Оценки надёжности эмпирических предсказаний (комбинаторный подход)

Воронцов Константин Вячеславович
voron@ccas.ru, http://www.ccas.ru/voron

Вычислительный Центр РАН,
Москва, Вавилова 40, 119991

Ломоносовские чтения, 17 апреля, 2008

Содержание

Слабая вероятностная аксиоматика

1. $X^L = \{x_i\}_{i=1}^L$ — конечная выборка объектов.
2. Все разбиения $X^L = X_n^\ell \cup X_n^k$, $n = 1, \ldots, N$, $N = C_L^k$ имеют равные шансы реализоваться $(L = \ell + k)$.

Тогда...

Слабая вероятностная аксиоматика

1. $X^L = \{x_i\}_{i=1}^L$ — конечная выборка объектов.
2. Все разбиения $X^L = X_n^\ell \cup X_n^k$, $n = 1, \ldots, N$, $N = C_L^k$ имеют равные шансы реализоваться $(L = \ell + k)$.

Тогда:

- *Вероятность события* A: $\{1, \ldots, N\} \to \{0, 1\}$:
$$\mathsf{P}_n A(n) = \frac{1}{N} \sum_{i=1}^{N} A(n).$$
- *Матожидание случайной величины* ξ: $\{1, \ldots, N\} \to \mathbb{R}$:
$$\mathsf{E}_n \xi(n) = \frac{1}{N} \sum_{i=1}^{N} \xi(n).$$
- *Распределение случайной величины* ξ: $\{1, \ldots, N\} \to \mathbb{R}$:
$$F_\xi(z) = \mathsf{P}_n \big[\xi(n) < z\big].$$

Задача эмпирического предсказания

- Реализуется разбиение (X_n^ℓ, X_n^k), $n \in \{1, \ldots, N\}$;
  выборка X_n^ℓ — *наблюдаемая*, выборка X_n^k — *скрытая*.
- Задана функция двух выборок $T\colon \mathbb{X}^k \times \mathbb{X}^\ell \to R$
- Требуется:

  1. Выбрать функцию $\hat{T}\colon \mathbb{X}^\ell \to R$ так, чтобы значение $\hat{T}_n = \hat{T}(X_n^\ell)$ предсказывало бы $T_n = T(X_n^k, X_n^\ell)$.

  2. Оценить точность предсказаний:

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

  где $d(\hat{r}, r)$ — отклонение предсказания $\hat{r}$ от истины r,
  например:
  $d(\hat{r}, r) = r - \hat{r}$ — для получения верхних оценок;
  $d(\hat{r}, r) = |r - \hat{r}|$ — для двусторонних оценок.

Пример 1: Закон больших чисел

Опр. *Частота* события $S \subset \mathbb{X}$ на конечной выборке $U \subset \mathbb{X}$:

$$\nu_S(U) = \frac{1}{|U|} \sum_{u \in U} [u \in S].$$

Положим: $R = \mathbb{R}$, $\hat{T}(U) = T(U) = \nu_S(U)$.

Теорема (известный классический факт)

Частота $\nu_S(X_n^\ell)$ *предсказывает частоту* $\nu_S(X_n^k)$,
причём справедлива точная оценка

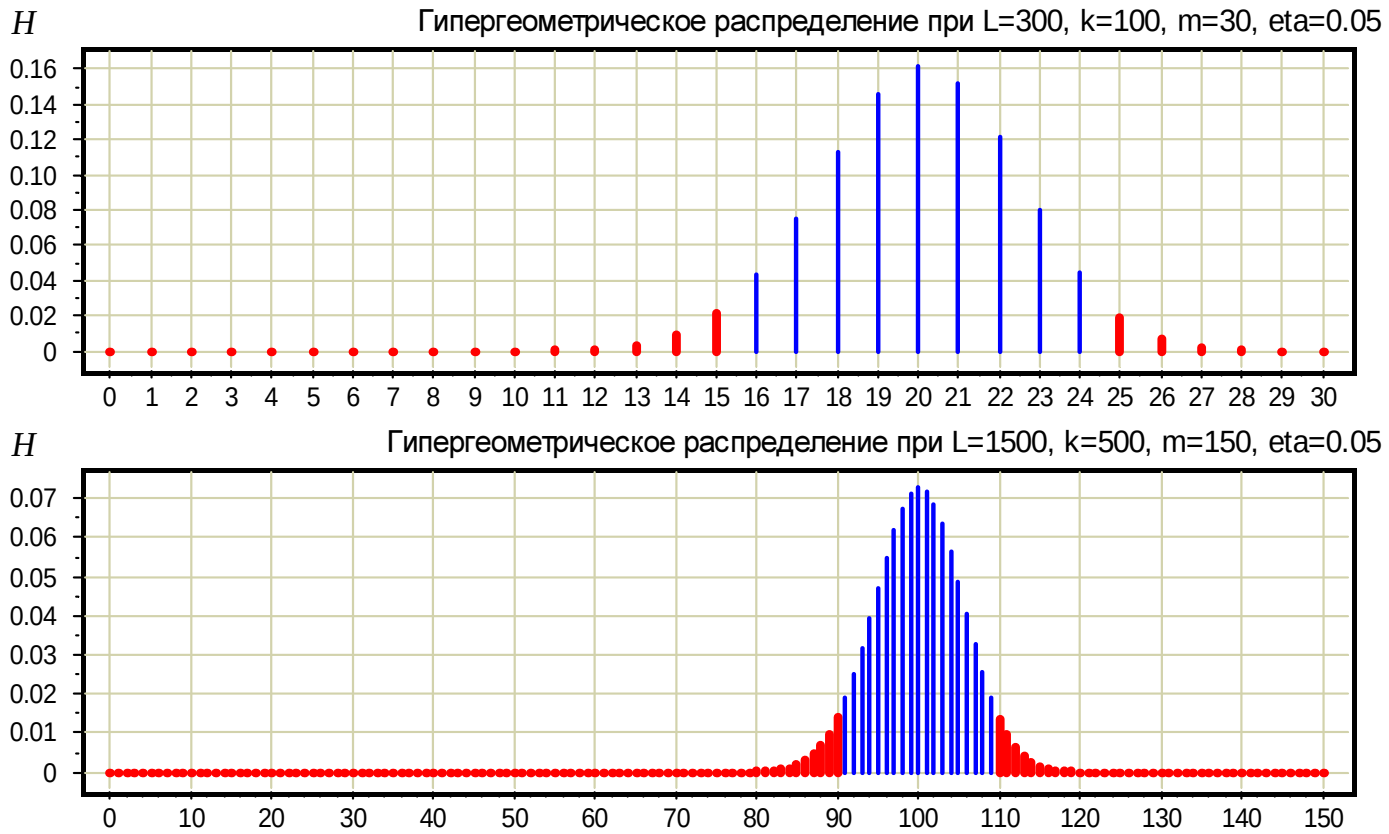
$$\mathsf{P}_n\left[\nu_S(X_n^k) - \nu_S(X_n^\ell) \geqslant \varepsilon\right] = H_L^{\ell,m}\left(s(\varepsilon)\right),$$

где $H_L^{\ell,m}(s)$ *— левый хвост гипергеометрического распределения,* $s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m - \varepsilon k) \right\rfloor$, $m = L\nu_S(X^L)$.

Пример 1: Закон больших чисел

Левый хвост гипергеометрического распределения:

$$H_L^{\ell,m}(s(\varepsilon)) = \sum_{t=s_0}^{s(\varepsilon)} \frac{C_m^t C_{L-m}^{\ell-t}}{C_L^\ell}, \quad s_0 = \max\{0, m-k\}$$



Пример 2: Сходимость эмпирических распределений

Опр. *Эмпирическое распределение* функции $\xi\colon \mathbb{X} \to \mathbb{R}$ на конечной выборке $U \subseteq \mathbb{X}$ есть

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

Положим: R — пр-во кус.-пост. невозр. функций $F\colon \mathbb{R} \to [0, 1]$,
$T(U) = \hat{T}(U) = F_\xi(z, U)$, $d(\hat{r}, r) = \max\limits_{z \in \mathbb{R}} |r(z) - \hat{r}(z)|$.

Теорема (малоизвестный факт)

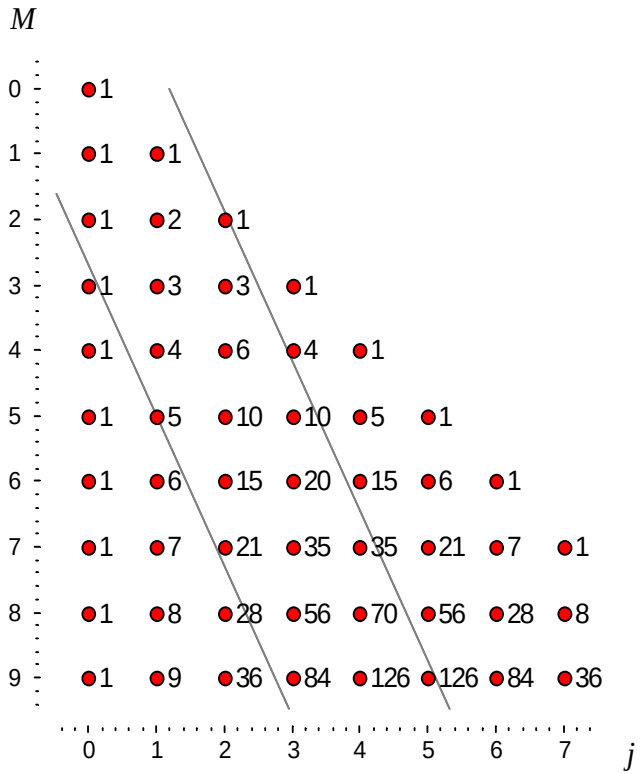
Если значения $\xi(x_i)$ попарно различны на X^L, то

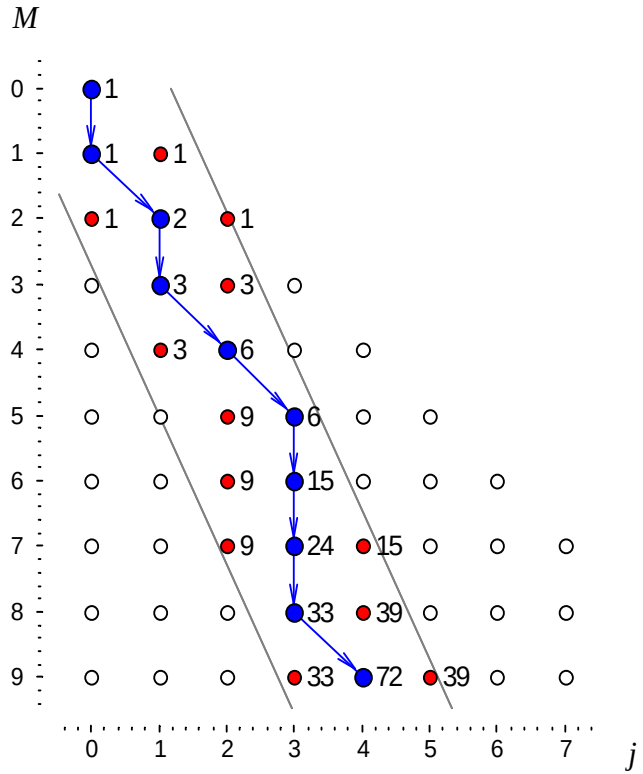
$$\mathsf{P}_n\Big\{\max_{z \in \mathbb{R}} |F_\xi(z, X_n^k) - F_\xi(z, X_n^\ell)| > \varepsilon\Big\} = \frac{G_L^k(\varepsilon)}{C_L^k},$$

где $G_L^k(\varepsilon)$ — значение из усечённого треугольника Паскаля.

Пример 2: Сходимость эмпирических распределений

Усечённый треугольник Паскаля $G_m^j(\varepsilon)$ — между двумя прямыми $j^+(m) = \frac{k}{L}(m + \varepsilon\ell)$, $j^-(m) = \frac{k}{L}(m - \varepsilon\ell)$.





Здесь $L = 16$, $k = 7$, $\varepsilon = 0.3$.

Слабая вероятностная аксиоматика

- Достаточна для доказательства фундаментальных фактов:
  - закон больших чисел: точная оценка скорости сходимости;
  - критерий Смирнова: точная оценка скорости сходимости;
  - многие непараметрические критерии;
  - оценки обобщающей способности (Вапника-Червоненкиса).
- Основана на более слабых предположениях:
  - нет определения вероятности как меры на $\mathbb{X}$;
  - нет определения вероятности как частоты при $L \to \infty$;
  - нет необходимости определять понятие «вероятность»!
    а что же осталось?
  - предположение о *независимости* объектов выборки X^L

Связь с классической аксиоматикой Колмогорова

Теорема (Принцип соответствия)

Если в слабой аксиоматике для некоторой функции $\phi(X^\ell, X^k)$ получена оценка

$$Q_\varepsilon(X^L) = \mathsf{P}_n\big[\phi(X_n^\ell, X_n^k) > \varepsilon\big] \leqslant \eta(\varepsilon, X^L)$$

то аналогичная оценка верна и в аксиоматике Колмогорова

$$\mathsf{E}_{X^L} Q_\varepsilon(X^L) = \mathsf{P}_{X^L}\big[\phi(X^\ell, X^k) > \varepsilon\big] \leqslant \mathsf{E}_{X^L}\eta(\varepsilon, X^L)$$

Если $\eta(\varepsilon, X^L) \equiv \eta(\varepsilon)$, то оценка справедлива для любой X^L.

Философу на заметку:
«трансдукция по форме, индукция по содержанию» ;))

Связь с эмпирическим оцениванием

Что делать, когда теоретические оценки не известны или слишком сильно завышены?

$$Q_\varepsilon(X^L) = \mathsf{P}_n\big[d(\hat{T}_n, T_n) > \varepsilon\big] \leqslant \boxed{???}$$

Остаётся возможность измерить Q_ε эмпирически:

$$Q_\varepsilon(X^L) \leqslant \frac{1}{|N'|}\sum_{n\in N'}\big[d(\hat{T}_n, T_n) > \varepsilon\big] + \underbrace{\delta(N, |N'|)}_{\to 0 \text{ при } |N'|\to N},$$

где множество разбиений $N' \subset \{1, \ldots, N\}$ выбирается
— либо случайно (метод Монте-Карло),
— либо по блокам (k-fold Cross-Validation)

Слабая вероятностная аксиоматика: за и против

\+ Лучше подходит для задач анализа данных и обучения по прецедентам

\− . . . но хуже — для континуальных стохастических явлений

\+ Даёт не асимптотические, не завышенные оценки

\− . . . требующие сложных комбинаторных вычислений

\+ Удовлетворяет «принципу соответствия»

\− . . . однако не все классические теоремы имеют аналоги в слабой аксиоматике

Слабая вероятностная аксиоматика: за и против

\+ Лучше подходит для задач анализа данных и обучения по прецедентам

\− . . . но хуже — для континуальных стохастических явлений

\+ Даёт не асимптотические, не завышенные оценки

\− . . . требующие сложных комбинаторных вычислений

\+ Удовлетворяет «принципу соответствия»

\− . . . однако не все классические теоремы имеют аналоги в слабой аксиоматике

Открытая проблема:

Какую часть математической статистики можно воспроизвести в рамках слабой аксиоматики?

Задача обучения по прецедентам

- $\mathbb{X}$ — множество объектов, $\mathbb{Y}$ — множество ответов.
- $y^*: \mathbb{X} \to \mathbb{Y}$ — неизвестная *целевая зависимость*.
- *Обучающая выборка* $X^\ell = \{x_i, y_i\}_{i=1}^\ell \subset \mathbb{X} \times \mathbb{Y}$, $y_i = y^*(x_i)$.
- *Метод обучения* $\mu: X^\ell \mapsto f$.
- *Частота ошибок* алгоритма $f: \mathbb{X} \to \mathbb{Y}$ на выборке $U \subset \mathbb{X}$:
$$\nu(f, U) = \frac{1}{|U|} \sum_{u \in U} \big[|f(x_i) - y_i| > \delta\big].$$
- *Обобщающая способность*:
средняя ошибка $\nu(\mu X^\ell, U)$ должна быть достаточно мала для большинства выборок $U \in \mathbb{X}^*$.

Оценки Вапника-Червоненкиса [1971]

- Для семейства алгоритмов F в сильной аксиоматике:

$$P_\varepsilon(F) = \mathsf{P}_{X^L}\Big[\sup_{f\in F}\big(\nu(f,X^k)-\nu(f,X^\ell)\big) > \varepsilon\Big]$$
$$\leqslant \Delta^F(L)\cdot 1.5\, e^{-\varepsilon^2\ell} \quad \text{(при условии } \ell = k\text{)};$$

- $\Delta^F(L)$ — *функция роста* (shatter coefficient) семейства F — макс. число функций $f \in F$, попарно различимых на X^L;
$\Delta^F(L) \leqslant 1.5\,\frac{L^h}{h!}$, $h = \mathrm{VCdim}(F)$ — *ёмкость* семейства F.

Оценки Вапника-Червоненкиса [1971]

- Для семейства алгоритмов F в сильной аксиоматике:

$$P_\varepsilon(F) = \mathsf{P}_{X^L}\Big[\sup_{f\in F}\big(\nu(f,X^k)-\nu(f,X^\ell)\big) > \varepsilon\Big]$$
$$\leqslant \Delta^F(L)\cdot 1.5\, e^{-\varepsilon^2\ell} \quad \text{(при условии } \ell = k\text{)};$$

- $\Delta^F(L)$ — *функция роста* (shatter coefficient) семейства F — макс. число функций $f \in F$, попарно различимых на X^L;
  $\Delta^F(L) \leqslant 1.5\,\frac{L^h}{h!}$, $h = \mathrm{VCdim}(F)$ — *ёмкость* семейства F.
- Оценка крайне завышена, т. к. не учитывает X^ℓ, $y^*(x)$, μ.
- **Цель:**
  сравнить причины завышенности количественно.

Оценки Вапника-Червоненкиса — теперь в слабой аксиоматике

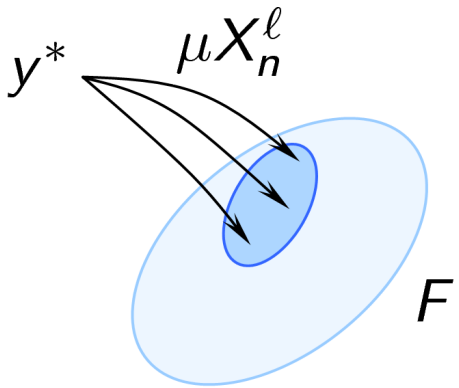
- В слабой вероятностной аксиоматике:

$$Q_\varepsilon(\mu, X^L) = \mathsf{P}_n\Big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \varepsilon\Big]$$
$$\leqslant \Delta_L^\ell(\mu, X^L) \cdot \max_m H_L^{\ell,m}(s(\varepsilon))$$
$$\left(\leqslant \Delta^F(L) \cdot 1.5\, e^{-\varepsilon^2 \ell}\right);$$

  где алгоритм $f_n = \mu X_n^\ell$ — результат обучения на X_n^ℓ;
- $\Delta_L^\ell(\mu, X^L)$ — *локальный коэффициент разнообразия (local shatter coefficient)* множества алгоритмов $F_L^\ell(\mu, X^L) = \left\{f_n = \mu X_n^\ell \;\middle|\; n = 1, \ldots, N\right\}$, которые могут быть результатом обучения для данной задачи $\langle \mu, X^L \rangle$:

Оценки, зависящие от данных

- *Эффект локализации:*
  Если y^*, μ, X^L фиксированы,
  то в результате обучения
  $F_L^\ell(\mu, X^L) \Subset F$



- Равномерная сходимость [Вапник, Червоненкис, 1969]
- Theory of learnable (PAC-learning) [Valiant, 1982]
- Concentration inequalities [Talagrand, 1995]
- Data-dependent bounds [Haussler, 1992; Bartlett, 1998;...]
- Self-bounding learning algorithms [Freund, 1998]
- PAC-Bayes bounds [McAllester, 1999; Langford, 2005]
- Microchoice bounds [Langford, Blum, 2001]
- Algorithmic luckiness [Herbrich, Williamson, 2002]
- Shell bounds [Langford, 2002]

Оценки, зависящие от данных: дальнейшее уточнение

- **Идея:** скалярная характеристика разнообразия содержит слишком мало информации о процессе обучения.
- Вводится *профиль разнообразия (shatter profile)* $\{D_m\}_{m=0}^{L}$:

$$\Delta_L^\ell(\mu, X^L) = \sum_{m=1}^{L} D_m(\mu, X^L),$$

где $D_m(\mu, X^L)$ — локальный коэффициент разнообразия множества алгоритмов $\left\{f_n \;\middle|\; \nu(f_n, X^L) = \frac{m}{L},\ n = 1, \ldots, N\right\}$.

Оценки, зависящие от данных: дальнейшее уточнение

- **Идея:** скалярная характеристика разнообразия содержит слишком мало информации о процессе обучения.
- Вводится *профиль разнообразия (shatter profile)* $\{D_m\}_{m=0}^{L}$:

$$\Delta_L^\ell(\mu, X^L) = \sum_{m=1}^{L} D_m(\mu, X^L),$$

где $D_m(\mu, X^L)$ — локальный коэффициент разнообразия множества алгоритмов $\left\{f_n \;\middle|\; \nu(f_n, X^L) = \frac{m}{L},\ n = 1, \ldots, N\right\}$.

Теорема

В слабой аксиоматике справедлива оценка:

$$Q_\varepsilon(\mu, X^L) \leqslant \sum_{m=1}^{L} D_m(\mu, X^L) \cdot H_L^{\ell,m}\left(s(\varepsilon)\right);$$

Эффективный локальный профиль разнообразия

- **Обратная задача**:
  при каком профиле D_m оценка была бы точной?
- **Теорема**

$$Q_{\varepsilon,m}(\mu, X^L) = \mathsf{P}_n\Big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \varepsilon\Big]\Big[\nu(f_n, X^L) = \tfrac{m}{L}\Big]$$
$$\leqslant D_m(\mu, X^L) \cdot H_L^{\ell,m}\left(s(\varepsilon)\right);$$

Заменим здесь «$\leqslant$» на «$=$» и выразим D_m:

Эффективный локальный профиль разнообразия

- **Обратная задача**:
  при каком профиле D_m оценка была бы точной?
- **Теорема**

$$Q_{\varepsilon,m}(\mu, X^L) = \mathsf{P}_n\Big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \varepsilon\Big]\Big[\nu(f_n, X^L) = \tfrac{m}{L}\Big]$$
$$\leqslant D_m(\mu, X^L) \cdot H_L^{\ell,m}\left(s(\varepsilon)\right);$$

  Заменим здесь «$\leqslant$» на «$=$» и выразим D_m:
- *Эффективный локальный профиль разнообр.* $\{\hat{D}_m(\varepsilon)\}_{m=0}^L$:

$$\hat{D}_m(\varepsilon) = \frac{\frac{1}{|N'|}\sum_{n\in N'}\big[\nu(f_n, X_n^k) - \nu(f_n, X_n^\ell) > \varepsilon\big]\big[\nu(f_n, X^L) = \frac{m}{L}\big]}{H_L^{\ell,m}\left(s(\varepsilon)\right)}.$$

- *Эффективный локальный коэффициент разнообразия*:
  $\hat{\Delta}_L^\ell(\varepsilon) = \hat{D}_0(\varepsilon) + \cdots + \hat{D}_L(\varepsilon)$.

Факторы завышенности оценок ТВЧ

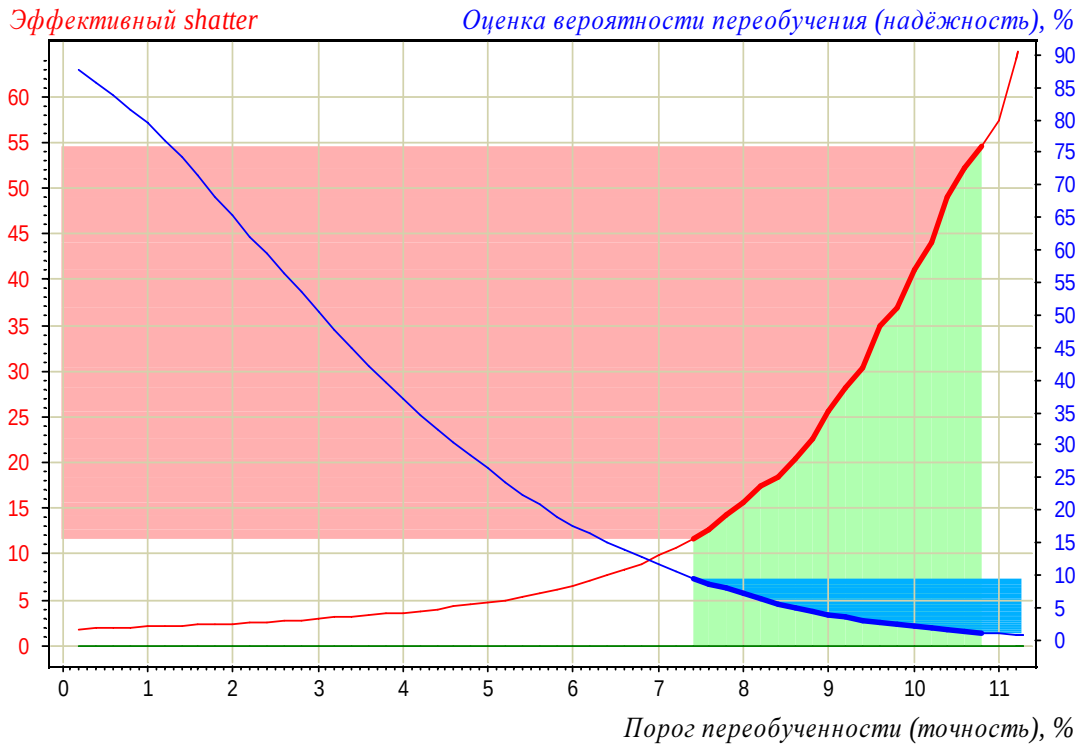
- r_1: пренебрежение эффектом локализации
- r_2: пренебрежение степенью различности алгоритмов (выделение коэффициента D_m как сомножителя)
- r_3: свёртка профиля разнообразия
- r_4: экспоненциальная аппроксимация гипергеометрического распределения $H = \max\limits_m H_L^{\ell,m}(s(\varepsilon))$

Степень завышенности оценки ТВЧ раскладывается в виде:

$$\frac{\Delta^F(L)\cdot\frac{3}{2}\,e^{-\varepsilon^2\ell}}{\hat{Q}_\varepsilon} = \overbrace{\frac{\Delta^F(L)}{\Delta_L^\ell}}^{r_1}\cdot\overbrace{\frac{\Delta_L^\ell}{\hat{\Delta}_L^\ell(\varepsilon)}}^{r_2(\varepsilon)}\cdot\overbrace{\frac{\hat{\Delta}_L^\ell(\varepsilon)H}{\hat{Q}_\varepsilon}}^{r_3(\varepsilon)}\cdot\overbrace{\frac{\frac{3}{2}\,e^{-\varepsilon^2\ell}}{H}}^{r_4(\varepsilon)}$$

Методика измерения $\hat{\Delta}_L^\ell(\varepsilon)$

- фиксируется диапазон надёжности $\hat{Q}_\varepsilon \in [0.01, 0.1]$;
- для него определяется диапазон точности ε;
- на котором определяется минимум и максимум $\hat{\Delta}_L^\ell(\varepsilon)$.



Логические алгоритмы классификации

- *Закономерность* (правило) — предикат $\phi_y : X \to \{0,1\}$, выделяющий преимущественно объекты класса y
- *Взвешенное голосование* правил:

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

где $\phi_y^t(x)$ — t-ое правило класса y, w_y^t — вес правила

- *Метод обучения закономерностей* класса y: $\mu_y : X^\ell \mapsto \left\{ \phi_y^t(x) \;\middle|\; t = 1, \ldots, T_y \right\}$,
- Преимущества:
  - известна функция роста $\Delta^F(L)$
  - легко оценить снизу локальный профиль $\Delta_L^\ell(\mu, X^L)$
  - легко оценить эффективный локальный профиль

Экспериментальный стенд

- 7 задач классификации из репозитория UCI, $|\mathbb{Y}| = 2$
- 20×2-кратный скользящий контроль, $\ell = k$
- Алгоритм Forecsys ScoringAce® [Кочедыков, Ивахненко,...]

Задача	L	n	C4.5	C5.0	RIPPER	SLIPPER	Forecsys
crx	690	15	15.5	14.0	15.2	15.7	14.3 ± 0.2
german	1000	20	27.0	28.3	28.7	27.2	28.5 ± 1.0
hepatitis	155	19	18.8	20.1	23.2	17.4	16.7 ± 1.7
horse-colic	300	25	16.0	15.3	16.3	15.0	16.4 ± 0.5
hypothyroid	3163	25	0.4	0.4	0.9	0.7	0.8 ± 0.04
liver	345	6	37.5	31.9	31.3	32.2	29.2 ± 1.6
promoters	106	57	18.1	22.7	19.0	18.9	12.0 ± 2.0

L — длина выборки; n — число признаков;

по алгоритмам: процент ошибок на контроле.

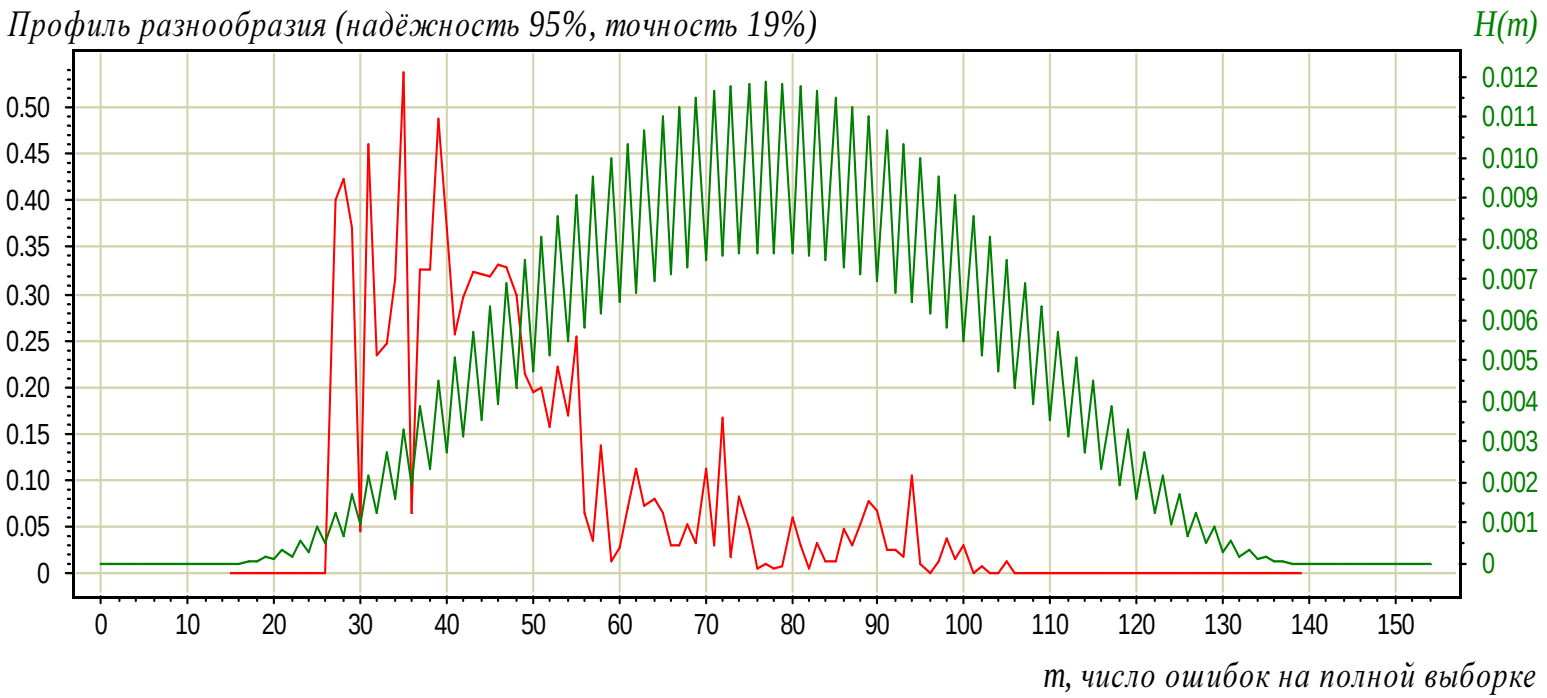
Результаты

Степени завышенности при значении точности ε_0, соответствующей надёжности $\hat{Q}_\varepsilon = 0.05$.

Задача	y	r_1	$r_2(\varepsilon)$	$r_3(\varepsilon)$	$r_4(\varepsilon)$	$\hat{\Delta}_L^\ell[\varepsilon_1, \varepsilon_2]$	$\hat{\Delta}_L^\ell(\varepsilon_0)$
crx	0	890	680	3.1	32.6	[10; 41]	24
	1	690	1700	1.6	11.6	[11; 180]	12
german	1	8 950	1500	1.7	10.9	[38; 530]	54
	2	37 000	9000	1.2	9.9	[1.0; 2.2]	1.9
hepatitis	0	23	280	13.4	9.5	[11; 148]	83
	1	55	680	2.4	22.5	[12; 27]	15
horse-colic	1	72	4500	2.1	7.2	[2; 9]	7
	2	140	3400	3.6	7.3	[3; 6]	6
hypothyroid	0	61 000	400	32.2	16.5	[3; 220]	21
	1	153 000	460	3.8	28.7	[2; 44]	30
promoters	0	94	340	5.9	9.8	[36; 230]	72
	1	150	790	3.4	6.9	[9; 22]	18

Как выглядит эффективный локальный профиль?

- Зависимость эффективного локального профиля $\hat{D}_m(\varepsilon)$ и функции $H(m) = H_L^{\ell,m}(s(\varepsilon))$ от числа ошибок m на X^L.
- Выбрано значение $\varepsilon = 0.19$, соответствующее $\hat{Q}_\varepsilon = 0.05$.
- Задача UCI:hepatitis, класс $y = 0$.



Выводы и дальнейшие исследования

- Выводы
  - Эффективная локальная ёмкость $\leqslant 1$
  - Оценки сложности Δ порядка 10^2 науке пока не известны
  - Основные пути улучшения оценок:
    Локализация $+$ Учёт сходства алгоритмов
- Дальнейшие исследования
  - Мета-обучение критериев информативности для улучшения качества правил в логических алгоритмах
  - Явный учёт степени сходства алгоритмов для оптимизации поиска в локально-переборных алгоритмах обучения