Измерение локальной эффективной функции роста в задачах поиска логических закономерностей

К. В. Воронцов

voron@ccas.ru

http://www.ccas.ru/voron

Москва
Вычислительный Центр РАН

Проблема оценивания обобщающей способности является, пожалуй, наиболее интригующей в машинном обучении. Окончательного решения её до сих пор не найдено, несмотря на значительные усилия многих учёных на протяжении последних 40 лет.

Ставшая уже классической теория Вапника-Червоненкиса даёт слишком завышенные оценки, не способные объяснить некоторые эмпирические факты. Например, почему алгоритм бустинга строит взвешенное голосование из сотен и даже тысяч решающих деревьев, а качество классификации независимых тестовых данных только улучшается с ростом сложности композиции? Причём во многих задачах улучшается, а в некоторых нет.

Мой доклад преследует цель прояснить взаимосвязи между качеством обучения и сложностью. А также представить новый экспериментальный материал, позволяющий сделать несколько интересных выводов о причинах завышенности классических оценок и о причинах высокой эффективности логических алгоритмов классификации.

Задача обучения по прецедентам

- Восстановление зависимости $y^* : X \to Y$
- Выборка $X^l = \{x_1, ..., x_l\}$ с известными ответами $y^*(x_i)$
- *Метод обучения* — отображение $\mu : X^l \mapsto a$, $a \in A$,
  где $A = \{a : X \to Y\}$ — заданное семейство алгоритмов
- Частота ошибок алгоритма a на выборке X^l:

  $\nu(a, X^l) = \frac{1}{l}\sum_{i=1}^{l} I(a(x_i), y_i^*)$, где $I(y, y^*)$ — функция потерь.

Проблема:

Оценить *обобщающую способность* $\nu\left(\mu(X^l), X^k\right)$,

где X^k — произвольная (неизвестная) выборка.

Введем следующие стандартные обозначения.

Пусть имеется множество объектов X, множество допустимых ответов Y и требуется восстановить неизвестную зависимость y^* по конечной обучающей выборке X^l.

В отличие от теории Вапника-Червоненкиса, мы вводим понятие метода обучения. Это некое отображение, которое по выборке строит алгоритм.

Частота ошибок алгоритма a на выборке X^l определяется стандартным образом.

Основной вопрос, который нас интересует — какой будет частота ошибок обученного алгоритма на новой выборке X^k, неизвестной в момент обучения.

Функционалы обобщающей способности

Статистическая теория Вапника-Червоненкиса:

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

Комбинаторная теория:

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

где $a_n = \mu(X_n^l)$,

$X^L = X_n^l \bigcup X_n^k$, $n = \overline{1, N}$ — всевозможные разбиения,

$N = C_L^l$, $L = l + k$.

«Принцип соответствия»:

$$EQ_\varepsilon(\mu, X^L) = \mathrm{P}_{X^l, X^k}\left\{\nu\left(\mu(X^l), X^k\right) - \nu\left(\mu(X^l), X^l\right) > \varepsilon\right\} \le P_\varepsilon(A).$$

Для оценивания обобщающей способности в теории Вапника-Червоненкиса вводится функционал равномерной сходимости частот в двух выборках. Основная претензия к этому функционалу — почему супремум берётся по всему семейству алгоритмов, в то время как реально используется только часть семейства. Проделаем такой мысленный эксперимент: пополним семейство A дополнительными алгоритмами, но пользоваться ими не будем. Понятно, что на существо дела это не повлияет — задача как решалась, так и будет решаться, однако значение функционала увеличится. Значит, функционал Вапника не вполне адекватно формализует само понятие обобщающей способности.

В комбинаторной теории вводится другой функционал, который зависит не от семейства алгоритмов, а от метода обучения. Он основан на принципе скользящего контроля. Это доля всевозможных разбиений выборки на обучение и контроль, при которых частота ошибок на контроле заметно превышает частоту ошибок на обучении. «Заметно», значит более чем на ε.

Отметим, что комбинаторный функционал не требует никаких теоретико-вероятностных предположений о природе данных, и вообще не опирается на понятие вероятности. Тем не менее, никаких радикальных противоречий между двумя теориями нет. Их примиряет следующий простой факт. Матожидание комбинаторного функционала равно вероятности того, что частота ошибок на контроле превысит частоту ошибок на обучении более чем на ε. Этот функционал уже предельно похож на вапниковский, только вместо супремума по всему семейство в нём появляется конкретный метод обучения.

Оценки обобщающей способности

$$P_\varepsilon(A) \le \Delta^A(L) \cdot \exp\left(-2\varepsilon^2 \frac{lk}{l+k}\right)$$
$$Q_\varepsilon(\mu, X^L) \le \Delta_L^l(\mu, X^L) \cdot \Gamma_L^l(\varepsilon)$$

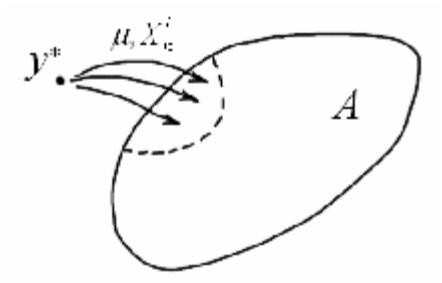
Новая мера *сложности*:

$\Delta^A(L) = \max_{X^L} \#\left\{ I(a, x_i)_{i=1}^l \,\middle|\, a \in A \right\}$ — функция роста по Вапнику

$\Delta_L^l(\mu, X^L) = \#\left\{ I(a_n, x_i)_{i=1}^l \,\middle|\, n = 1, \ldots, N \right\}$ — *локальная ф. роста*

Причины завышенности оценок Вапника-Червоненкиса:

1. Пренебрежение *эффектом локализации*
2. Погрешность экспоненциальной оценки
3. Погрешность разложения (переход от анализа качества к анализу сложности)



Верхние оценки этих двух функционалов также очень похожи по своей структуре. Они имеют вид произведения функции роста на множитель, экспоненциально убывающий с ростом длины выборки.

Функция роста по Вапнику — это число способов по-разному разделить выборку на классы всеми алгоритмами семейства A. То есть это характеристика сложности или разнообразия семейства.

Локальная функция роста определяется аналогично. Но учитываются только те алгоритмы, которые являются результатом настройки по всевозможным подвыборкам длины l.

Это позволяет учесть так называемый *эффект локализации* — зафиксировав задачу, то есть восстанавливаемую зависимость и метод обучения, мы уже не можем получить любой алгоритм семейства. Реально мы всегда пользуемся только каким-то узким подсемейством, состоящим из алгоритмов, более-менее похожих на восстанавливаемую зависимость.

Ещё одна причина завышенности заключается в том, что экспоненциальная верхняя оценка комбинаторного множителя Гамма завышена в несколько раз.

Третья причина завышенности связана с самим разложением функционала качества в произведение функции роста и комбинаторного множителя. Суть проблемы заключается примерно в следующем. При доказательстве этих оценок вероятность объединения событий грубо оценивается сверху суммой их вероятностей, несмотря на то, что эти события существенно совместны. Этот источник завышенности является результатом нашего упорного желания выделить сложность — функцию роста — в виде отдельного множителя. Есть подозрение, что любые сложностные оценки такого вида, будь то вапниковские или локальные, неминуемо окажутся завышенными, причём довольно сильно.

Итак, возникает вопрос: по какой причине мы всё-таки проигрываем в точности оценок — из-за пренебрежения эффектом локализации или из-за выделения функции роста?

Комбинаторный подход позволяет ответить на этот вопрос экспериментально.

Преимущества комбинаторного подхода

1. Отказ от избыточно сильной аксиоматики:
   — принципа равномерной сходимости;
   — гипотезы i.i.d.

2. Учёт метода обучения μ позволяет описать *эффект локализации* $\Rightarrow$
   $\Rightarrow$ снимается «запрет на сложность»

3. Функционал $Q_\varepsilon(\mu, X^L)$ можно измерять:

$$\widehat{Q}_\varepsilon(\mu, X^L) = \frac{1}{|\widehat{N}|}\sum_{n\in\widehat{N}}\Big[\nu(a_n, X_n^k) - \nu(a_n, X_n^l) > \varepsilon\Big],$$

где $a_n = \mu(X_n^l)$

$\widehat{N} \subset \{1,\ldots,N\}$ — случайное подмножество разбиений

Основное преимущество комбинаторного подхода — в более чёткой, неизбыточной аксиоматике.

Но для проведения экспериментов гораздо важнее другое преимущество — возможность измерить функционал качества с помощью стандартной техники скользящего контроля. Для измерения с любой желаемой степенью точности достаточно взять некоторое случайное подмножество разбиений, для каждого разбиения провести обучение и контроль, и усреднить результат по всем разбиениям.

Это позволяет провести эксперимент по измерению функции роста и ответить на вопрос: «Какой должна быть функция роста, чтобы оценка обобщающей способности получилась не завышенной?» Или, что то же самое: «Какое значение на самом деле принимает функция роста в данной задаче? Какова эффективная сложность той части семейства, которой мы реально пользуемся?»

Понятие эффективной локальной функции роста

Теорема. Пусть $\mu(X^l) = a = \text{const}$. Тогда $Q_\varepsilon(\mu, X^L) = \Gamma_L^l(\varepsilon, m) = \sum_{s=0}^{\lceil (m-\varepsilon k) l/L \rceil} \frac{C_m^s C_{L-m}^{l-s}}{C_L^l}$, где

$m = Lv(a, X^L)$.

Следствие (разновидность Закона Больших Чисел).

Пусть $\mu(X^l) = a = \text{const}$ и X^L — i.i.d. Тогда

$EQ_\varepsilon(\mu, X^L) = \mathrm{P}\left\{ v(a, X^k) - v(a, X^l) > \varepsilon \right\} \le \max_m \Gamma_L^l(\varepsilon, m) = \Gamma_L^l(\varepsilon)$.

Определение. *Эффективная локальная функция роста*:

$Q_\varepsilon(\mu, X^L) = \Delta_{\text{эфф}}(\mu, X^L) \cdot \Gamma_L^l(\varepsilon, m)$, при некотором m.

Интерпретация 1. $\Delta_{\text{эфф}}$ — такой должна быть функция роста, чтобы оценка получалась не завышенной.

Интерпретация 2. $\Delta_{\text{эфф}}$ — это не мера сложности, а коэффициент, показывающий, во сколько раз падает надёжность оценки $Q_\varepsilon(\mu, X^L)$ по сравнению с Законом Больших Чисел, вследствие переобучения.

Такая постановка вопроса приводит к понятию эффективной локальной функции роста.

Фактически, это не новая мера сложности, а просто коэффициент, показывающий, во сколько раз падает надёжность оценивания частоты ошибок на контроле, по сравнению с обычным законом больших чисел.

Причиной падения надёжности является, конечно же, эффект переобучения.

Методика измерения эффективной локальной функции роста

1. Измеряется $\widehat{Q}_\varepsilon(\mu, X^L)$, оценивается доверительный интервал:

$$\widehat{Q}_{\min} \le \widehat{Q}_\varepsilon(\mu, X^L) \le \widehat{Q}_{\max}$$

2. Поскольку m не известно, $\Gamma_L^l(\varepsilon, m)$ оценивается сверху и снизу.

Результат:

двусторонняя эмпирическая оценка
локальной эффективной функции роста:

$$\frac{\widehat{Q}_{\min}}{\max\limits_m \Gamma_L^l(\varepsilon, m)} \le \Delta_{\text{эфф}}(\mu, X^L) \le \frac{\widehat{Q}_{\max}}{\min\limits_m \Gamma_L^l(\varepsilon, m)}$$

Методика измерения эффективной локальной функции роста довольно проста.

Сначала вычисляется оценка скользящего контроля, вместе с доверительным интервалом. Фактически, это стандартное стохастическое измерение типа Монте-Карло.

Затем оценивается комбинаторный множитель. При этом также возникает доверительный интервал, связанный с принципиальным незнанием m — общего числа ошибок, допускаемых на полной выборке.

Наконец, эффективная локальная функция роста оценивается как их отношение.

Заметим, что такая методика измерения не проходит в рамках теории Вапника-Червоненкиса, так как довольно трудно оценить функционал, содержащий супремум по всему потенциально бесконечному семейству алгоритмов.

Идея эксперимента

Оценить, какая из 3^x причин завышенности более существенна.

Для этого:

— вычислить $\Delta^A(L)$ — функцию роста по Вапнику;

— измерить $\Delta_L^l(\mu, X^L)$ — локальную функцию роста;

— измерить $\widehat{Q}_\varepsilon(\mu, X^L)$;

— оценить $\Delta_{\text{эфф}}(\mu, X^L)$.

Тогда можно оценить факторы завышенности:

$$R_1 = \frac{\Delta^A(L)}{\Delta_L^l(\mu, X^L)}$$ (пренебрежение эффектом локализации)

$$R_3 = \frac{\Delta_L^l(\mu, X^L)}{\Delta_{\text{эфф}}(\mu, X^L)}$$ (погрешность разложения)

Осталось выбрать A и μ …

Цель эксперимента состояла в том, чтобы сравнить количественно три основные причины завышенности вапниковских оценок и понять, на чём мы теряем больше: на пренебрежении эффектом локализации или на выделении сложностного множителя — функции роста.

Для этого оценивается функция роста по Вапнику; локальная функция роста как число алгоритмов, реально построенных в результате обучения; и эффективная локальная функция роста — по нашей методике.

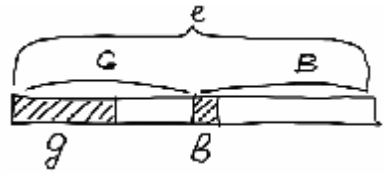
Логические алгоритмы классификации

Логические закономерности класса $c \in Y$:

$\varphi_c : X \to \{0,1\}$,

$b(\varphi_c) / g(\varphi_c) \ll B / G$

φ_c — конъюнкции ранга $\leq K$



Частота ошибок закономерности φ_c:

$$\nu(\varphi_c, X^l) = \frac{1}{l} \sum_{i=1}^{l} [\varphi_c(x_i) \neq [y_i = c]]$$

Метод поиска закономерности по обучающей выборке:

$\mu : X^l \to \varphi_c$

Понятия *обобщающей способности* и *функции роста* легко распространяются на методы поиска закономерностей

В качестве полигона для исследований были выбраны логические алгоритмы классификации, основанные на поиске конъюнктивных логических закономерностей. Тому были две причины.

Первая. Для них легко оценивается локальная функция роста. Это просто число закономерностей, просмотренных и оцененных в процессе поиска. Для этого в алгоритм поиска достаточно встроить счетчик. Кроме того, для них также легко посчитать и вапниковскую функцию роста — это число всех конъюнкций, которые возможно сконструировать из имеющихся признаков.

Вторая. У нас имелась под рукой библиотека логических алгоритмов классификации, о которой расскажут следующие докладчики (Кочедыков, Ивахненко).

Ещё один важный момент. Для простоты исследовалась обобщающая способность не всего классификатора, а отдельных логических закономерностей. Понятия частоты ошибок, локальной функции роста и обобщающей способности очевидным образом обобщаются на этот случай.

Результаты измерения функции роста

Задача*	число признаков	число термов	объектов		Оценки функции роста**		
			обуч.	тест	теорети-ческие	локаль-ные	эмпири-ческие
crx	15	1552	345	345	$1.1 \cdot 10^{11}$	$3.5 \cdot 10^{4}$	3.9
german	24	531	500	500	$5.7 \cdot 10^{9}$	$3.1 \cdot 10^{4}$	1.5
hepatits	19	134	77	78	$1.2 \cdot 10^{8}$	$1.8 \cdot 10^{4}$	2.6
liver	6	885	172	173	$7.9 \cdot 10^{10}$	$2.9 \cdot 10^{4}$	12.1

* Реальные задачи классификации из репозитория UCI

** При ограничении на максимальный ранг конъюнкций K=5

Результаты измерений несколько шокируют.

Теоретические оценки вапниковской функции роста имеют порядки 10 в 10-й степени.

Локальная функция роста — число реально просмотренных закономерностей — имеет порядки 10 в 4-й степени.

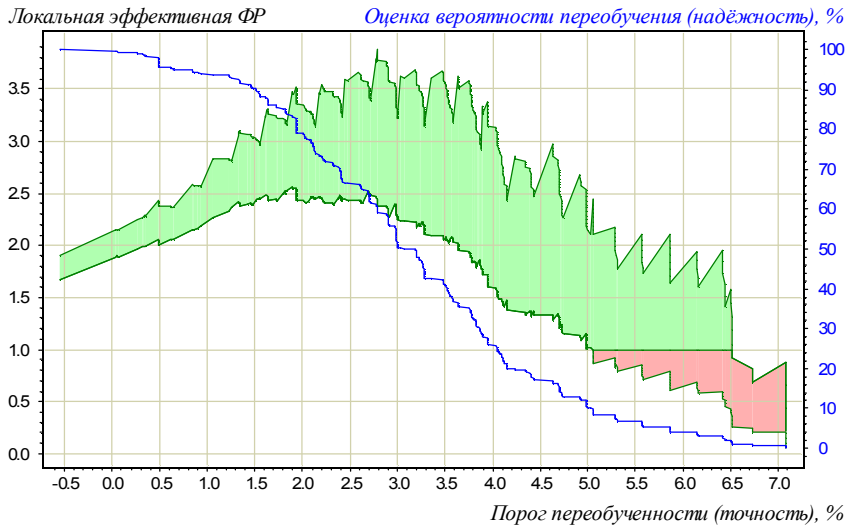
Эффективная локальная функция роста — порядок единицы!

Таким образом, *даже локальные оценки настолько завышены, что их невозможно применять для точного количественного оценивания обобщающей способности. Для этого необходимо отказываться от концепции сложности и искать другие источники обучаемости.*

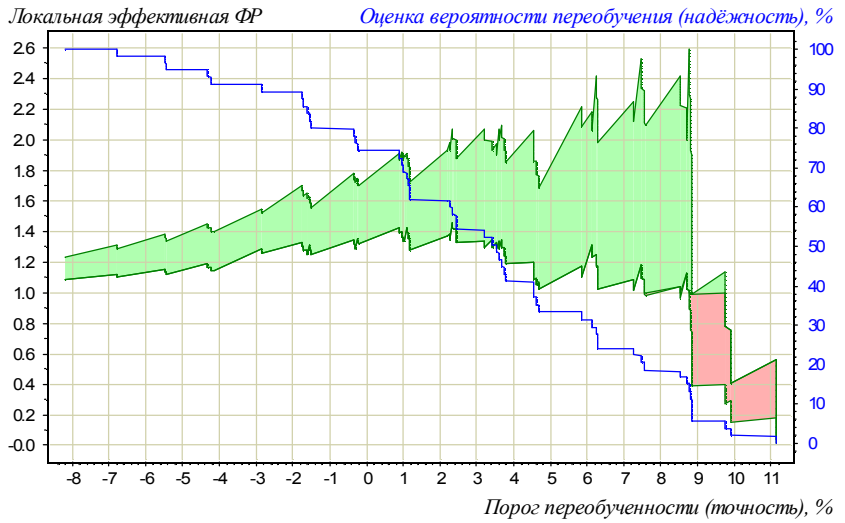
На следующих графиках показана зависимость эффективной локальной функция роста от параметр точности ε. По синей кривой на этих графиках можно определять, какое значение должен принимать этот параметр, если исходить из желаемого значения функционала качества (надёжности).

Зависимость э.л.ф.р. от параметра точности ε

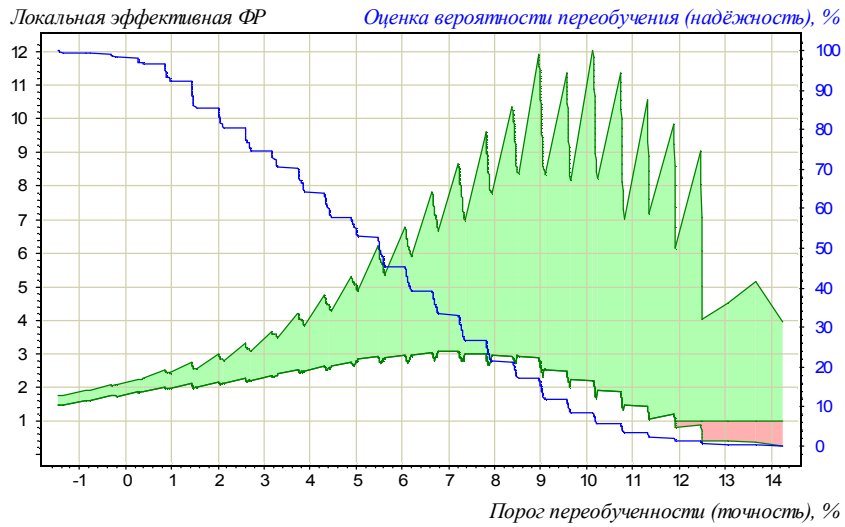
Задача: crx



Задача: hepatitis



Задача: liver



Выводы

1. Существенны обе причины завышенности оценок теории Вапника-Червоненкиса:
   — пренебрежение эффектом локализации
   — погрешность разложения
2. В логических алгоритмах классификации
   э.л.ф.р. имеет порядок единицы на реальных задачах
3. Интерпретация:
   если закономерности объективно проявляются в данных
   и если применяемый метод их находит,
   то переобучения почти нет,
   независимо от того, насколько сложно семейство
4. Данная методика позволяет вычислять *поправку на переобучение*
   при оценке вероятности ошибки отдельных правил
