по выборке X^L к полученной вероятности $p(X^L)$. В случаях, когда эта вероятность не зависит от выборки X^L, результат переносится непосредственно. Таким образом, соблюдается «принцип соотвествия» — две теории приводят к одинаковым результатам на пересечении областей их применимости.

В сильной аксиоматике вероятности, функции распределения, математические ожидания являются ненаблюдаемыми — они выражаются либо через предельный переход к бесконечной выборке, либо через вероятностную меру P на множестве объектов $\mathbb{X}$. На практике ни то, ни другое, как правило, не известно. В слабой аксиоматике рассматриваются исключительно *статистики* — функции конечных выборок $z\colon \mathbb{X}^* \to Z$. При анализе данных оценивание ненаблюдаемых вероятностей является искусственной задачей, в некоторой степени оторванной от практики.

2. Задачи эмпирического предсказания

Задача эмпирического предсказания состоит в том, чтобы, получив выборку данных, предсказать определённые свойства аналогичных данных, которые станут известны позже, и оценить точность предсказания.

Пусть задано множество R и функция $T\colon \mathbb{X}^* \times \mathbb{X}^* \to R$. Рассмотрим эксперимент, в котором реализуется одно из равновероятных разбиений выборки X^L на две подвыборки X_n^ℓ и X_n^k, $n = 1, \ldots, N$. После реализации разбиения n наблюдателю сообщается подвыборка X_n^ℓ. Не зная скрытой подвыборки X_n^k, он должен предсказать значение $T_n = T(X_n^k, X_n^\ell)$, существенно зависящее от X_n^k. Необходимо также оценить надёжность предсказания, то есть вероятность того, что неизвестное истинное значение T_n не сильно отличается от сделанного предсказания.

Задача 2.1. Построить *предсказывающую функцию* $\hat{T}\colon \mathbb{X}^* \to R$, значение которой $\hat{T}_n = \hat{T}(X_n^\ell)$ на наблюдаемой подвыборке X_n^ℓ приближало бы неизвестное истинное значение $T_n = T(X_n^k, X_n^\ell)$, и оценить надёжность предсказания, указав невозрастающую *оценочную функцию* $\eta(\varepsilon)$ такую, что

$$\mathsf{P}_n\big[d(\hat{T}_n, T_n) \geqslant \varepsilon\big] \leqslant \eta(\varepsilon), \tag{2.1}$$

где $d\colon R \times R \to \mathbb{R}$ — заданная функция, характеризующая величину отклонения $d(\hat{T}_n, T_n)$ предсказанного значения $\hat{T}_n$ от неизвестного истинного значения T_n.

Параметр ε называется *точностью*, а величина $(1-\eta(\varepsilon))$ — *надёжностью* предсказания. Если в (2.1) достигается равенство, то $\eta(\varepsilon)$ называется *точной оценкой*. Оценка $\eta(\varepsilon)$ может зависеть от ℓ и k, а также от вида функций T и $\hat{T}$. Обычно предполагается, что $\varepsilon > 0$ и $0 < \eta < 1$. Эмпирическое предсказание можно считать состоятельным, если (2.1) выполняется при достаточно малых ε и η.

Замечание 2.1. Если функция $T(U, V)$ зависит только от U, то второй аргумент V условимся опускать. В некоторых задачах полагают $T(U) = \hat{T}(U)$. Тем не менее, роль функций T и $\hat{T}$ принципиально различна. Функция T предполагается заданной заранее и входит в постановку задачи. Функцию предсказания $\hat{T}$ наблюдатель имеет право выбирать по собственному усмотрению.

Замечание 2.2. Предсказание некоторого свойства выборки на основании свойств другой выборки называется *трансдукцией* или переходом от частного к частному. Принято считать, что трансдукция более примитивна и ограничена, чем *индукция* — переход от частного к общему. В нашем случае это не совсем так. Если удаётся получить оценку $\eta(\varepsilon)$, справедливую *для любой* выборки X^L или хотя бы для широкого класса выборок, то трансдукция приобретает общность индукции.

Примеры задач эмпирического предсказания. Выбирая множество R, функции T и $\hat{T}$, семейство Ω_ε или функцию d, можно получать постановки различных задач теории вероятностей, математической статистики, машинного обучения.

Задача 2.2 (оценивание частоты события). Пусть $S \subseteq \mathbb{X}$ — некоторое множество объектов; будем называть его «событием». Введём функцию *частоты события* S на конечной выборке U:

$$\nu_S(U) = \frac{1}{|U|} \sum_{x \in U} [x \in S], \quad U \in \mathbb{X}^*.$$

Положим $R = \mathbb{R}$, $T(U) = \hat{T}(U) = \nu_S(U)$, $d(\hat{r}, r) = |r - \hat{r}|$.

Требуется предсказать частоту события S на скрытой выборке X_n^k по его частоте на наблюдаемой выборке X_n^ℓ и оценить надёжность предсказания:

$$\mathsf{P}_n\big[|\nu_S(X_n^k) - \nu_S(X_n^\ell)| \geqslant \varepsilon\big] \leqslant \eta(\varepsilon); \tag{2.2}$$

В некоторых случаях требуется получить одностороннюю, скажем, верхнюю оценку. Тогда надо положить $d(\hat{r}, r) = r - \hat{r}$:

$$\mathsf{P}_n\big[\nu_S(X_n^k) - \nu_S(X_n^\ell) \geqslant \varepsilon\big] \leqslant \eta(\varepsilon). \tag{2.3}$$

Эта задача имеет фундаментальное значение для теории вероятностей и тесно связана с законом больших чисел и предельными теоремами. Далее для (2.2) и (2.3) будут получены точные оценки. Они возникают и в практических приложениях, например, в выборочном контроле качества [2].

Задача 2.3 (оценивание функции распределения). Определим для произвольной функции $\xi\colon \mathbb{X} \to \mathbb{R}$ и произвольной конечной выборки $U \in \mathbb{X}^*$ *эмпирическую функцию распределения* $F_\xi\colon \mathbb{R} \to [0,1]$. Она показывает, на какой доле объектов выборки значение $\xi(x)$ не превосходит z:

$$F_\xi(z, U) = \frac{1}{|U|} \sum_{x \in U} [\xi(x) \leqslant z].$$

В качестве R возьмём множество всех неубывающих кусочно-постоянных функций $F\colon \mathbb{R} \to [0,1]$. Введём на R равномерную метрику $d(\hat{r}, r) = \max\limits_{z \in \mathbb{R}} |r(z) - \hat{r}(z)|$. Положим $T(U) = \hat{T}(U) = F_\xi(z, U)$.

Требуется предсказать максимальное отклонение функции распределения на скрытой выборке $F_\xi(z, X_n^k)$ от известной функции распределения на наблюдаемой выборке $F_\xi(z, X_n^\ell)$ и оценить надёжность предсказания:

$$\mathsf{P}_n\Big[\max_{z \in \mathbb{R}} |F_\xi(z, X_n^k) - F_\xi(z, X_n^\ell)| \geqslant \varepsilon\Big] \leqslant \eta(\varepsilon).$$

Данная задача тесно связана со сходимостью эмпирических распределений и имеет фундаментальное значение для математической статистики. На этой оценке основан критерий Смирнова, применяемый для проверки гипотезы однородности (одинаковой распределённости) двух выборок [3, 4]. Для данной задачи также существует точная оценка, но её рассмотрение выходит за рамки работы.

Задача 2.4 (обучение по прецедентам). Задано множество допустимых ответов $\mathbb{Y}$. Существует неизвестная *целевая зависимость* (target function) $y^*\colon \mathbb{X} \to \mathbb{Y}$, которая каждому объекту $x \in \mathbb{X}$ ставит в соответствие правильный ответ $y^*(x)$. Задана *функция потерь* (loss function) $\mathscr{L}\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$, значение которой $\mathscr{L}(y, y')$ характеризует величину ошибки ответа y при правильном ответе y'. Функции вида $a\colon \mathbb{X} \to \mathbb{Y}$, допускающие эффективную компьютерную реализацию, называются *алгоритмами*. *Средняя ошибка* алгоритма $a\colon \mathbb{X} \to \mathbb{Y}$ на конечной выборке U есть

$$\nu(a, U) = \frac{1}{|U|} \sum_{x \in U} \mathscr{L}\big(a(x), y^*(x)\big), \quad U \in \mathbb{X}^*.$$

Метод обучения (learning algorithm) $\mu\colon \mathbb{X}^* \to \mathbb{Y}^{\mathbb{X}}$, по наблюдаемой *обучающей* (training) выборке X_n^ℓ с известными ответами $y_i = y^*(x_i),\ x_i \in X_n^\ell$ строит алгоритм $a_n = \mu X_n^\ell$. Когда средняя ошибка на скрытой *контрольной* (testing) выборке $\nu(a_n, X_n^k)$ существенно превосходит среднюю ошибку обучения $\nu(a_n, X_n^\ell)$, говорят, что алгоритм a_n *переобучен* [5, 6].

Введём функцию разности средней ошибки алгоритма a_n на двух выборках:

$$\delta(a_n, X_n^\ell, X_n^k) = \nu(a_n, X_n^k) - \nu(a_n, X_n^\ell).$$

Опр. 2.1. *Переобученностью (overfitting) алгоритма $a_n = \mu X_n^\ell$ будем называть разность его средней ошибки на контрольной и обучающей выборках $\delta(\mu X_n^\ell, X_n^\ell, X_n^k)$.*

Положим $R = \mathbb{R},\ T_n = \nu(a_n, X_n^k),\ \hat{T}_n = \nu(a_n, X_n^\ell),\ d(\hat{r}, r) = r - \hat{r}$. Требуется предсказать верхнюю границу переобученности и оценить надёжность предсказания:

$$\mathsf{P}_n\big[\nu(a_n, X_n^k) - \nu(a_n, X_n^\ell) \geqslant \varepsilon\big] \leqslant \eta(\varepsilon). \tag{2.4}$$

Предотвращение переобучения является серьёзной проблемой при построении алгоритмов классификации и регрессии по конечным выборкам данных [7].

Эмпирическое оценивание вероятности. Результаты, полученные в слабой аксиоматике, всегда могут быть проверены экспериментально. Пусть имеется совокупность значений $\varphi_n = \varphi(X_n^\ell, X_n^k),\ n = 1, \ldots, N$. Чтобы оценить значение суммы

$$Q_N = \mathsf{P}_n\, \varphi_n \equiv \frac{1}{N} \sum_{n=1}^{N} \varphi_n,$$

заменим суммирование по всем N разбиениям суммированием по некоторому подмножеству разбиений $N' \subset \{1, \ldots, N\}$, не слишком большому — чтобы сумма вычислялась за приемлемое время:

$$Q_N \approx Q(N') = \hat{\mathsf{P}}_n \varphi_n \equiv \frac{1}{|N'|} \sum_{n \in N'} \varphi_n.$$

Например, в методе Монте-Карло подмножество разбиений N' выбирается случайно и независимо из равномерного распределения на $\{1, \ldots, N\}$. В этом случае оценивание точности приближения $|Q(N') - Q_N|$ сводится к Задаче 2.2, только теперь в качестве объектов рассматриваются разбиения.

Будем называть $\hat{\mathsf{P}}_n \equiv \frac{1}{N'} \sum_{n \in N'}$ *эмпирической оценкой вероятности*.

Эмпирическое оценивание имеет несколько существенных недостатков. Оно требует знания полной выборки X^L, и потому не может быть использовано непосредственно для эмпирического предсказания. Оно не позволяет получать оценки в аналитическом виде. Наконец, оно может потребовать большого объёма вычислений.

Таким образом, область применимости эмпирического оценивания довольно ограничена. На практике оно используется для экспериментального исследования зависимости Q_N от некоторых параметров задачи (например, от длины выборки). В задачах обучения по прецедентам эмпирическое оценивание называют *скользящим контролем* (cross-validation) и используют для оценивания качества метода обучения μ, а не отдельного алгоритма. Оно незаменимо в тех случаях, когда теоретические оценки Q_N не известны или сильно завышены. В данной работе эмпирическое оценивание применяется для анализа точности теоретических оценок.

3. Задача оценивания частоты события

Рассмотрим Задачу 2.2 о предсказании частоты события $S \subseteq \mathbb{X}$. Пусть число элементов события S во всей выборке X^L фиксировано и равно $m = L\nu_S(X^L)$. Тогда число элементов S в наблюдаемой подвыборке $\ell\nu_S(X_n^\ell)$ и число элементов S в скрытой подвыборке $k\nu_S(X_n^k)$ подчиняются гипергеометрическому распределению:

$$\mathsf{P}_n\big[\ell\nu_S(X_n^\ell) = s\big] = \mathsf{P}_n\big[k\nu_S(X_n^k) = m - s\big] = h\big(\tbinom{\ell\ s}{L\ m}\big) = \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}, \tag{3.1}$$

где s принимает значения от $s_0(m) = \max\{0, m-k\}$ до $s_1(m) = \min\{\ell, m\}$.

Введём сокращённые обозначения $\nu_n^\ell = \nu_S(X_n^\ell)$, $\nu_n^k = \nu_S(X_n^k)$.

Теорема 3.1. *Для любого $\varepsilon \in [0,1)$ справедливы точные оценки:*

$$\mathsf{P}_n\big[\nu_n^\ell \leqslant \varepsilon\big] = H\big(\tbinom{\ell\ \lfloor\varepsilon\ell\rfloor}{L\ m}\big);$$

$$\mathsf{P}_n\big[\nu_n^k \geqslant \varepsilon\big] = H\big(\tbinom{\ell\ \lfloor m-\varepsilon k\rfloor}{L\ m}\big);$$

$$\mathsf{P}_n\big[\nu_n^k - \nu_n^\ell \geqslant \varepsilon\big] = H\big(\tbinom{\ell\ s_m^-(\varepsilon)}{L\ m}\big), \qquad s_m^-(\varepsilon) = \big\lfloor \tfrac{\ell}{L}(m - \varepsilon k)\big\rfloor; \tag{3.2}$$

$$\mathsf{P}_n\big[|\nu_n^k - \nu_n^\ell| \geqslant \varepsilon\big] = H\big(\tbinom{\ell\ s_m^-(\varepsilon)}{L\ m}\big) + \bar{H}\big(\tbinom{\ell\ s_m^+(\varepsilon)}{L\ m}\big), \qquad s_m^+(\varepsilon) = \big\lceil \tfrac{\ell}{L}(m + \varepsilon k)\big\rceil. \tag{3.3}$$

Замечание 3.1. В условии теоремы под $\lfloor z\rfloor$ понимается целая часть действительного числа z, то есть наибольшее целое, *меньшее или равное* z. Аналогично, $\lceil z\rceil$ — наименьшее целое, *большее или равное* z. Если в левой части поменять нестрогие неравенства на строгие, то все оценки останутся в силе с одной оговоркой: $\lfloor z\rfloor$ надо будет понимать как наибольшее целое, *меньшее* z; оно отличается от функции целой части только тем, что $\lfloor z\rfloor = z - 1$ при целых z. Соответственно, и $\lceil z\rceil$ надо будет понимать как наименьшее целое, *большее* z, тогда $\lceil z\rceil = z + 1$ при целых z.

Доказательство. Первые два неравенства являются непосредственным следствием (3.1), поэтому начнём сразу с доказательства (3.2). Сгруппируем все слагаемые с одинаковыми значениями $s = \ell\nu_n^\ell$ и просуммируем их по-отдельности:

$$\mathsf{P}_n\big[\nu_n^k - \nu_n^\ell \geqslant \varepsilon\big] = \sum_{s=s_0}^{s_1} \mathsf{P}_n\Big[\nu_n^\ell = \frac{s}{\ell}\Big]\Big[\frac{m-s}{k} - \frac{s}{\ell} \geqslant \varepsilon\Big].$$

Наибольшее целое, для которого выполняется неравенство $\frac{m-s}{k} - \frac{s}{\ell} \geqslant \varepsilon$, как раз и есть $s_m^-(\varepsilon)$, поэтому полученное выражение можно переписать короче:

$$\mathsf{P}_n\big[\nu_n^k - \nu_n^\ell \geqslant \varepsilon\big] = \sum_{s=s_0}^{s_m^-(\varepsilon)} \mathsf{P}_n\Big[\nu_n^\ell = \frac{s}{\ell}\Big] = \sum_{s=s_0}^{s_m^-(\varepsilon)} h\big(\tbinom{\ell\ s}{L\ m}\big) = H\big(\tbinom{\ell\ s_m^-(\varepsilon)}{L\ m}\big).$$

Двусторонняя оценка (3.3) доказывается аналогично, если разбить множество разбиений на два непересекающихся подмножества:

$$\mathsf{P}_n\big[|\nu_n^k - \nu_n^\ell| \geqslant \varepsilon\big] = \mathsf{P}_n\big[\nu_n^k - \nu_n^\ell \geqslant \varepsilon\big] + \mathsf{P}_n\big[\nu_n^\ell - \nu_n^k \geqslant \varepsilon\big] = H\big(\tbinom{\ell\ s_m^-(\varepsilon)}{L\ m}\big) + \bar{H}\big(\tbinom{\ell\ s_m^+(\varepsilon)}{L\ m}\big).$$

Теорема доказана. ■

Верхняя оценка. Число элементов m события S в полной выборке X^L невозможно узнать, пока неизвестна скрытая часть данных. В то же время, от него зависят оценочные функции (3.2) и (3.3). Простейшее решение этой проблемы — взять максимум по m и получить вместо точной оценки завышенную верхнюю оценку:

$$\mathsf{P}_n\big[\nu_n^k - \nu_n^\ell \geqslant \varepsilon\big] \leqslant \max_{m=0,\ldots,L} H\big(\tbinom{\ell}{L}, s_m^-(\varepsilon)\big) \equiv \Gamma_L^\ell(\varepsilon). \tag{3.4}$$

Здесь максимум достаточно брать по всем m от $\lceil \varepsilon k \rceil$ до $\lfloor L - \varepsilon\ell \rfloor$, так как при остальных значениях m левая часть неравенства равна нулю.

Известно асимптотическое поведение величины $\Gamma_L^\ell(\varepsilon)$ [5]: при любом $\varepsilon > 0$

$$\Gamma_L^\ell(\varepsilon) \sim \exp\left(-2\varepsilon^2 \frac{\ell k}{\ell + k}\right), \quad \ell, k \to \infty,$$

откуда следует, что вероятности $\mathsf{P}_n\big[\nu_n^k - \nu_n^\ell \geqslant \varepsilon\big]$ и $\mathsf{P}_n\big[|\nu_n^k - \nu_n^\ell| \geqslant \varepsilon\big]$ стремятся к нулю при одновременном стремлении ℓ и k к бесконечности. Это означает, что равенства (3.2) и (3.3) являются аналогом закона больших чисел в слабой аксиоматике.

4. Задачи обучения по прецедентам

Уточним постановку Задачи 2.4 о предсказании качества обучения по прецедентам. Будем рассматривать только бинарные функции потерь, полагая

$$\mathscr{L}(y, y') = \big[\text{ответ } y \text{ является ошибочным при правильном ответе } y'\big].$$

Выбор функции потерь зависит от конкретной задачи, в первую очередь от множества допустимых ответов $\mathbb{Y}$. В задачах классификации, когда $\mathbb{Y}$ — конечное множество классов, полагают $\mathscr{L}(y, y') = [y \neq y']$. В задачах восстановления регрессии, когда $\mathbb{Y} = \mathbb{R}$, принято использовать гладкие функции потерь, например $\mathscr{L}(y, y') = (y - y')^2$. Однако введение бинарной функции потерь также возможно: $\mathscr{L}(y, y') = \big[|y - y'| \geqslant d\big]$, где d — фиксированное пороговое значение.

Для дальнейшего вид бинарной функции потерь не важен. Основные результаты справедливы для широкого класса задач, включая и классификацию, и регрессию.

Классическая теория Вапника-Червоненкиса [8, 9, 5], далее ТВЧ, основана на колмогоровской вероятностной аксиоматике. Предполагается, что множество объектов $\mathbb{X}$ является вероятностным пространством с некоторым неизвестным распределением вероятностей, и все рассматриваемые выборки случайны и независимы.

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

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

Такой способ обучения алгоритма называется *методом минимизации эмпирического риска*. В семействе может существовать много алгоритмов, минимизирующих эмпирический риск. Предполагается, что в качестве решения может быть выбран любой из этих алгоритмов. Другие методы обучения в ТВЧ не рассматриваются.

Качество алгоритма a^* характеризуется вероятностью ошибки $P(a^*)$. Достаточным условием *обучаемости* (learnability) является малое отклонение частоты ошибки $\nu(a, X^\ell)$ от её вероятности $P(a)$ для любого $a \in A$. Точнее, при достаточно малых значениях точности ε и надёжности η должна быть справедлива оценка

$$P_\varepsilon(A) = \mathsf{P}\Big\{\sup_{a\in A} \big|P(a) - \nu(a, X^\ell)\big| > \varepsilon\Big\} \leqslant \eta(\varepsilon). \tag{4.1}$$

Введение супремума даёт гарантированную оценку, справедливую независимо от того, какой именно алгоритм a^* будет получен в результате обучения. Если правая часть (4.1) стремится к нулю при $\ell, k \to \infty$, то говорят, что имеет место *равномерная сходимость* частоты ошибок к их вероятности.

Можно также характеризовать качество алгоритма a^* частотой ошибок $\nu(a^*, X^k)$ на случайной независимой контрольной выборке X^k. Тогда оценки получаются более точными в силу «основной леммы», доказанной в [9, стр. 219] при $\ell = k$:

$$\mathsf{P}\Big\{\sup_{a\in A} |\,\mathsf{P}(a) - \nu(a, X^\ell)| > \varepsilon\Big\} \leqslant 2\,\mathsf{P}\Big\{\max_{a\in A} |\nu(a, X^k) - \nu(a, X^\ell)| > \tfrac{1}{2}\varepsilon\Big\}. \tag{4.2}$$

Позже эта оценка была уточнена [5]: в правой части $\frac{1}{2}\varepsilon$ заменено на $\varepsilon - \frac{1}{\ell}$.

Если правая часть стремится к нулю при $\ell, k \to \infty$, то говорят, что имеет место *равномерная сходимость* частоты ошибок в двух выборках.

Вполне достаточно учитывать только положительные отклонения частот, поскольку отрицательные отклонения $\nu(a, X^k) - \nu(a, X^\ell) < 0$ свидетельствуют о хорошей обучаемости. При этом точность повышается ещё вдвое, и приходим к функционалу равномерного одностороннего отклонения частот в двух выборках:

$$P_\varepsilon(A) = \mathsf{P}\Big\{\max_{a\in A} \big(\nu(a, X^k) - \nu(a, X^\ell)\big) > \varepsilon\Big\}. \tag{4.3}$$

При $\ell = k$ для любого распределения вероятностей на $\mathbb{X}$ и любой целевой зависимости y^* справедлива оценка скорости равномерной сходимости [9]:

$$P_\varepsilon(A) \leqslant \Delta^A(2\ell) \cdot \tfrac{3}{2}\, e^{-\varepsilon^2\ell}, \tag{4.4}$$

где $\Delta^A(L)$ — функция роста семейства алгоритмов A. Чтобы дать определение функции роста, введём два вспомогательных понятия.

Опр. 4.1. *Алгоритмы a и a' неразличимы на выборке X^L, если они допускают ошибки на одних и тех же объектах: $\mathscr{L}(a(x_i), y_i) = \mathscr{L}(a'(x_i), y_i)$ для всех $x_i \in X^L$.*

Неразличимость является отношением эквивалентности на множестве A.

Опр. 4.2. *Коэффициент разнообразия (shatter coefficient) $\Delta(A, X^L)$ множества алгоритмов A на выборке X^L — это число классов эквивалентности, индуцируемых на множестве A отношением неразличимости алгоритмов на выборке X^L.*

Можно сказать и так: коэффициент разнообразия — это количество различных бинарных векторов вида $\big[\mathscr{L}(a(x_i), y^*(x_i))\big]_{i=1}^L$, порождаемых всевозможными алгоритмами $a \in A$ на заданной выборке X^L. В задачах классификации на два класса коэффициент разнообразия равен числу различных *дихотомий* (способов разделить

выборку на два класса), реализуемых всевозможными алгоритмами из A. Отметим, что в работах 70-х годов [10, 8, 9] коэффициент разнообразия назывался *индексом системы событий*. Алгоритм a индуцирует событие $S_a = \big\{x \in \mathbb{X} \;\big|\; \mathscr{L}(a(x), y^*(x)) = 1\big\}$. Семейство A индуцирует систему событий $S = \{S_a | \, a \in A\}$. Индекс системы событий S есть число различных подмножеств вида $S_a \cap X^L$, $a \in A$, что равносильно Определению 4.2. В англоязычных работах прижился термин shattering — число разбиений всеми возможными способами, буквальный перевод — «вдребезги».

Опр. 4.3. *Функцией роста (growth function) множества алгоритмов A называется максимальное значение коэффициента разнообразия $\Delta^A(X^L)$ по всем возможным выборкам длины L:*

$$\Delta^A(L) = \max_{X^L} \Delta(A, X^L), \quad L = 1, 2, 3, \ldots$$

Функция роста не зависит ни от выборки, ни от метода обучения, и является мерой сложности множества алгоритмов A. Очевидна верхняя оценка $\Delta^A(L) \leqslant 2^L$.

Минимальное число h, при котором $\Delta^A(h) < 2^h$, называется *ёмкостью* или *размерностью Вапника-Червоненкиса* (VC-dimension) семейства алгоритмов A. Если такого h не существует, то говорят, что ёмкость A бесконечна. Доказано, что если A имеет конечную ёмкость h, то его функция роста зависит от L полиномиально:

$$\Delta^A(L) \leqslant C_L^0 + C_L^1 + \cdots + C_L^h \leqslant \frac{3}{2}\frac{L^h}{h!}. \tag{4.5}$$

В этом случае имеет место равномерная сходимость, и семейство A является обучаемым. Таким образом, в ТВЧ для оценивания качества обучения по прецедентам достаточно знать только длину выборки и ёмкость семейства алгоритмов.

Практическому применению описанного подхода препятствует чрезвычайная завышенность оценки (4.4). Чтобы в этом убедиться, достаточно выполнить численный расчёт требуемой длины обучающей выборки ℓ как функции от (h, η, ε). Она имеет порядки 10^5–10^9, что существенно превышает количество объектов, с которым обычно приходится иметь дело на практике [7].

Причина завышенности оценок ТВЧ — в их чрезмерной общности. Они справедливы при любом распределении вероятностей на $\mathbb{X}$, любой целевой зависимости $y^*(x)$ и любом методе обучения μ. Это «пессимистичные» оценки, ориентированные на худший случай, который вряд ли когда-нибудь встретится на практике.

С введением понятия *метода обучения* μ становится очевидно, что сам функционал равномерной сходимости уже является завышенной верхней оценкой:

$$\mathsf{P}\Big\{\delta(\mu X^\ell, X^\ell, X^k) > \varepsilon\Big\} \leqslant \mathsf{P}\Big\{\max_{a \in A} \delta(a, X^\ell, X^k) > \varepsilon\Big\}. \tag{4.6}$$

Малое значение левой, а не правой, части этого неравенства следовало бы принимать за определение *обучаемости* по прецедентам. В ТВЧ много внимания уделяется необходимым и достаточным условиям равномерной сходимости. Однако равномерная сходимость, как следует из (4.6), является лишь достаточным условием обучаемости. Если ёмкость бесконечна и равномерной сходимости нет, то рано делать вывод, что нет обучаемости. Распространённой ошибкой в интерпретации ТВЧ является вывод о необходимости ограничивать сложность семейства алгоритмов. Такой вывод был бы справедлив, если бы оценки ТВЧ были достаточно точны.

Теория Вапника-Червоненкиса в слабой аксиоматике. Допустим, что в Задаче 2.4 (обучения по прецедентам) метод μ выдаёт для любой выборки один и тот же фиксированный алгоритм a. В такой упрощённой постановке задача сводится к оцениванию частоты события $S = \{x \in \mathbb{X} \mid \mathscr{L}(a(x), y^*(x)) = 1\}$, и из (3.4) следует

Утверждение 4.1. *Для любого $a\colon \mathbb{X} \to \mathbb{Y}$ и любого $\varepsilon \in [0,1)$ справедлива оценка*

$$\mathsf{P}_n\big[\nu(a, X_n^k) - \nu(a, X_n^\ell) \geqslant \varepsilon\big] < \Gamma_L^\ell(\varepsilon). \tag{4.7}$$

В этом частном случае работает закон больших чисел: частоту события S на контрольной выборке можно предсказать по его частоте на обучающей выборке, и точность предсказания увеличивается с ростом длины выборок.

Теперь рассмотрим общий случай, когда метод μ на разных обучающих выборках может выдавать различные алгоритмы. Обозначим через A_L^ℓ множество алгоритмов, порождаемых методом μ по всевозможным подвыборкам $X_n^\ell \subset X^L$:

$$A_L^\ell \equiv A_L^\ell(\mu, X^L) = \{a_n = \mu X_n^\ell \mid n = 1, \ldots, N\}.$$

Мощность множества A_L^ℓ не превосходит $N = C_L^\ell$. Она может быть и меньше N, если на различных подвыборках метод μ строит одинаковые алгоритмы. Коэффициент разнообразия множества A_L^ℓ может оказаться ещё меньше, если некоторые алгоритмы, не совпадающие как отображения $\mathbb{X} \to \mathbb{Y}$, неразличимы на выборке X^L.

Опр. 4.4. *Коэффициент разнообразия множества алгоритмов $A_L^\ell(\mu, X^L)$ будем называть локальным коэффициентом разнообразия (local shatter coefficient) метода μ на выборке X^L и обозначать $\Delta_L^\ell \equiv \Delta_L^\ell(\mu, X^L) = \Delta\big(A_L^\ell(\mu, X^L), X^L\big)$.*

Множество алгоритмов A_L^ℓ разбивается на $L+1$ подмножеств алгоритмов A_m, допускающих на генеральной выборке X^L заданное число ошибок $m = 0, 1, \ldots, L$:

$$A_m \equiv A_{L,m}^\ell(\mu, X^L) = \{a_n = \mu X_n^\ell \mid \nu(a_n, X^L) = \tfrac{m}{L},\ n = 1, \ldots, N\};$$
$$A_L^\ell = A_0 \cup A_1 \cup \cdots \cup A_L.$$

Опр. 4.5. *Последовательность коэффициентов разнообразия $D_m \equiv \Delta_{L,m}^\ell(\mu, X^L) = \Delta\big(A_{L,m}^\ell(\mu, X^L), X^L\big)$, $m = 0, 1, \ldots, L$ будем называть локальным профилем разнообразия (local shatter profile) метода μ на выборке X^L.*

Множества A_m не пересекаются и в объединении дают A_L^ℓ. Поэтому

$$\Delta_L^\ell = D_0 + D_1 + \cdots + D_L. \tag{4.8}$$

Напомним, что в Задаче 2.4 (обучения по прецедентам) нас интересуют верхние оценки функционала (2.4):

$$Q_\varepsilon \equiv Q_\varepsilon(\mu, X^L) = \mathsf{P}_n\big[\delta_n \geqslant \varepsilon\big],$$

где $\delta_n = \delta(a_n, X_n^\ell, X_n^k)$ — переобученность алгоритма $a_n = \mu X_n^\ell$.

Разобьём функционал Q_ε на $L+1$ слагаемых $Q_{\varepsilon,m} \equiv Q_{\varepsilon,m}(\mu, X^L)$ по параметру m:

$$Q_\varepsilon = \sum_{m=0}^{L} \mathsf{P}_n\big[\delta_n \geqslant \varepsilon\big]\big[\nu(a_n, X^L) = \tfrac{m}{L}\big] = \sum_{m=0}^{L} Q_{\varepsilon,m}.$$

Теорема 4.2. *Для любых $\varepsilon \in [0,1)$ и $m = 0, 1, \ldots, L$ справедлива оценка*

$$Q_{\varepsilon,m} \leqslant D_m H\tbinom{\ell}{L}{\bar{s}_m(\varepsilon)}{m}. \tag{4.9}$$

Доказательство. Отношение неразличимости алгоритмов на выборке X^L разбивает множество алгоритмов A_m на классы эквивалентности A_{md}, где $d = 1, \ldots, D_m$ — порядковый номер класса среди всех D_m классов, алгоритмы которых допускают m ошибок. Запишем P_n через сумму разбиений, взятых отдельно для каждого класса эквивалентности:

$$Q_{\varepsilon,m} = \sum_{d=1}^{D_m} \frac{1}{N} \sum_{n=1}^{N} \big[a_n \in A_{md}\big] \big[\nu(a_n, X_n^k) - \nu(a_n, X_n^\ell) \geqslant \varepsilon\big].$$

Значение функционала не изменится, если алгоритм $a_n = \mu X_n^\ell$ заменить на произвольный элемент a_{md} из класса эквивалентности A_{md}. Применим тот же приём, что и в доказательстве теоремы 3.1 — перегруппируем слагаемые по значениям числа ошибок s на обучающей выборке:

$$\begin{aligned} Q_{\varepsilon,m} &= \sum_{d=1}^{D_m} \sum_{s=s_0}^{\min\{\ell,m\}} \frac{1}{N} \sum_{n=1}^{N} \Big[a_n \in A_{md}\Big] \Big[\nu(a_{md}, X_n^\ell) = \frac{s}{\ell}\Big] \Big[\frac{m-s}{k} - \frac{s}{\ell} \geqslant \varepsilon\Big] = \\ &= \sum_{d=1}^{D_m} \sum_{s=s_0}^{\bar{s}_m(\varepsilon)} \underbrace{\frac{1}{N} \sum_{n=1}^{N} \Big[a_n \in A_{md}\Big] \Big[\nu(a_{md}, X_n^\ell) = \frac{s}{\ell}\Big]}_{\gamma(m,s)}. \end{aligned}$$

Оценим сверху внутреннюю сумму $\gamma(m,s)$, заменив $[a_n \in A_{md}]$ единицей. Рассуждая аналогично доказательству теоремы 3.1, получим $\gamma(m,s) \leqslant h\tbinom{\ell}{L}{s}{m}$. Эта величина уже не зависит от d, поэтому её можно вынести за знак суммирования по d:

$$Q_{\varepsilon,m} \leqslant D_m \sum_{s=s_0}^{\bar{s}_m(\varepsilon)} h\tbinom{\ell}{L}{s}{m} = D_m H\tbinom{\ell}{L}{\bar{s}_m(\varepsilon)}{m}.$$

Теорема доказана. ■

Теорема 4.3. *Для любого $\varepsilon \in [0,1)$ справедлива оценка $Q_\varepsilon \leqslant \Delta_L^\ell \Gamma_L^\ell(\varepsilon)$.*

Доказательство вытекает непосредственно из предыдущей теоремы:

$$Q_\varepsilon = \sum_{m=0}^{L} Q_{\varepsilon,m} \leqslant \sum_{m=0}^{L} D_m H\tbinom{\ell}{L}{\bar{s}_m(\varepsilon)}{m} \leqslant \left(\sum_{m=0}^{L} D_m\right) \max_m H\tbinom{\ell}{L}{\bar{s}_m(\varepsilon)}{m} = \Delta_L^\ell \Gamma_L^\ell(\varepsilon). \tag{4.10}$$

Теорема доказана. ■

Оценка (4.10) отличается от (4.7) сомножителем Δ_L^ℓ, то есть надёжность предсказания может ухудшиться по сравнению с законом больших чисел во столько раз, сколько классов различимых алгоритмов содержит в себе множество A_L^ℓ. Как видно из доказательства, эта оценка может оказаться сильно завышенной.

В конкретной задаче целевая зависимость $y^*(x)$, обучающая выборка X^ℓ и метод обучения μ фиксированы. Поэтому в результате обучения могут быть получены только те алгоритмы семейства A, которые метод μ сочтёт подходящими для данной задачи. Остальные алгоритмы остаются незадействованными. Этот эффект будем называть *локализацией семейства алгоритмов*. Для обеспечения обучаемости не обязательно ограничивать сложность семейства. Достаточно применить метод обучения, способный подстраиваться под задачу, выделяя в семействе A соответствующее ей локальное подсемейство $A_L^\ell(\mu, X^L)$. Это свойство *локализующей способности* метода μ является важной составной частью его обобщающей способности.

Классическая оценка ТВЧ (4.4) получается, если к (4.10) применить операцию матожидания E_{X^L}, матожидание коэффициента разнообразия оценить сверху функцией роста, а гипергеометрическое распределение — экспоненциальной оценкой:

$$\begin{aligned}\mathsf{E}_{X^L} Q_\varepsilon(\mu, X^L) = \mathsf{P}_{X^\ell, X^k}\big[\delta(\mu X^\ell, X^\ell, X^k) \geqslant \varepsilon\big] = \mathsf{E}_{X^L}\, \mathsf{P}_n\big[\delta_n \geqslant \varepsilon\big] \leqslant \\ \leqslant \mathsf{E}_{X^L} \Delta_L^\ell(\mu, X^L) \cdot \Gamma_L^\ell(\varepsilon) \leqslant \Delta^A(2\ell) \cdot \tfrac{3}{2} e^{-\varepsilon^2 \ell},\end{aligned}$$

где последнее неравенство справедливо в предположении $\ell = k$.

С другой стороны, аналогично Теоремам 4.2 и 4.3, нетрудно доказать полный аналог оценки Вапника-Червоненкиса (4.4) в слабой аксиоматике:

$$\begin{aligned}P_\varepsilon(A) = \mathsf{P}_n\Big[\max_{a \in A} \delta(a, X_n^\ell, X_n^k) \geqslant \varepsilon\Big] \leqslant \Delta^A(L) \cdot \Gamma_L^\ell(\varepsilon) \leqslant \\ \leqslant \Delta^A(2\ell) \cdot \tfrac{3}{2} e^{-\varepsilon^2 \ell}. \qquad (4.11)\end{aligned}$$

Таким образом, верхние оценки функционалов $P_\varepsilon(A)$ и $Q_\varepsilon(\mu, X^L)$ совпадают, Однако функционал Q_ε точнее формализует понятие «обучаемости».

Отметим, что идея доказательства Теорем 4.2 и 4.3 в целом та же, что и в Теореме П2 из [9, стр. 221], но здесь доказательства «очищены от излишних вероятностных допущений». Именно в этих теоремах сосредоточена основная суть ТВЧ. Многие понятия и построения ТВЧ в слабой аксиоматике оказываются избыточными: семейство алгоритмов, равномерная сходимость частоты к вероятности, равномерная сходимость частот в двух подвыборках, «основная лемма» (4.2), необходимые условия равномерной сходимости.

Оценки (4.10) и (4.9) всё ещё завышены, и тому есть две причины.

Во-первых, коэффициент разнообразия не учитывает степень сходства алгоритмов. Два неразличимых алгоритма вносят в Δ_L^ℓ суммарный вклад, равный 1. Два алгоритма, различающихся на половине объектов выборки X^L, суммарно вносят 2. Два алгоритма, различающихся только на одном объекте, также вносят 2, хотя эта ситуация гораздо ближе к случаю неразличимых алгоритмов. В результате обучения по схожим подвыборкам X_n^ℓ, как правило, образуется много схожих алгоритмов. Каждый из них вносит в Δ_L^ℓ вклад, равный 1, что приводит к завышенности локального коэффициента разнообразия.

Во-вторых, коэффициент разнообразия не учитывает, что алгоритмы, получаемые в результате обучения, не равновероятны. Обозначим через N_{md} подмножество разбиений, при которых получаются алгоритмы из класса эквивалентности A_{md}:

$$N_{md} = \big\{n \in \{1, \ldots, N\} \;\big|\; a_n \in A_{md}\big\}$$

Если $|N_{md}| \gg 1$, то алгоритмы класса A_{md} будем называть *типичными*. Если мощность $|N_{md}|$ близка к 1, то алгоритмы класса A_{md} будем называть *нетипичными*. Среди них, скорее всего, окажутся алгоритмы относительно низкого качества, полученные при неслучайных разбиениях выборки, доля которых невелика и ограничена уровнем значимости η. Однако их число вполне может оказаться сравнимым с локальным коэффициентом разнообразия. Каждый нетипичный алгоритм вносит в Δ_L^ℓ вклад, равный 1, хотя такие алгоритмы можно было бы вообще не учитывать.

5. Методика эмпирического анализа завышенности

Основные причины завышенности оценок ТВЧ видны из доказательства Теорем 4.2 и 4.3. Слабая аксиоматика позволяет оценить вклад каждой из причин эмпирически, измерив значения функционалов Q_ε и $Q_{\varepsilon,m}$ с помощью скользящего контроля. Пусть $N' \subset \{1,\ldots,N\}$ — подмножество разбиений, $\hat{\mathsf{P}}_n \equiv \frac{1}{N'}\sum_{n\in N'}$ — эмпирическая оценка вероятности. Соответственно,

$$\hat{Q}_\varepsilon = \hat{\mathsf{P}}_n\big[\delta_n \geqslant \varepsilon\big];$$
$$\hat{Q}_{\varepsilon,m} = \hat{\mathsf{P}}_n\big[\delta_n \geqslant \varepsilon\big]\big[\nu(a_n, X^L) = m/L\big];$$

где $a_n = \mu X_n^\ell$ — алгоритм, построенный методом μ по обучающей подвыборке X_n^ℓ, $\delta_n = \nu(a_n, X_n^k) - \nu(a_n, X_n^\ell)$ — его переобученность.

Зададимся вопросом: какие значения должен принимать локальный профиль разнообразия, чтобы оценка (4.9) не была завышенной.

Опр. 5.1. *Эффективным локальным профилем разнообразия будем называть последовательность значений*

$$\hat{D}_m(\varepsilon) = \frac{\hat{Q}_{\varepsilon,m}}{H\binom{\ell\ s_m^-(\varepsilon)}{L\ m}}, \quad m = 0,\ldots,L.$$

Опр. 5.2. *Эффективным локальным коэффициентом разнообразия будем называть величину* $\hat{\Delta}_L^\ell(\varepsilon) = \hat{D}_0(\varepsilon) + \hat{D}_1(\varepsilon) + \cdots + \hat{D}_L(\varepsilon)$.

Это обратная задача — по известным оценкам надёжности обучения $\hat{Q}_{\varepsilon,m}$ и $\hat{Q}_\varepsilon$ оценивается профиль разнообразия D_m и коэффициент разнообразия Δ_L^ℓ. Разумеется, такие оценки не могут быть использованы в дальнейшем для решения основной (прямой) задачи. Цель их введения — выделить и сравнить различные факторы завышенности ТВЧ и показать, к каким значениям коэффициентов разнообразия нужно стремиться в теоретических исследованиях.

Эффективные коэффициенты разнообразия могут принимать нецелые значения. Кроме того, они зависят от параметра точности ε. Для обоснованного выбора ε сначала задаётся надёжность η_0 или диапазон значений надёжности $[\eta_1, \eta_2]$, в наших экспериментах 0.05 и $[0.01, 0.1]$. Точность и надёжность связаны невозрастающей зависимостью $\eta(\varepsilon) = Q_\varepsilon \approx \hat{Q}_\varepsilon$. Это позволяет вычислить соответствующее значение точности $\varepsilon = \eta^{-1}(\eta_0)$ или диапазон значений точности $[\varepsilon_1, \varepsilon_2] = [\eta^{-1}(\varepsilon_2), \eta^{-1}(\varepsilon_1)]$, который, в свою очередь, определяет диапазон коэффициента разнообразия:

$$\hat{\Delta}_L^\ell \in \Big[\min_{\varepsilon\in[\varepsilon_1,\varepsilon_2]} \hat{\Delta}_L^\ell(\varepsilon),\ \max_{\varepsilon\in[\varepsilon_1,\varepsilon_2]} \hat{\Delta}_L^\ell(\varepsilon)\Big].$$

Теперь рассмотрим основные факторы завышенности оценок ТВЧ и способы их эмпирического измерения. *Степенью завышенности* будем называть число, показывающее, во сколько раз завышена верхняя оценка.

1. Пренебрежение эффектом локализации. Локальный коэффициент разнообразия Δ_L^ℓ вполне может оказаться много меньше глобального коэффициента разнообразия $\Delta(A, X^L)$, и, тем более, функции роста $\Delta^A(L)$. Степень завышенности

$$r_1 = \frac{\Delta^A(L)}{\Delta_L^\ell}$$

может быть вычислена, если известны и теоретическая оценка функции роста $\Delta^A(L)$, и локальный коэффициент Δ_L^ℓ. Число разбиений $|N'|$ является тривиальной и, как правило, сильно заниженной оценкой Δ_L^ℓ.

2. Выделение коэффициента разнообразия в виде сомножителя. В доказательстве Теоремы 4.2 оценка сверху делается единственный раз с целью выделить коэффициент разнообразия D_m в виде сомножителя. Эффективный локальный профиль $\hat{D}_m(\varepsilon)$ определяет, какими должны были бы быть сомножители D_m, чтобы оценка не была завышенной. Степень завышенности определяется отношением

$$r_2(\varepsilon) = \frac{\Delta_L^\ell}{\hat{\Delta}_L^\ell(\varepsilon)}.$$

3. Свёртка профиля разнообразия $\{D_m\}_{m=0}^L$ в скалярную характеристику сложности — коэффициент разнообразия $\Delta_L^\ell = \sum_{m=0}^L D_m$. Этот шаг был сделан при доказательстве Теоремы 4.3. Степень завышенности определяется отношением

$$r_3(\varepsilon) = \frac{\sum_{m=0}^L \hat{D}_m(\varepsilon)\Gamma_L^\ell(\varepsilon)}{\sum_{m=0}^L \hat{D}_m(\varepsilon) H\binom{\ell\ \bar{s_m}(\varepsilon)}{L\ m}} = \frac{\hat{\Delta}_L^\ell(\varepsilon)\Gamma_L^\ell(\varepsilon)}{\hat{Q}_\varepsilon}.$$

4. Экспоненциальная аппроксимация функции гипергеометрического распределения оправдана только стремлением получить «более красивую» формулу. Когда все вычисления выполняются на компьютере, экспоненциальная оценка лишается практической целесообразности. Степень завышенности определяется отношением

$$r_4(\varepsilon) = \frac{\frac{3}{2}\, e^{-\varepsilon^2 \ell}}{\Gamma_L^\ell(\varepsilon)}.$$

Произведение введённых отношений равно степени завышенности оценки ТВЧ:

$$r_1 \cdot r_2(\varepsilon) \cdot r_3(\varepsilon) \cdot r_4(\varepsilon) = \frac{\Delta^A(L) \cdot \frac{3}{2}\, e^{-\varepsilon^2 \ell}}{\hat{Q}_\varepsilon}.$$

Эффективная ёмкость по Вапнику. Эффект локализации связан с фиксацией целевой зависимости y^*, метода обучения μ и выборки X^L. В работах [11, 12] вводится понятие *эффективной ёмкости* (effective VC-dimension), которая учитывает μ и X^L, но не учитывает y^*. Следовательно, отношение эффективной функции роста к эффективному локальному коэффициенту разнообразия даёт оценку той доли завышенности r_1, которая связана только с целевой зависимостью y^*.

Следуя [11], ограничимся задачей классификации с двумя классами, $\mathbb{Y} = \{0, 1\}$, при функции потерь $\mathscr{L}(y, y') = [y \neq y']$ и при $\ell = k$.

Эффективная функция роста определяется как значение $\Delta^A(L)$, при котором оценка (4.11) становится точной (не завышенной):

$$\hat{\Delta}^A_{\text{eff}}(L, \varepsilon) = \frac{1}{\Gamma^\ell_L(\varepsilon)} \hat{\mathsf{P}}_n \Big[\max_{a \in A} \delta(a, X^\ell_n, X^k_n) \geqslant \varepsilon \Big].$$

Эффективная ёмкость h определяется как величина, связанная с эффективной функцией роста формулой (4.5): $\hat{\Delta}^A_{\text{eff}}(L) = \frac{3}{2}\frac{L^h}{h!}$. Для её измерения в [11] предлагается оценивать $\hat{\Delta}^A_{\text{eff}}$ при различных L и подбирать такое значение h, при котором зависимость $\hat{\Delta}^A_{\text{eff}}$ от L наиболее точно аппроксимируется функцией $\frac{3}{2}\frac{L^h}{h!}$. Достаточно высокая точность аппроксимации свидетельствует о корректности методики.

Поиск алгоритма $\tilde{a}_n \in A$, на котором достигается максимум $\delta(a, X^\ell_n, X^k_n)$, эквивалентен минимизации эмпирического риска $\nu(a, \tilde{X}^L_n)$ по модифицированной полной выборке $\tilde{X}^L_n$. Модификация заключается в том, что на всех объектах $x_i \in X^k_n$ правильный ответ y_i заменяется ошибочным ответом $1 - y_i$:

$$\tilde{a}_n = \arg\max_{a \in A} \delta(a, X^\ell_n, X^k_n) = \arg\min_{a \in A} \left(\frac{1}{\ell} \sum_{x_i \in X^\ell_n} \big[a(x_i) \neq y_i\big] + \frac{1}{k} \sum_{x_i \in X^k_n} \big[a(x_i) = y_i\big] \right),$$

Для получения алгоритма $\tilde{a}_n$ тот же метод обучения μ применяется к модифицированной полной выборке $\tilde{X}^L_n$. Алгоритм фактически обучается делать ошибки на случайной половине объектов. Тем самым устраняется фиксация целевой зависимости y^* и связанная с ней часть эффекта локализации.

Степень завышенности, связанная с игнорированием целевой зависимости y^*:

$$r'_1(\varepsilon) = \frac{\hat{\mathsf{P}}_n \big[\delta(\tilde{a}_n, X^\ell_n, X^k_n) \geqslant \varepsilon\big]}{\hat{\mathsf{P}}_n \big[\delta(a_n, X^\ell_n, X^k_n) \geqslant \varepsilon\big]} = r_3(\varepsilon) \frac{\hat{\Delta}^A_{\text{eff}}(L, \varepsilon)}{\hat{\Delta}^\ell_L(\varepsilon)}.$$

Приведём две интерпретации коэффициента $r'_1(\varepsilon)$.

1. Эксперименты с линейным пороговым классификатором, описанные в [11], дали вполне ожидаемый результат: эффективная ёмкость приблизительно равна размерности подпространства, в котором сосредоточены объекты выборки. Коэффициент $r'_1(\varepsilon)$ показывает, во сколько раз завышена эта оценка.

2. *Эффективная функция роста* определяется через функционал равномерной сходимости P_ε, который сам является заведомо завышенной оценкой. *Эффективный локальный коэффициент разнообразия* определяется через функционал полного скользящего контроля Q_ε, который более точно формализует понятие обучаемости. Отсюда вторя интерпретация: $r'_1(\varepsilon)$ — это степень завышенности, возникающая в результате применения принципа равномерной сходимости.

6. Обобщение на случай закономерностей

Логические алгоритмы классификации, основанные на *индукции правил* (rule induction), особенно удобны для проведения экспериментов по эмпирическому измерению степени завышенности. Во-первых, для них легко выписываются оценки функции роста. Во-вторых, они основаны на явном переборе большого количества элементарных классификаторов (правил), что позволяет эффективно оценивать локальные коэффициенты разнообразия. В-третьих, эти алгоритмы широко применяются на практике, поэтому проблема переобученности как самого алгоритма, так и составляющих его правил, представляет значительный практический интерес.

Рассматриваются задачи классификации, $\mathbb{Y}$ — конечное множество.

Говорят, что предикат $\varphi\colon \mathbb{X} \to \{0,1\}$ *выделяет* (covers) объект x, если $\varphi(x) = 1$. Предикат φ характеризуется относительно класса $y \in \mathbb{Y}$ и выборки X^ℓ двумя величинами: числом положительных примеров p_y (выделяемых объектов класса y) и числом отрицательных примеров b_y (выделяемых объектов других классов):

$$p_y(\varphi_y, X^\ell) = \#\big\{x_i \in X^\ell \;\big|\; \varphi_y(x_i) = 1,\ y_i = y\big\};$$
$$b_y(\varphi_y, X^\ell) = \#\big\{x_i \in X^\ell \;\big|\; \varphi_y(x_i) = 1,\ y_i \neq y\big\};$$

Закономерностью или правилом (rule) класса $y \in \mathbb{Y}$ называется предикат $\varphi_y\colon X \to \{0,1\}$, выделяющий достаточно много объектов класса y и достаточно мало объектов всех остальных классов: $p_y(\varphi_y, X^\ell) \geqslant p_{y0},\ \ b_y(\varphi_y, X^\ell) \leqslant b_{y0}$, где p_{y0} и b_{y0} — заданные пороговые константы.

Качество закономерности можно характеризовать не только парой показателей (p_y, b_y), но и одним показателем информативности $I(p_y, b_y)$. На практике его вводят по-разному, в частности, используется энтропийный критерий *выигрыша информации* (information gain), статистические критерии ξ^2, ω^2, точный тест Фишера [13], легко вычисляемый критерий $I(p_y, b_y) = \sqrt{p_y} - \sqrt{b_y}$ [14], и другие.

Логический алгоритм есть линейная композиция закономерностей:

$$a(x) = \arg\max_{y \in Y} \sum_{t=1}^{T_y} w_y^t \varphi_y^t(x),$$

где $\varphi_y^t(x)$ — закономерности класса y, w_y^t — веса закономерностей, T_y — число закономерностей класса y. В таком виде представимы многие логические алгоритмы: взвешенное голосование (weighted voting) правил [14], решающие списки (decision list) [15], решающие деревья (decision tree) [16], машины покрывающих множеств (set covering machine) [17], и другие.

Эмпирическое измерение степени завышенности удобно производить не для алгоритмов, а для закономерностей. Для этого придётся немного изменить основные определения. Модификации носят технический характер, так что формулировки основных теорем остаются практически теми же.

Методом обучения закономерностей класса y называется отображение μ_y, которое по обучающей выборке X^ℓ строит набор закономерностей:

$$\mu_y X^\ell = \big\{\varphi_y^t(x) \;\big|\; t = 1, \ldots, T_y\big\}.$$

Частота ошибок закономерности φ_y на выборке X^ℓ есть

$$\nu_y(\varphi_y, X^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell}\big[\varphi_y(x_i) \neq [y_i = y]\big] = \frac{1}{\ell}\big(b_y(\varphi_y, X^\ell) + P_y - p_y(\varphi_y, X^\ell)\big), \qquad (6.1)$$

где P_y — число объектов класса y в выборке X^ℓ. В случае $\varphi_y(x_i) = 0$ и $y_i = y$ закономерность допускает ошибку I рода — не выделяет объект своего класса. В случае $\varphi_y(x_i) = 1$ и $y_i \neq y$ закономерность допускает ошибку II рода — выделяет объект чужого класса. Ошибки I рода, как правило, менее опасны в логических алгоритмах, так как пропущенный объект может быть выделен другими закономерностями.

Переобученностью закономерности $\varphi \in \mu_y X^\ell$ при заданной контрольной выборке X^k называется разность частот её ошибок на контроле и на обучении:

$$\delta_y(\varphi, X^\ell, X^k) = \nu_y(\varphi, X^k) - \nu_y(\varphi, X^\ell).$$

Функционал полного скользящего контроля $Q_\varepsilon(\mu_y, X^L)$ определяется как доля переобученных закономерностей среди всех закономерностей класса y, построенных методом μ_y по всевозможным подвыборкам $X^\ell_n \subset X^L$:

$$Q_\varepsilon(\mu_y, X^L) = \mathsf{P}_n \frac{1}{|\mu_y X^\ell_n|}\sum_{\varphi\in\mu_y X^\ell_n}\big[\delta_y(\varphi, X^\ell_n, X^k_n) \geqslant \varepsilon\big].$$

Предикаты $\varphi, \varphi' \colon X \to \{0, 1\}$ называются *неразличимыми* или эквивалентными на выборке X^L, если $\varphi(x) = \varphi'(x)$ для всех $x \in X^L$. *Коэффициентом разнообразия* $\Delta(\Phi, X^L)$ множества предикатов Φ на выборке X^L называется число классов эквивалентности, индуцируемых на Φ отношением неразличимости. Рассмотрим множество закономерностей, получаемых методом μ_y по всевозможным обучающим подвыборкам: $\Phi^\ell_L = \bigcup_{n=1}^N \mu_y X^\ell_n$. Его коэффициент разнообразия $\Delta^\ell_L = \Delta(\Phi^\ell_L, X^L)$ назовём *локальным коэффициентом разнообразия* метода μ_y на выборке X^L. Множество закономерностей Φ^ℓ_L разбивается на $L+1$ подмножеств Φ_m, состоящих из закономерностей с фиксированным числом ошибок m на полной выборке X^L:

$$\Phi_m = \big\{\varphi \in \Phi^\ell_L \;\big|\; \nu_y(\varphi, X^L) = \tfrac{m}{L}\big\},\ m = 0, \ldots, L.$$

Локальный профиль разнообразия метода μ_y на выборке X^L есть последовательность коэффициентов разнообразия $D_m = \Delta(\Phi_m, X^L),\ m = 0, \ldots, L$.

Очевидно, что $\Delta^\ell_L = D_0 + \cdots + D_L$.

Наряду с функционалом Q_ε определим функционал $Q_{\varepsilon,m}$ как долю переобученных закономерностей, допускающих m ошибок на X^L:

$$Q_{\varepsilon,m}(\mu_y, X^L) = \mathsf{P}_n \frac{1}{|\mu_y X^\ell_n|}\sum_{\varphi\in\mu_y X^\ell_n}\big[\delta_y(\varphi, X^\ell_n, X^k_n) \geqslant \varepsilon\big]\big[\nu_y(\varphi, X^L) = \tfrac{m}{L}\big].$$

В этих обозначениях теоремы 4.2 и 4.3 остаются непосредственно справедливы для случая закономерностей. Модификация коснулась, главным образом, определения локального множества алгоритмов A^ℓ_L — теперь его роль играет локальное множество закономерностей Φ^ℓ_L. Смысл модификации предельно прост: надо учитывать все закономерности $\varphi^t_y(x)$, построенные при всех разбиениях n. Методика эмпирического измерения величин Q_ε, $Q_{\varepsilon,m}$, $\hat{\Delta}^\ell_L$, $\hat{D}_m$ остаётся прежней.

Отметим, что все эти величины определяются для каждого класса $y \in \mathbb{Y}$ в отдельности, и, вообще говоря, могут сильно отличаться для разных классов.

Предложенная методика существенно уточняет более ранние варианты [18, 19].

Алгоритм 6.1. Обучение конъюнкций методом усечённого поиска в ширину.

Вход:
X^ℓ — обучающая выборка; $y \in \mathbb{Y}$ — класс, для которого строятся конъюнкции; K — максимальный ранг конъюнкций; T_1 — число лучших конъюнкций, отбираемых на каждом шаге; T_0 — число лучших конъюнкций, отбираемых на последнем шаге; $I_{\min}$ — порог информативности; $E_{\max}$ — допустимая доля ошибок;

Выход:
список конъюнкций $R_y = \{\varphi_y^t(x) | \, t = 1, \ldots, T_y\}$;

$R_y := \varnothing$;
для всех $\beta \in \mathscr{B}_j$, $j = 1, \ldots, n$
  `Добавить_в_список` (R_y, β);
для всех $k = 2, \ldots, K$
  для всех конъюнкций $\varphi \in R_y$ ранга $(k-1)$
    для всех $\beta \in \mathscr{B}_j$, $j = 1, \ldots, n$
      если признака f_j нет в конъюнкции φ и $I_y(\varphi \wedge \beta) \geqslant I_{\min}$ **то**
        `Добавить_в_список` $(R_y, \varphi \wedge \beta)$;
оставить в R_y не более T_0 конъюнкций с наибольшими $I_y(\varphi)$ и $E_y(\varphi) \leqslant E_{\max}$;

ПРОЦЕДУРА `Добавить_в_список` (R_y, φ);
если $|R_y| < T_1$ **то**
  $R_y := R_y \cup \{\varphi\}$
иначе
  найти худшую конъюнкцию в списке: $\psi := \arg\min\limits_{\psi \in R_y} I_y(\psi)$;
  если $I_y(\varphi) > I_y(\psi)$ **то**
    заменить в списке R_y худшую конъюнкцию ψ на φ;

Синтез логических закономерностей. Для экспериментов использовался метод обучения закономерностей, основанный на применении усечённого поиска в ширину [20], бустинга закономерностей [14], и точного теста Фишера в роли критерия информативности [13]. Алгоритм реализован Д. Кочедыковым и А. Ивахненко и применяется в системе кредитного скоринга Forecsys ScoringAce® [21, 18, 19]. Здесь приводится его упрощённое описание.

Пусть объекты $x \in \mathbb{X}$ описываются n дискретными признаками $f_j \colon \mathbb{X} \to D_j$, $j = 1, \ldots, n$. Номинальные признаки порождают *элементарные предикаты (термы)* двух видов: $\beta_j(x) = [f_j(x) = c]$ и $\beta_j(x) = [f_j(x) \neq c]$ при всевозможных $c \in D_j$. Порядковые признаки, в дополнение к этим двум, порождают ещё два вида термов: $\beta_j(x) = [f_j(x) \leqslant c]$ и $\beta_j(x) = [f_j(x) \geqslant c]$, $c \in D_j$. Обозначим через $\mathscr{B}_j$ множество всех термов, порождаемых признаком f_j. Поиск закономерностей производится среди конъюнкций ранга не выше K, составленных из термов:

$$\Phi[K] = \Big\{ \varphi(x) = \bigwedge_{j \in J} \beta_j(x) \;\Big|\; \beta_j \in \mathscr{B}_j,\ J \subseteq \{1, \ldots, n\},\ |J| \leqslant K \Big\}.$$

Алгоритм 6.1 начинает поиск закономерностей с построения конъюнкций ранга 1. Для этого отбираются не более T_1 самых информативных термов. На всех последующих шагах к каждой из имеющихся конъюнкций добавляется один терм всеми

возможными способами. Получается расширенное множество конъюнкций, из которых снова отбираются T_1 самых информативных. Наращивание конъюнкций прекращается либо при достижении максимального ранга K, либо когда ни одну из конъюнкций не удаётся улучшить путём добавления терма. Лучшие конъюнкции, собранные со всех шагов, заносятся в списки R_y. Параметр T_1 позволяет управлять *шириной поиска* и находить компромисс между качеством и скоростью работы алгоритма.

Качество предиката $\varphi(x)$ относительно обучающей выборки X^ℓ и класса y оценивается двумя критериями: долей ошибочно выделенных объектов $E_y(\varphi) = \frac{b_y}{p_y+b_y}$ и информативностью $I_y(\varphi) = \ln C_{P_y+B_y}^{p_y+b_y} - \ln C_{P_y}^{p_y} C_{B_y}^{b_y}$, где P_y — число объектов класса y в выборке X^ℓ, B_y — число объектов всех остальных классов в X^ℓ.

После выполнения Алгоритма 6.1 в выборке могут остаться объекты, не выделенные ни одной закономерностью из списков R_y, либо ошибочно выделенные закономерностями «чужих» классов. Этим объектам назначаются бóльшие веса согласно формулам бустинга [14] и Алгоритм 6.1 запускается заново. Веса объектов учитываются при вычислении критерия $I_y(\varphi)$, что позволяет находить новые закономерности, существенно отличающиеся от найденных на предыдущих итерациях.

Функция роста $\Delta^{\Phi[K]}(L)$ множества $\Phi[K]$ не превосходит его мощности. Пусть j-й признак порождает $d_j = |\mathscr{B}_j|$ термов, $j = 1, \ldots, n$. Тогда число конъюнкций ранга r, построенных из признаков подмножества $J = \{1, \ldots, j\}$, не превосходит

$$H_{r,j} = \sum_{\substack{J' \subseteq J \\ |J'|=r}} \prod_{j \in J'} d_j.$$

Возможно быстрое вычисление чисел $H_{r,j}$ за $O(Kn)$ операций, если воспользоваться рекуррентными формулами: $H_{0,j} = 1$, $H_{r,j} = 0$ при $r > j$ и

$$H_{r,j+1} = H_{r,j} + d_j H_{r-1,j}, \quad j = 1, \ldots, n, \quad r = 1, \ldots, K.$$

Функция роста не превосходит суммарного числа конъюнкций ранга от 1 до K:

$$\Delta^{\Phi[K]}(L) \leqslant H_{1,n} + \cdots + H_{K,n}.$$

Локальный коэффициент разнообразия Δ_L^ℓ оценивается суммарным числом конъюнкций, попавших в списки R_y по всем обучающим выборкам X_n^ℓ, $n \in N'$:

$$\underline{\Delta}_L^\ell = \sum_{n \in N'} |\mu_y X_n^\ell| \leqslant |N'| \cdot T_0.$$

Эта оценка может оказаться сильно заниженной, поскольку $|N'| \ll N$. Более адекватной оценкой является число $\underline{\Delta}_L^\ell$ всех проанализированных конъюнкций, удовлетворяющих критериям высокой информативности $I_y(\varphi) \geqslant I_{\min}$ и низкой доли ошибок $E_y(\varphi) \leqslant E_{\max}$. Его легко вычислить в ходе перебора.

Обозначим через $\overline{\Delta}_L^\ell$ число всех конъюнкций φ, для которых в ходе перебора вычисляются характеристики $p_y(\varphi, X^\ell)$ и $b_y(\varphi, X^\ell)$. Тривиальная и несколько завышенная оценка $\overline{\Delta}_L^\ell \leqslant |N'|(T_1 K - T_1 + 1)(d_1 + \cdots + d_n)$. Точное число всех проанализированных конъюнкций также легко вычисляется в ходе перебора.

Задача	L	n	$d_1 \cdots d_n$	C4.5	C5.0	RIPPER	SLIPPER	Forecsys
crx	690	15	$2^4\,3^2\,4^1\,9^1\,14^1\,20^6$	15.5	14.0	15.2	15.7	14.3 ± 0.2
german	1000	20	$2^2\,3^3\,4^3\,5^5\,10^1\,11^1\,20^5$	27.0	28.3	28.7	27.2	28.5 ± 1.0
hepatitis	155	19	$2^{13}\,6^4\,8^1\,9^1$	18.8	20.1	23.2	17.4	16.7 ± 1.7
horse-colic	300	25	$2^3\,3^2\,4^6\,5^5\,6^2\,20^7$	16.0	15.3	16.3	15.0	16.4 ± 0.5
hypothyroid	3163	25	$2^{18}\,20^7$	0.4	0.4	0.9	0.7	0.8 ± 0.04
liver	345	6	$12^1\,20^5$	37.5	31.9	31.3	32.2	29.2 ± 1.6
promoters	106	57	57^4	18.1	22.7	19.0	18.9	12.0 ± 2.0

Таблица 1. Характеристики задач: длина выборки L; число признаков n; число порождаемых термов d_j, где запись 20^5 означает, что имеется 5 признаков, порождающих по 20 термов; процент ошибок на контроле для четырёх стандартных алгоритмов из [22, 14]; процент ошибок на контроле для Алгоритма 6.1.

Для Алгоритма 6.1 имеется ещё один способ оценить степень завышенности, связанную с локализацией целевой зависимости y^*. Это отношение числа всех проанализированных конъюнкций к числу конъюнкций, оказавшихся закономерностями:

$$r_1''(\varepsilon) = \frac{\overline{\Delta}_L^\ell}{\underline{\Delta}_L^\ell}.$$

Поскольку Алгоритм 6.1 ведёт направленный поиск наиболее информативных конъюнкций, это отношение может оказаться несколько заниженным.

7. Эксперименты, результаты, выводы

Алгоритм тестировался на 7 реальных задачах классификации из репозитория UCI [23]. Число классов во всех задачах равнялось двум. Выборка разбивалась 20 раз случайным образом на две равные части, $\ell = k$, со стратификацией классов. При каждом разбиении сначала первая половина выборки была обучающей, вторая — контрольной, затем они менялись ролями. Таким образом, $|N'| = 40$. В Таблице 1 показаны характеристики задач и средний процент ошибок на контрольных данных. Данные по алгоритмам C4.5, C5.0, RIPPER и SLIPPER взяты из работ [22, 14] и показывают, что качество реализованного алгоритма сопоставимо с аналогами (доказательство его превосходства не является целью данной работы).

В Таблице 2 показаны оценки коэффициентов разнообразия, вычисленные в процессе работы Алгоритма 6.1. В двух правых столбцах приведены оценки эффективного локального коэффициента разнообразия, вычисленные по Определению 5.2.

На Рис. 1 показаны графики зависимости коэффициента $\hat{\Delta}_L^\ell$ от точности ε. Убывающая кривая показывает зависимость надёжности $\hat{Q}_\varepsilon$ от ε. Для определения диапазона возможных значений $\hat{\Delta}_L^\ell(\varepsilon)$ сначала фиксируется диапазон «разумных» значений надёжности $\hat{Q}_\varepsilon \in [0.01, 0.1]$ (по правой вертикальной оси), для него определяется диапазон точности (по горизонтальной оси), на котором определяется минимальное и максимальное значение $\hat{\Delta}_L^\ell(\varepsilon)$.

В Таблице 3 показаны оценки степеней завышенности, вычисленные при фиксированном значении надёжности $\hat{Q}_\varepsilon = 0.05$.

Задача	T_1	K	y	$\lvert\Phi[K]\rvert$	$\frac{1}{\lvert N'\rvert}\overline{\Delta}^{\ell}_{L}$	$\frac{1}{\lvert N'\rvert}\underline{\Delta}^{\ell}_{L}$	$\frac{1}{\lvert N'\rvert}\underline{\underline{\Delta}}^{\ell}_{L}$	$\hat{\Delta}^{\ell}_{L}[\varepsilon_1,\varepsilon_2]$	$\hat{\Delta}^{\ell}_{L}(\varepsilon_0)$
crx	50	4	0	$1.4\cdot10^{7}$	$2.1\cdot10^{4}$	380	5	[10; 41]	24
			1			490	6	[11; 180]	12
german	50	5	1	$5.2\cdot10^{8}$	$3.0\cdot10^{4}$	1370	14	[38; 530]	54
			2			330	3	[1.0; 2.2]	1.9
hepatitis	50	4	0	$5.6\cdot10^{5}$	$0.9\cdot10^{4}$	570	7	[11; 148]	83
			1			240	3	[12; 27]	15
horse-colic	50	5	1	$1.9\cdot10^{6}$	$3.8\cdot10^{4}$	630	7	[2; 9]	7
			2			330	3	[3; 6]	6
hypothyroid	100	5	0	$5.3\cdot10^{8}$	$6.3\cdot10^{4}$	210	7	[3; 220]	21
			1			80	3	[2; 44]	30
liver	50	4	0	$1.9\cdot10^{6}$	$1.1\cdot10^{4}$	700	7	[4; 21]	12
			1			650	7	[3; 12]	5
promoters	50	3	0	$1.0\cdot10^{8}$	$2.2\cdot10^{4}$	480	5	[36; 230]	72
			1			300	3	[9; 22]	18

Таблица 2. Параметры метода: ширина поиска T_1; максимальный ранг конъюнкций K; номер класса в кодировке UCI. Оценки коэффициентов разнообразия: функция роста $|\Phi[K]|$; среднее число проанализированных конъюнкций $\frac{1}{|N'|}\overline{\Delta}^{\ell}_{L}$; среднее число информативных конъюнкций $\frac{1}{|N'|}\underline{\Delta}^{\ell}_{L}$; среднее число конъюнкций, отобранных в алгоритм $\frac{1}{|N'|}\underline{\underline{\Delta}}^{\ell}_{L}$; эффективный локальный коэффициент разнообразия $\hat{\Delta}^{\ell}_{L}(\varepsilon)$, соответствующий диапазону $\hat{Q}_\varepsilon \in [0.01, 0.1]$ и значению $\hat{Q}_\varepsilon = 0.05$.

Интерпретации и выводы. Среди четырёх факторов завышенности первые два оказались наиболее значимыми: r_1 — пренебрежение эффектом локализации и r_2 — выделение коэффициента разнообразия в виде отдельного сомножителя. На устранение первого фактора направлено большинство работ по *оценкам, зависящим от выборки* (data-dependent bounds) [24, 25, 6]. Однако все эти оценки имеют сомножитель, описывающий «сложность» некоторого множества алгоритмов, пусть даже и локального. Большие значения r_2 говорят о том, что «проклятие завышенности» присуще всем сложностным оценкам.

Коэффициенты r'_1 и r''_1 оценивают влияние целевой зависимости y^* на степень завышенности r_1. Оба они занижены, поэтому можно утверждать только, что соответствующая потеря точности составляет два порядка или более. Введённое Вапником понятие *эффективной ёмкости* не учитывает этот фактор, поскольку основывается на принципе равномерной сходимости.

Эффективный локальный коэффициент разнообразия во всех задачах не превосходит длины выборки L. Попытка использовать его для определения *эффективной локальной ёмкости* по формуле (4.5) приводит к вырожденному результату: такая ёмкость на практике не превышает единицу. Отсюда снова следует вывод, что оценки качества обучения, зависящие от коэффициентов разнообразия (даже локальных), по природе своей чрезвычайно завышены. Для обоснования обучаемости необходимо вводить какие-то другие, гораздо более тонкие, характеристики метода обучения.

Фактор r_3 в большинстве случаев сравнительно мал. При значениях числа ошибок $m = L\nu_y(\varphi, X^L)$, характерных для закономерностей, значения функции $H(m) = H\binom{\ell}{L}{m}{s^-_m(\varepsilon)}$ не сильно отличаются от максимума, см. Рис. 2. Однако при $m \to 0$ функция $H(m)$ стремится к нулю быстрее геометрической прогрессии. Поэто-

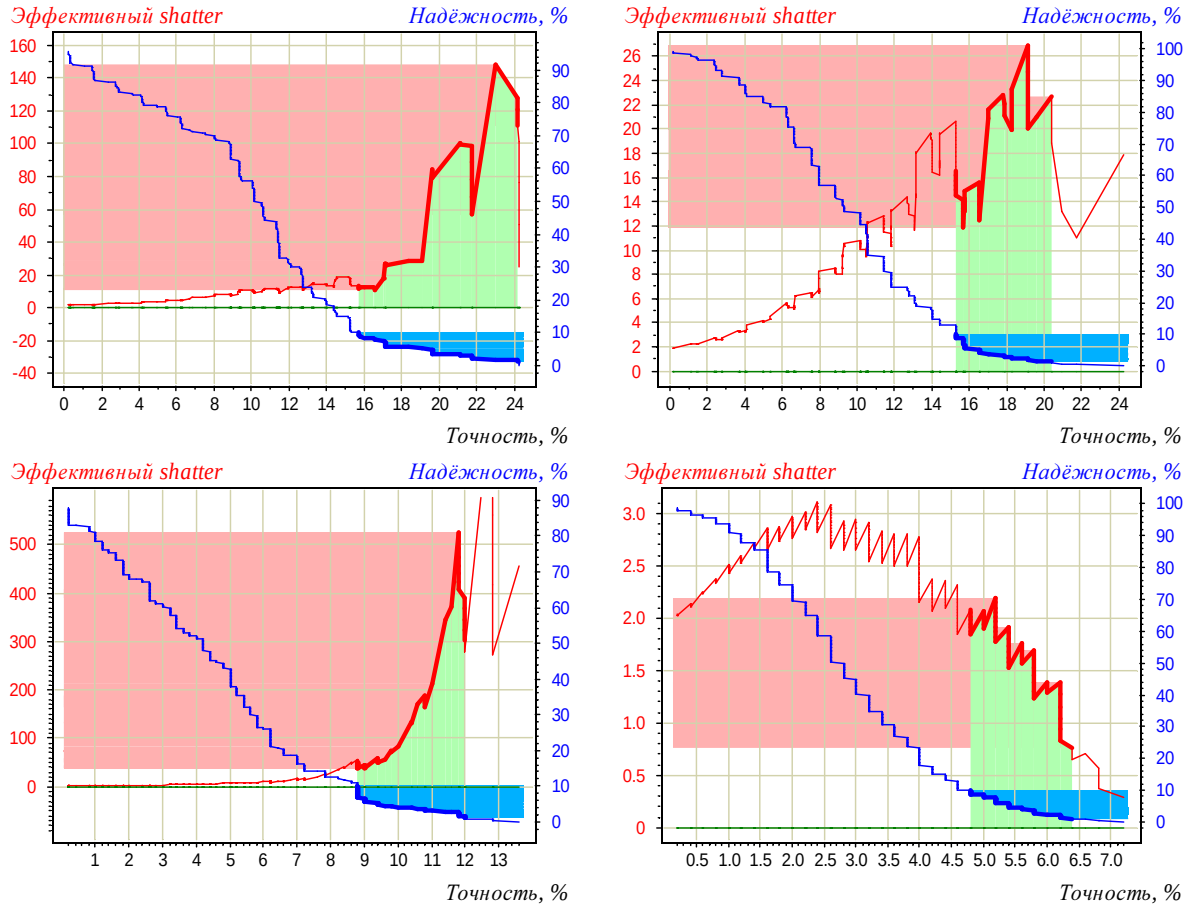

Рис. 1. Зависимость эффективного локального коэффициента разнообразия $\hat{\Delta}_L^\ell$ и надёжности $\hat{Q}_\varepsilon$ от точности ε для задач hepatitis (вверху, $y = 0, 1$), german (внизу, $y = 1, 2$). Полосы показывают определение диапазона возможных значений $\hat{\Delta}_L^\ell$ по заданному диапазону надёжности $\hat{Q}_\varepsilon \in [0.01, 0.1]$.

му для обычных алгоритмов классификации и «хороших» задач с частотой ошибок (ориентировочно) менее 10% фактор r_3 может оказаться значительным.

Фактор r_4 показывает, что экспоненциальная аппроксимация гипергеометрического распределения неточна, и на практике от неё следует отказаться.

Основные выводы и направления дальнейших исследований:

— В слабой вероятностной аксиоматике оценки обобщающей способности выводятся для функционалов, построенных по принципу полного скользящего контроля. Это упрощает эмпирический анализ теоретических оценок, позволяет контролировать точность оценок, разделять и анализировать причины их завышенности.

— Представляет интерес применение предложенной эмпирической методики к другим моделям алгоритмов и методам обучения с целью изучения их *локализующей способности*.

— При получении численно точных оценок обобщающей способности необходимо учитывать не только эффект локализации, но также неравномерность распределения и степень различности алгоритмов. В сложностных оценках необходимо ориентироваться на получение коэффициентов разнообразия порядка 10^1–10^2.

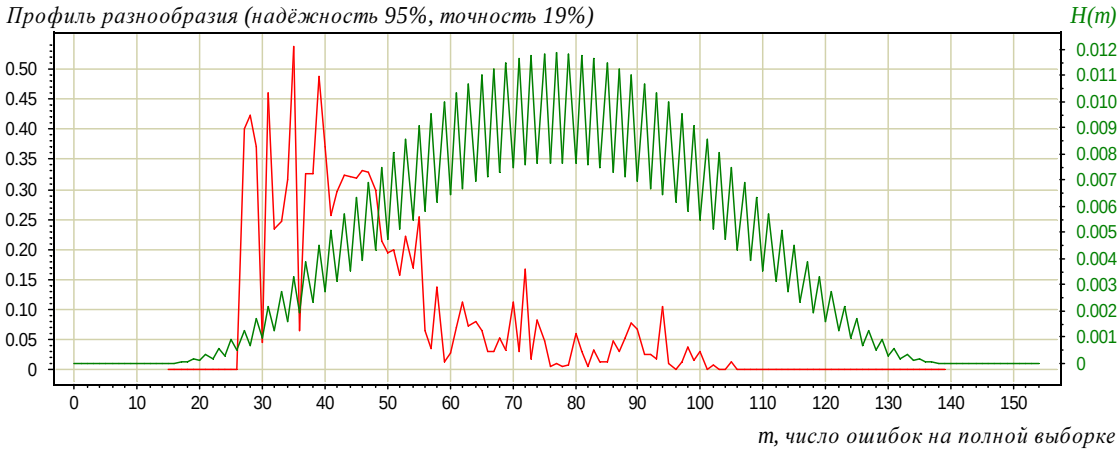

Рис. 2. Зависимость эффективного локального профиля $\hat{D}_m(\varepsilon)$ и функции $H(m)$ от числа ошибок $m = L\nu_y(\varphi, X^L)$, для задачи hepatitis, $y = 0$.

Задача	y	r_1	$r'_1(\varepsilon)$	$r''_1(\varepsilon)$	$r_2(\varepsilon)$	$r_3(\varepsilon)$	$r_4(\varepsilon)$
crx	0	890	20	55	680	3.1	32.6
	1	690	21	43	1700	1.6	11.6
german	1	8 950	18	22	1500	1.7	10.9
	2	37 000	22	92	9000	1.2	9.9
hepatitis	0	23	20	16	280	13.4	9.5
	1	55	20	37	680	2.4	22.5
horse-colic	1	72	19	60	4500	2.1	7.2
	2	140	20	115	3400	3.6	7.3
hypothyroid	0	61 000	21	310	400	32.2	16.5
	1	153 000	15	770	460	3.8	28.7
promoters	0	94	16	46	340	5.9	9.8
	1	150	23	73	790	3.4	6.9

Таблица 3. Степени завышенности при значении точности ε, соответствующей надёжности $\hat{Q}_\varepsilon = 0.05$.

Работа выполнена при поддержке РФФИ, проекты №05-01-00877, №08-07-00422 программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики» и Фонда содействия отечественной науке.

Литература

[1] *Колмогоров А. Н.* Теория информации и теория алгоритмов / Под ред. Ю. В. Прохоров. — М.: Наука, 1987.

[2] *Беляев Ю. К.* Вероятностные методы выборочного контроля. — М.: Наука, 1975.

[3] *Смирнов Н. В.* Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках // *Бюлл. Московского ун-та, серия А.* — 1939. — № 2. — С. 3–14.

[4] *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. — М.: Наука, 1983.

[5] *Vapnik V.* Statistical Learning Theory. — Wiley, New York, 1998.

[6] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances // *ESAIM: Probability and Statistics.* — 2005. — no. 9. — Pp. 323–375.

[7] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / Под ред. О. Б. Лупанов. — М.: Физматлит, 2004. — Т. 13. — С. 5–36.

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

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

[10] *Вапник В. Н., Червоненкис А. Я.* О равномерной сходимости частот появления событий к их вероятностям // *Теория вероятностей и ее применения.* — 1971. — Т. 16, № 2.

[11] *Vapnik V., Levin E., Cun Y. L.* Measuring the VC-dimension of a learning machine // *Neural Computation.* — 1994. — Vol. 6, no. 5. — Pp. 851–876.

[12] *Bottou L., Cortes C., Vapnik V.* On the effective VC dimension. — 1994.

[13] *Martin J. K.* An exact probability metric for decision tree splitting and stopping // *Machine Learning.* — 1997. — Vol. 28, no. 2-3. — Pp. 257–291.

[14] *Cohen W. W., Singer Y.* A simple, fast and effective rule learner // Proc. of the 16 National Conference on Artificial Intelligence. — 1999. — Pp. 335–342.

[15] *Rivest R. L.* Learning decision lists // *Machine Learning.* — 1987. — Vol. 2, no. 3. — Pp. 229–246.

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

[17] *Marchand M., Shawe-Taylor J.* Learning with the set covering machine // Proc. 18th International Conf. on Machine Learning. — Morgan Kaufmann, San Francisco, CA, 2001. — Pp. 345–352.

[18] *Воронцов К. В., Ивахненко А. А.* Эмпирические оценки локальной функции роста в задачах поиска логических закономерностей // *Искусственный Интеллект.* — 2006. — С. 281–284.

[19] *Ивахненко А. А., Воронцов К. В.* Верхние оценки переобученности и профили разнообразия логических закономерностей // Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 33–37.

[20] *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных. — Новосибирск: Наука, 1981.

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

[22] *Cohen W. W.* Fast effective rule induction // Proc. of the 12th International Conference on Machine Learning, Tahoe City, CA. — Morgan Kaufmann, 1995. — Pp. 115–123.

[23] *Asuncion A., Newman D.* UCI machine learning repository: Tech. rep.: University of California, Irvine, School of Information and Computer Sciences, 2007.

[24] *Koltchinskii V., Panchenko D.* Rademacher processes and bounding the risk of function learning // High Dimensional Probability, II / Ed. by D. E. Gine, J.Wellner. — Birkhauser, 1999. — Pp. 443–457.

[25] *Bartlett P. L., Mendelson S., Philips P.* Local complexities for empirical risk minimization // COLT: 17th Annual Conference on Learning Theory / Ed. by J. Shawe-Taylor, Y. Singer. — Springer-Verlag, 2004. — Pp. 270–284.