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

Задача обучения по прецедентам

- Восстановление зависимости $y^*: X \to Y$
- Выборка $X^l = \left\{ x_1, ..., x_l \right\}$ 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.

Функционалы обобщающей способности

Статистическая теория Вапника-Червоненкиса:

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

Комбинаторная теория:

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

где $a_n = \mu(X_n^l)$,

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

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

«Принцип соответствия»:

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

Оценки обобщающей способности

$$P_\varepsilon(A) \quad \le \quad \Delta^A(L) \quad \cdot \exp\left(-2\varepsilon^2 \tfrac{lk}{l+k}\right)$$

$$Q_\varepsilon(\mu, X^L) \quad \le \quad \Delta_L^l(\mu, X^L) \quad \cdot \Gamma_L^l(\varepsilon)$$

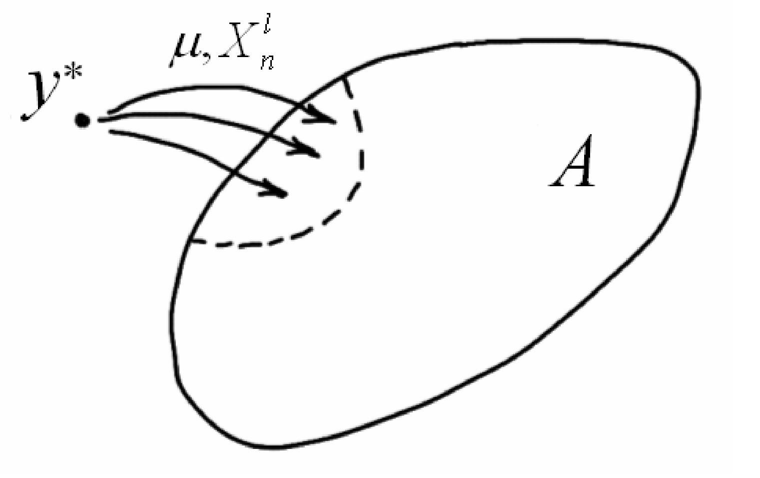
Новая мера *сложности*:

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

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



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

Понятие эффективной локальной функции роста

Теорема. Пусть $\mu(X^l) = a = \mathrm{const}$. 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 $m = L\nu(a, X^L)$.

Следствие (разновидность Закона Больших Чисел).

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

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

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 ЗБЧ, вследствие 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:

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

Идея эксперимента

Оценить, какая из 3-х 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…

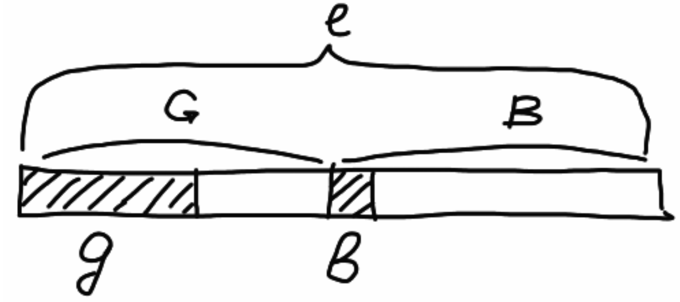
Логические алгоритмы классификации

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:

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

Метод поиска закономерности по обучающей выборке:

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

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

Результаты измерения функции роста

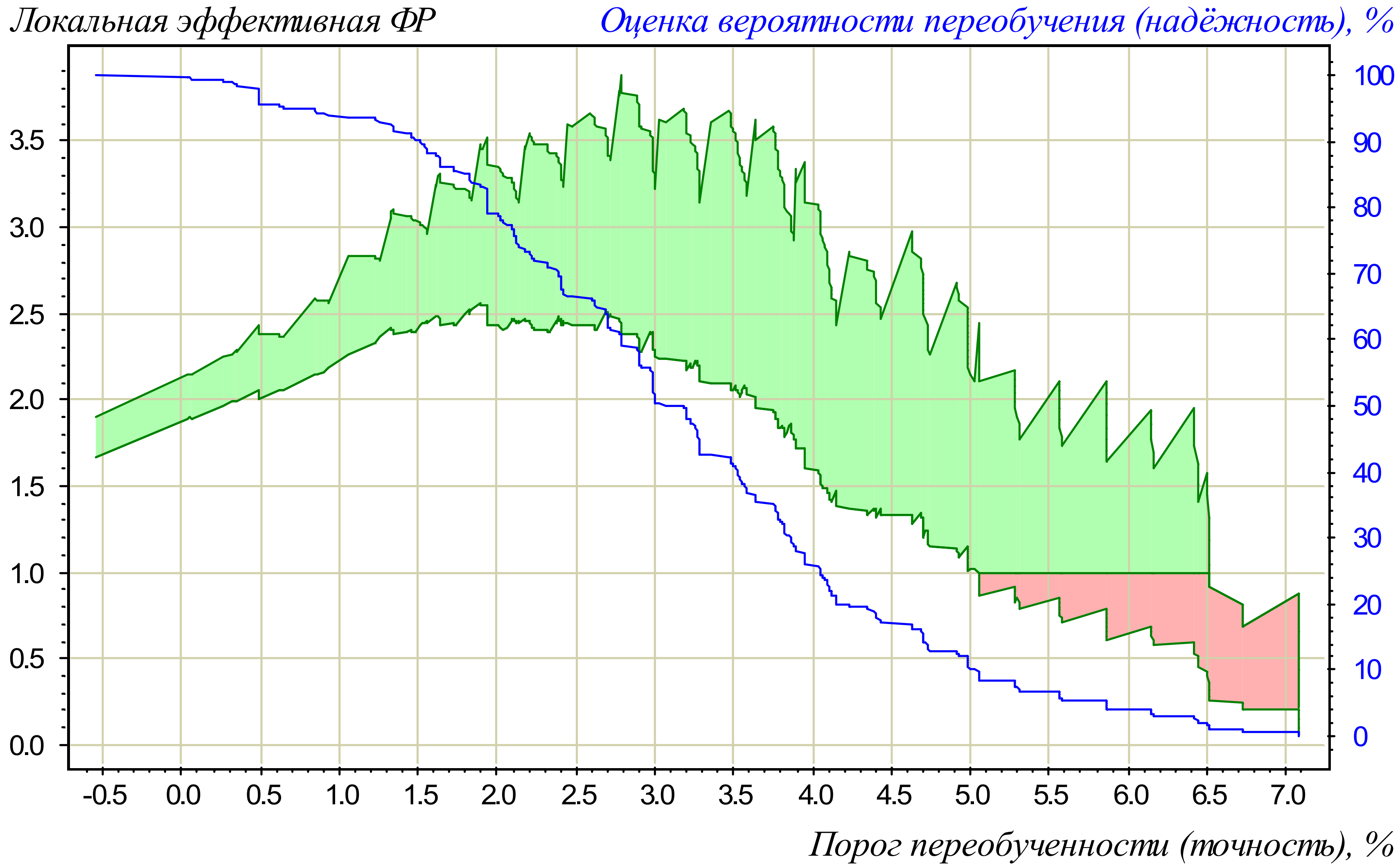
Задача*	число признаков	число термов	объектов обуч.	тест	Оценки функции роста** теоретические	локальные	эмпирические
crx	15	1552	345	345	$1.1 \cdot 10^{11}$	$3.5 \cdot 10^{4}$	3.9
german	24	531	500	500	$5.7 \cdot 10^{9}$	$3.1 \cdot 10^{4}$	1.5
hepatits	19	134	77	78	$1.2 \cdot 10^{8}$	$1.8 \cdot 10^{4}$	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

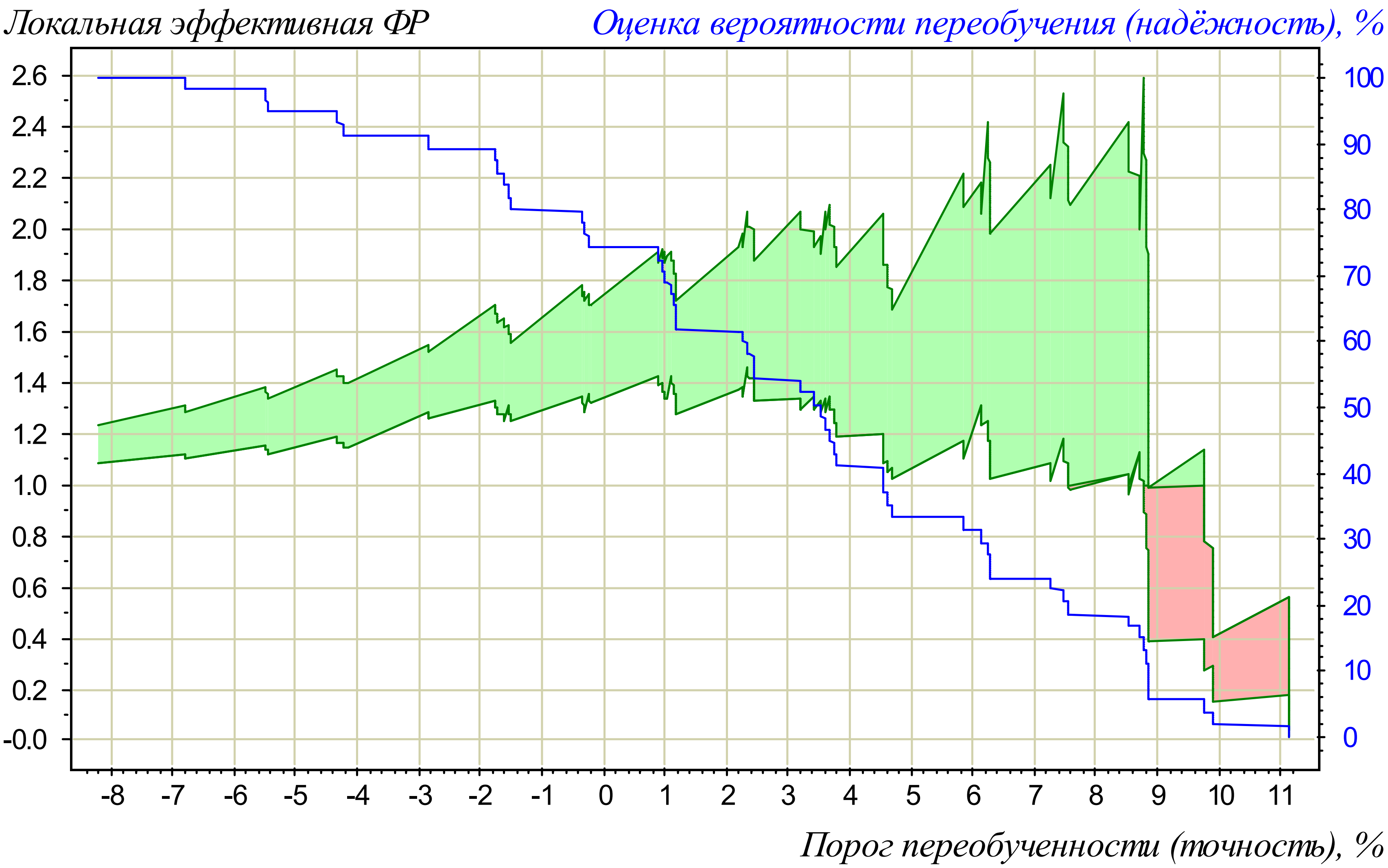
Зависимость э.л.ф.р. от параметра точности ε

Задача: crx



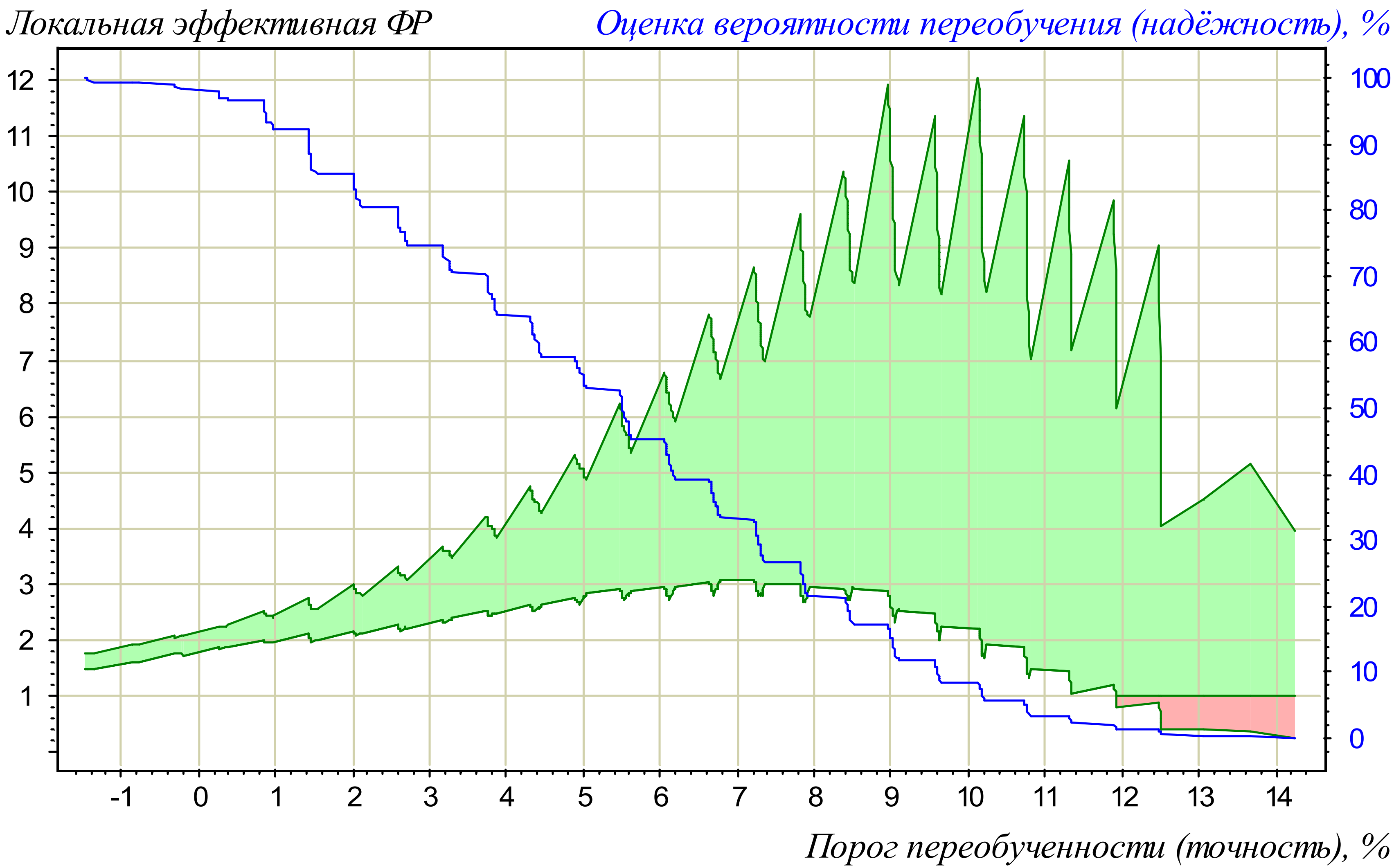
Зависимость э.л.ф.р. от параметра точности ε

Задача: hepatitis



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

Задача: liver



Выводы

1. Существенны обе причины завышенности оценок ТВЧ:
   — 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