Комбинаторная теория переобучения и поиск логических закономерностей

Воронцов Константин Вячеславович

Вычислительный центр им. А. А. Дородницына РАН

15-я всероссийская научная конференция
«Математические методы распознавания образов»
11–17 сентября 2011 • г. Петрозаводск

Содержание

Задача обучения по прецедентам. Матрица ошибок.

$\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечное *генеральное множество* объектов;
$\mathbb{A} = \{a_1, \ldots, a_D\}$ — конечное семейство *алгоритмов*;
$I(a, x) = [$алгоритм a ошибается на объекте $x]$;

$L \times D$-матрица ошибок с попарно различными столбцами:

	a_1	a_2	a_3	a_4	a_5	a_6	$\cdots$	a_D	
x_1	1	1	0	**0**	0	1	$\cdots$	1	X — наблюдаемая
$\ldots$	0	0	0	**0**	1	1	$\cdots$	1	(обучающая) выборка
x_ℓ	0	0	1	**0**	0	0	$\cdots$	0	длины ℓ
$x_{\ell+1}$	0	0	0	**1**	1	1	$\cdots$	0	$\bar{X}$ — скрытая
$\ldots$	0	0	0	**1**	0	0	$\cdots$	1	(контрольная) выборка
x_L	0	1	1	**1**	1	1	$\cdots$	0	длины $k = L - \ell$

$m(a, X) = \sum_{x \in X} I(a, x)$ — *число ошибок* $a \in \mathbb{A}$ на выборке $X \subset \mathbb{X}$;
$\nu(a, X) = m(a, X)/|X|$ — *частота ошибок* a на выборке X;

Задача оценивания вероятности переобучения

Опр. *Метод обучения* μ: $X \mapsto a$ произвольной выборке $X \subset \mathbb{X}$ ставит в соответствие некоторый алгоритм $a \in \mathbb{A}$.

Пример. Метод *минимизации эмпирического риска*:

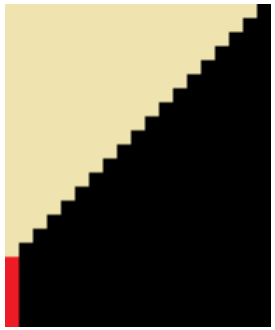
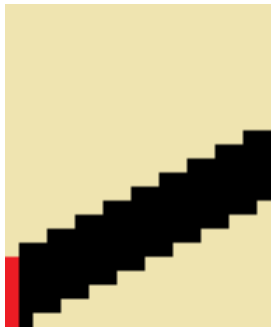
$$\mu X = \arg\min_{a \in \mathbb{A}} \nu(a, X).$$

Опр. *Вероятность переобучения* — доля разбиений $X \sqcup \bar{X} = \mathbb{X}$, при которых частота ошибок на контроле $\bar{X}$ существенно (на ε) превышает частоту ошибок на обучении X:

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\nu\big(\mu X, \bar{X}\big) - \nu\big(\mu X, X\big) \geqslant \varepsilon\big].$$

Здесь и далее $\mathsf{P} \equiv \dfrac{1}{C_L^\ell} \sum_{X \subset \mathbb{X}:\ |X|=\ell}$

Эксперимент: четыре семейства алгоритмов, заданных матрицами ошибок; лучший алгоритм у всех одинаков

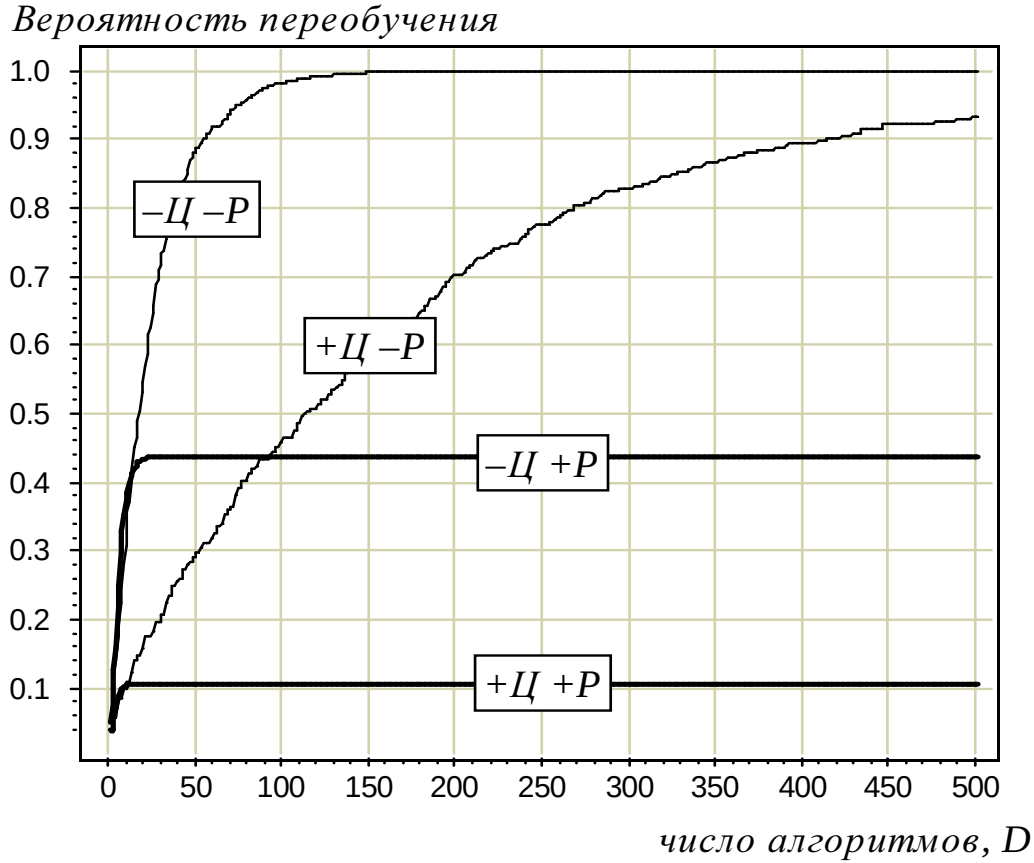
	с расслоением по числу ошибок	без расслоения по числу ошибок
каждая пара соседних алгоритмов отличается только на одном объекте (образуется *цепь*)		
соседние алгоритмы существенно различны, (*цепь* не образуется)		

Постепенно добавляя алгоритмы в $\{a_1, \ldots, a_D\}$, построим зависимости вероятности переобучения Q_ε от числа D.

Вывод: расслоение и связность снижают переобучение

Эксперимент с четырьмя семействами алгоритмов, $\ell = k = 100$, $\varepsilon = 0.05$:

+Ц — цепь;
−Ц — не цепь;
+Р — с расслоением;
−Р — без расслоения;



Вывод: получение точных оценок вероятности переобучения невозможно без учёта эффектов расслоения и связности.

Граф расслоения–связности множества алгоритмов

Определим бинарные отношения на множестве алгоритмов $\mathbb{A}$:

частичный порядок $a \leqslant b$: $I(a,x) \leqslant I(b,x)$ для всех $x \in \mathbb{X}$;

предшествование $a \prec b$: $a \leqslant b$ и $\|b-a\| = 1$.

Опр. *Граф расслоения–связности* $\langle \mathbb{A}, E \rangle$:

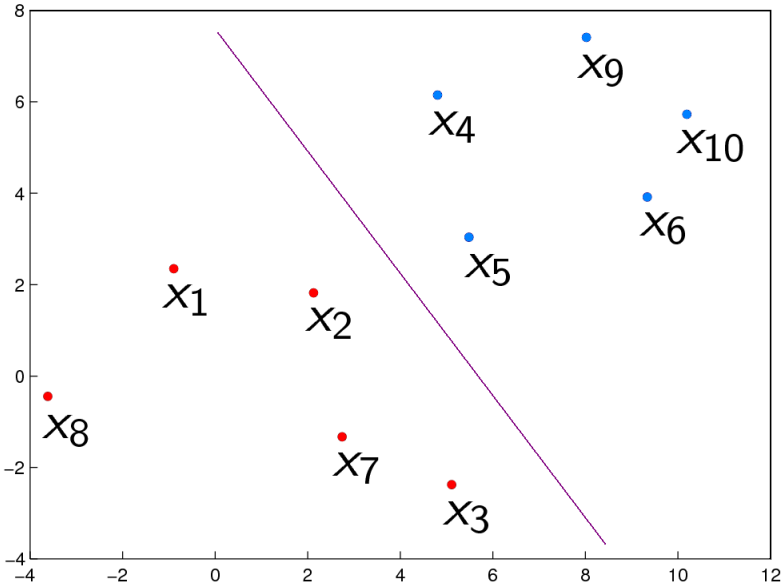
$\mathbb{A}$ — множество попарно различных векторов ошибок;

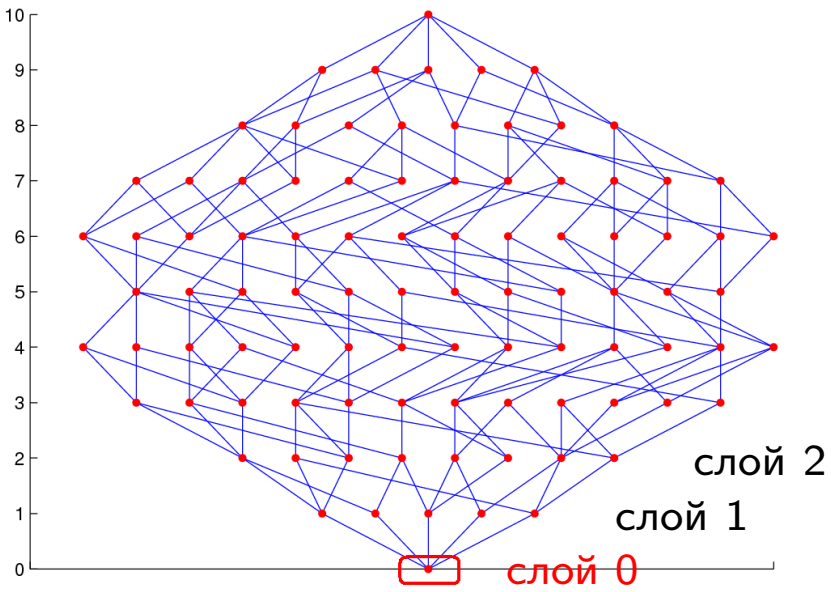
$E = \{(a,b) \colon a \prec b\}$.

Свойства графа расслоения–связности:

- это подграф графа Хассе отношения порядка $\leqslant$ на $\mathbb{A}$;
- каждому ребру (a,b) соответствует объект $x_{ab} \in \mathbb{X}$, такой, что $I(a,x_{ab}) = 0, \quad I(b,x_{ab}) = 1$;
- граф является многодольным со слоями $A_m = \{a \in \mathbb{A} \colon m(a,\mathbb{X}) = m\}, \quad m = 0,\ldots,L+1$;

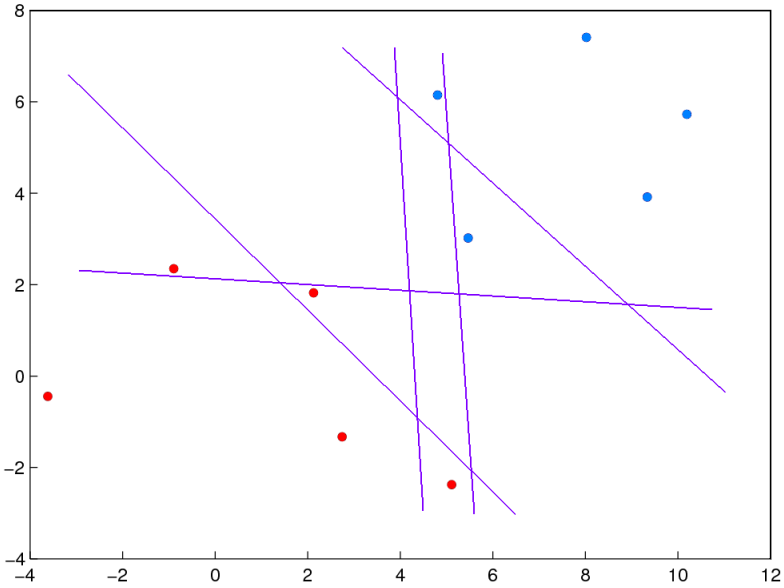
Пример. Семейство линейных алгоритмов классификации

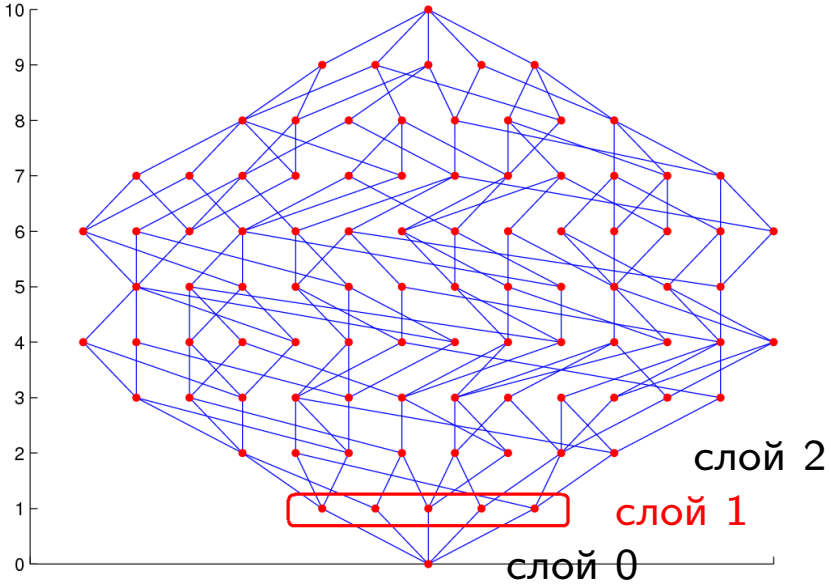




	слой 0
x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0

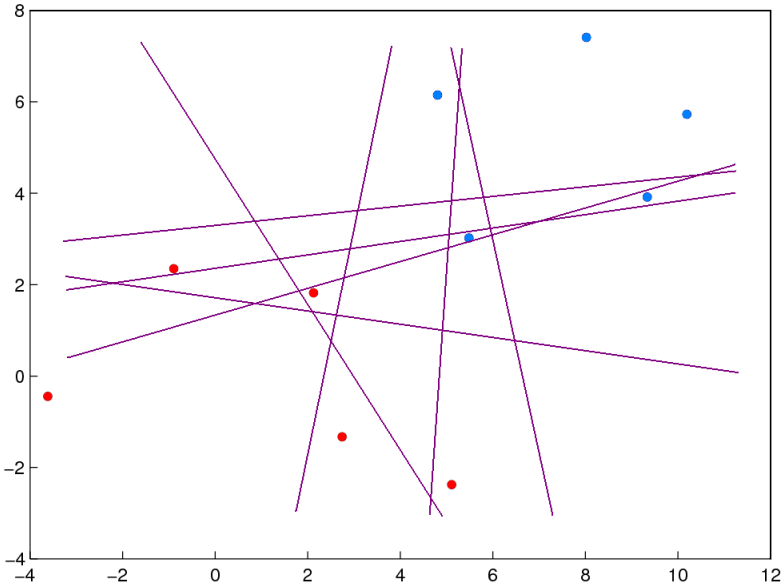
Пример. Семейство линейных алгоритмов классификации

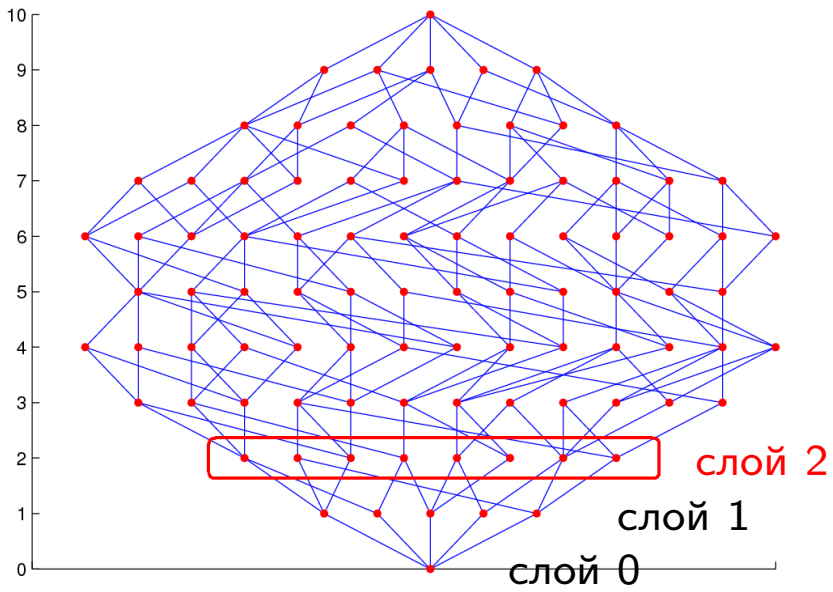




	слой 0	слой 1				
x_1	0	1	0	0	0	0
x_2	0	0	1	0	0	0
x_3	0	0	0	1	0	0
x_4	0	0	0	0	1	0
x_5	0	0	0	0	0	1
x_6	0	0	0	0	0	0
x_7	0	0	0	0	0	0
x_8	0	0	0	0	0	0
x_9	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0

Пример. Семейство линейных алгоритмов классификации





	слой 0	слой 1					слой 2								
x_1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	...
x_2	0	0	1	0	0	0	1	1	0	0	0	0	0	0	...
x_3	0	0	0	1	0	0	0	1	1	0	0	0	0	1	...
x_4	0	0	0	0	1	0	0	0	1	1	0	0	0	0	...
x_5	0	0	0	0	0	1	0	0	0	1	1	1	0	0	...
x_6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	...
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	...
x_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...

Характеристики расслоения и связности алгоритма

Множество объектов, соответствующих рёбрам, исходящим из a:

$$X_a = \left\{ x_{ab} \in \mathbb{X} \;\middle|\; a \prec b \right\},$$

Множество объектов, соответствующих всем рёбрам на путях, ведущих в a:

$$X'_a = \left\{ x \in \mathbb{X} \;\middle|\; \exists b \in \mathbb{A}\colon\ b < a,\ I(b,x) < I(a,x) \right\}.$$

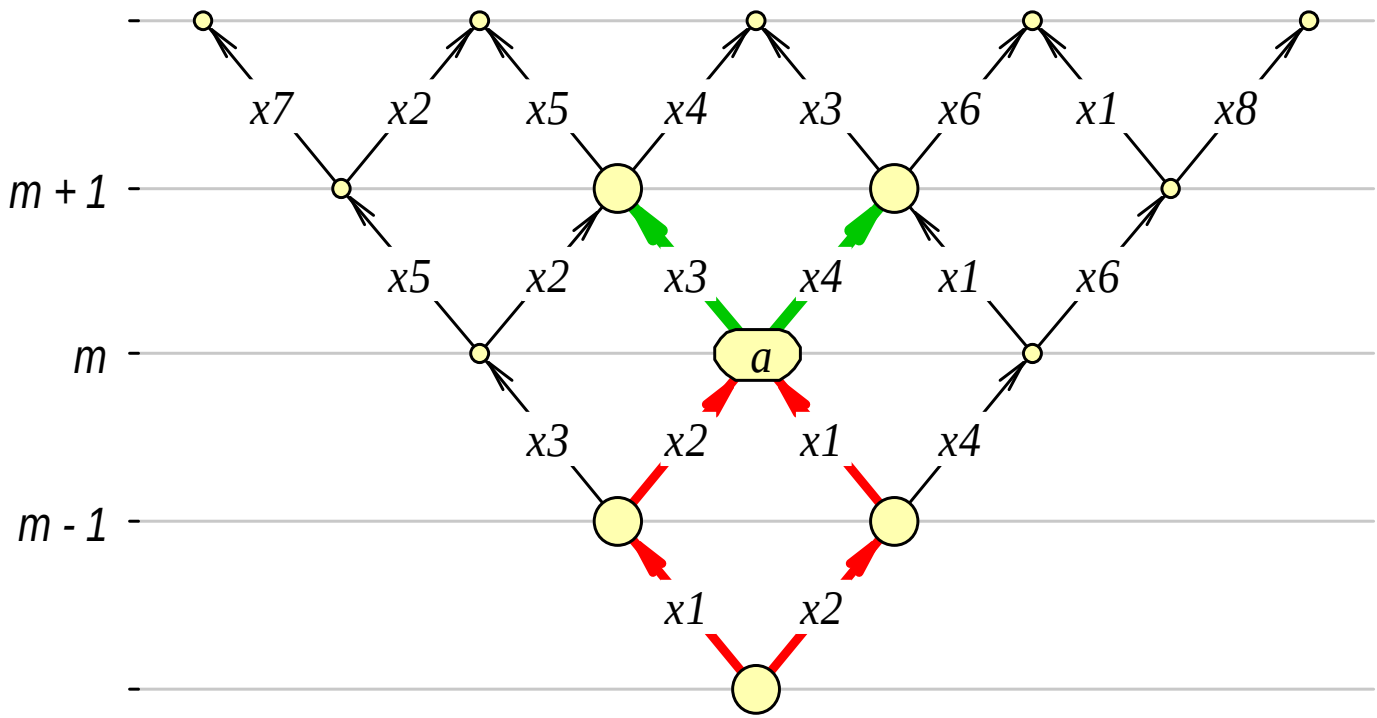
Опр. Характеристики расслоения и связности алгоритма a:

$u(a) = |X_a|$ — верхняя связность алгоритма a;

$q(a) = |X'_a|$ — неполноценность алгоритма a;

Пример: двумерная сеть алгоритмов

Верхняя связность: $u(a) = \#\{x3, x4\} = 2$;
Неполноценность: $q(a) = \#\{x1, x2\} = 2$;



Монотонные методы обучения

Опр. Метод обучения μ называется *монотонным*, если

$$\mu X = \arg\min_{a \in \mathbb{A}} K(a, X),$$

где $K(a, X)$ — строго монотонная функция вектора ошибок a:

$$\forall X \subset \mathbb{X},\ \forall a, b \in \mathbb{A} \quad a < b \ \rightarrow\ K(a, X) < K(b, X).$$

Пример 1. *Взвешенный эмпирический риск* с весами w_i:

$$K(a, X) = \sum_{x_i \in X} w_i\, I(a, x_i), \qquad w_i > 0.$$

Пример 2. Стандартные критерии информативности логических закономерностей (рассматриваются далее).

Верхняя оценка вероятности переобучения

Теорема (оценка расслоения–связности)

Пусть μ — монотонный метод обучения.
Тогда для любых $\mathbb{X}$, $\mathbb{A}$ и $\varepsilon \in (0,1)$

$$Q_\varepsilon \leqslant \sum_{a \in \mathbb{A}} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u,\, m-q}\left(\tfrac{\ell}{L}\left(m - \varepsilon k\right)\right)$$

где $u = |X_a|$ — *верхняя связность алгоритма* a,
$q = |X'_a|$ — *неполноценность алгоритма* a,
$m = m(a, X)$ — *число ошибок алгоритма* a,

$$\mathcal{H}_L^{\ell,\, m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}, \quad z = 0, \ldots, \ell$$

— *функция гипергеометрического распределения:*

Свойства оценки $Q_\varepsilon \leqslant \sum_{a\in\mathbb{A}} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u,\,m-q}\left(\frac{\ell}{L}\left(m-\varepsilon k\right)\right)$

1. Вклад алгоритма $a \in \mathbb{A}$ убывает экспоненциально
   по $u(a) \Rightarrow$ **связные семейства меньше переобучаются**;
   по $q(a) \Rightarrow$ **только нижние слои вносят вклад в Q_ε.**
2. Оценка обращается в равенство в случае многомерных монотонных сетей алгоритмов.
3. Вероятность получить алгоритм в результате обучения
$$\mathrm{P}\left[\mu X = a\right] \;\leqslant\; P_a = \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell}.$$
4. Если $q(a) > k$, то $P_a = 0$ и вклад алгоритма a равен 0
   $\Rightarrow$ при малых k оценка вырождается.
5. $\sum_{a\in\mathbb{A}} P_a$ — оценка степени завышенности.

Свойства оценки $Q_\varepsilon \leqslant \sum_{a\in\mathbb{A}} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u,\,m-q}\left(\frac{\ell}{L}\left(m-\varepsilon k\right)\right)$

6. При $|\mathbb{A}| = 1$ это оценка скорости сходимости частот в двух выборках (вариант закона больших чисел):

$$Q_\varepsilon = \mathcal{H}_L^{\ell,\,m}\left(\tfrac{\ell}{L}\left(m-\varepsilon k\right)\right) \to 0 \quad \text{при} \quad \ell, k \to \infty.$$

7. При $q = u = 0$ и $\ell = k$ это оценка Вапника-Червоненкиса:

$$Q_\varepsilon \;\leqslant\; \sum_{a\in\mathbb{A}} \mathcal{H}_L^{\ell,\,m}\left(\tfrac{\ell}{L}\left(m-\varepsilon k\right)\right) \;\leqslant\; |\mathbb{A}| \cdot \exp(-\varepsilon\ell^2).$$

8. При замене неполноценности q на нижнюю связность d это верхняя оценка *функционала равномерной сходимости*

$$\mathsf{P}\Big[\sup_{a\in\mathbb{A}}\big(\nu(a,\bar{X}) - \nu(a,X)\big) \geqslant \varepsilon\Big],$$

которая учитывает связность, но не учитывает расслоение.

Взвешенное голосование пороговых конъюнкций

Взвешенное голосование правил:

$$a(x) = \arg\max_{y \in Y} \sum_{r \in R_y} w_r r(x),$$

где $r\colon X \to \{0,1\}$ — правило, w_r — вес правила r,
R_y — множество правил класса y, Y — множество классов,

Будем выбирать R_y из **семейства пороговых конъюнкций** R:

$$r(x) = \bigwedge_{j \in J} \big[f_j(x) \leqslant \theta_j\big],$$

где $f_j(x)$ — числовые признаки, θ_j — пороги, $j = 1, \ldots, n$;
$J \subseteq \{1, \ldots, n\}$ — подмножество признаков, обычно $|J| \lesssim 7$.

Критерии информативности правил

Задача поиска закономерностей класса y в семействе правил R:

$$\begin{cases} n(r,X) = \#\{x_i \in X \mid r(x_i) = 1,\ y_i \neq y\} \to \min\limits_{r \in R}; \\ p(r,X) = \#\{x_i \in X \mid r(x_i) = 1,\ y_i = y\} \to \max\limits_{r \in R}. \end{cases}$$

Задача максимизации *информативности*: $\mathscr{H}(p,n) \to \max\limits_{r \in R}$.

Введём индикатор ошибки для правил класса y:

$$I(r,x_i) = \big[r(x_i) \neq [y_i{=}y]\big], \quad x_i \in \mathbb{X}.$$

Он индуцирует векторы ошибок и отношения $\leqslant$, $\prec$ на R.

Теорема

Если $\mathscr{H}(p,n)$ строго возрастает по p и строго убывает по n, то максимизация информативности $\mathscr{H}\big(p(r,X), n(r,X)\big)$ является монотонным методом обучения правил.

Примеры критериев информативности

$P = \left|\left\{x_i : y_i = y\right\}\right|$ — число «своих» в обучающей выборке;
$N = \left|\left\{x_i : y_i \neq y\right\}\right|$ — число «чужих» в обучающей выборке.

- точность: $\mathscr{H}(p,n) = \dfrac{p + N - n}{P + N}$;
- взвешенная точность: $\mathscr{H}(p,n) = p - \lambda n$;
- энтропийный критерий информационного выигрыша:

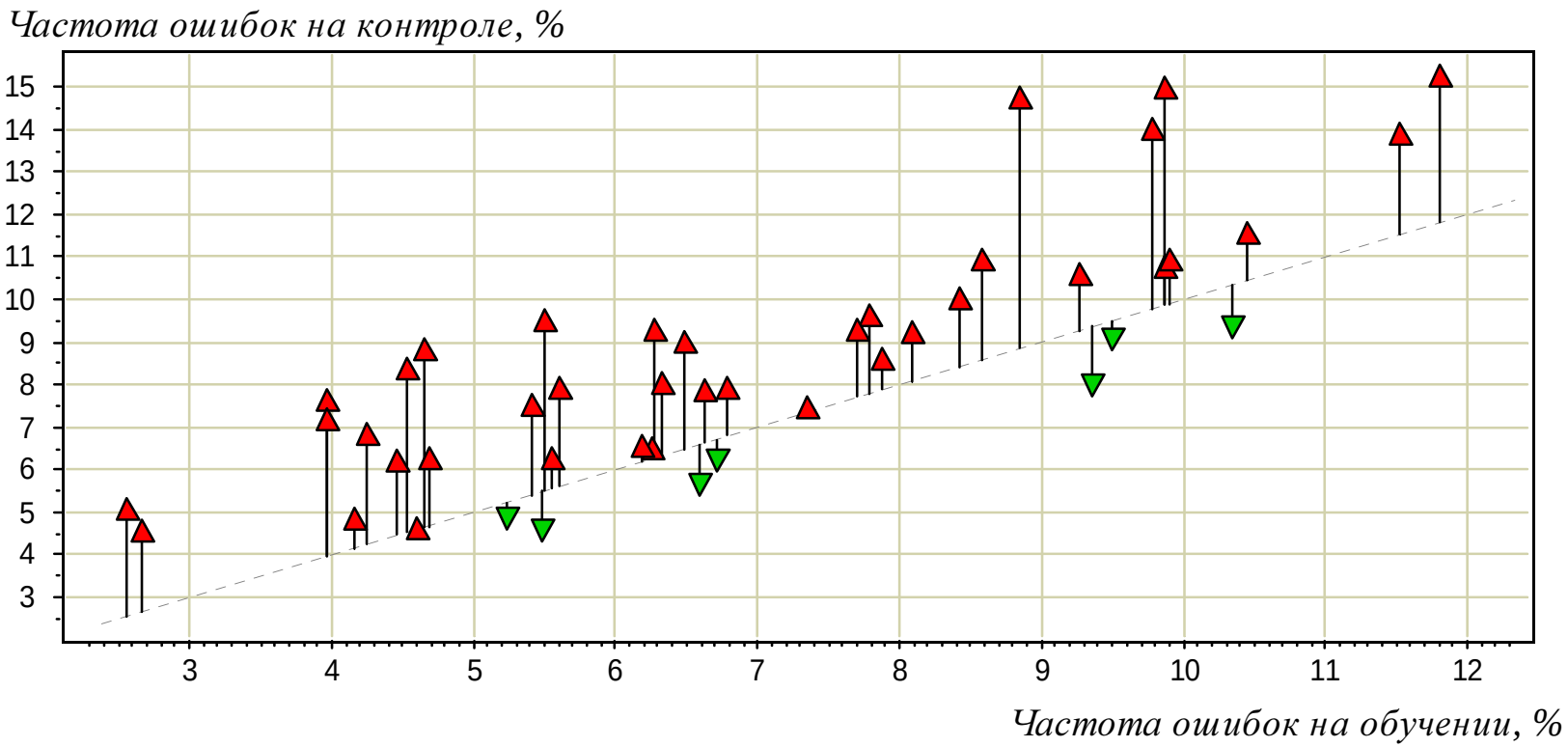
$$\mathscr{H}(p,n) = h\left(\frac{P}{\ell}\right) - \frac{p+n}{\ell}\, h\left(\frac{p}{p+n}\right) - \frac{\ell-p-n}{\ell}\, h\left(\frac{P-p}{\ell-p-n}\right),$$

  где $h(q) = -q\log_2 q - (1-q)\log_2(1-q)$;
- индекс Джини (Gini impurity): то же, при $h(q) = 2q(1-q)$;
- точный тест Фишера: $\mathscr{H}(p,n) = -\log C_P^p C_N^n / C_{P+N}^{p+n}$;
- критерий бустинга: $\mathscr{H}(p,n) = \sqrt{p} - \sqrt{n}$;

Проблема переобучения закономерностей

Общий недостаток всех критериев информативности — они не учитывают переобучение порогов θ_j и наборов J.

Пример. Задача предсказания результата хирургического лечения атеросклероза, $L = 98$, точки — закономерности.



Идея уменьшения переобучения закономерностей

План:

1. Вывести оценки расслоения–связности для метода обучения порогов θ при фиксированном наборе признаков J, отдельно для:
   — числа ошибок I рода $m' = P - p$ (на объектах класса y)
   — числа ошибок II рода $m'' = n$ (на остальных классах).
2. С их помощью получить оценки $\hat{p}$, $\hat{n}$ на неизвестной контрольной выборке по известным p, n на обучающей выборке.
3. Подставить эти оценки в критерий $\mathscr{H}$ и получить *критерий предсказанной информативности* $\mathscr{H}(\hat{p}, \hat{n})$.
4. По этому критерию отбирать подмножества признаков J.

Ошибки I и II рода относительно класса y

Два определения индикатора ошибки, числа и частоты ошибок:

I рода: $I'(r,x_i) = \big[r(x_i)=0\big]\big[y_i=y\big], \quad m' = P-p, \quad \nu';$

II рода: $I''(r,x_i) = \big[r(x_i)=1\big]\big[y_i\neq y\big], \quad m'' = n, \quad \nu''.$

Оценки вероятности переобучения для $r=\mu X$:

$$Q'_\varepsilon = \mathsf{P}\big[\nu'(r,\bar{X}) - \nu'(r,X) \geqslant \varepsilon\big] \;\leqslant\; \eta'_{\ell,k}(\varepsilon);$$

$$Q''_\varepsilon = \mathsf{P}\big[\nu''(r,\bar{X}) - \nu''(r,X) \geqslant \varepsilon\big] \;\leqslant\; \eta''_{\ell,k}(\varepsilon).$$

Обращение оценок: с вероятностью не менее $(1-\eta)$

$$\nu'(r,\bar{X}) \leqslant \nu'(r,X) + \varepsilon'_{\ell,k}(\eta),$$

$$\nu''(r,\bar{X}) \leqslant \nu''(r,X) + \varepsilon''_{\ell,k}(\eta).$$

Критерий предсказанной информативности:

$$\widehat{\mathscr{H}}(p,n) \;=\; \mathscr{H}\Big(\tfrac{k}{\ell}p - k\varepsilon'_{\ell,k}\big(\tfrac{1}{2}\big),\;\; \tfrac{k}{\ell}n + k\varepsilon''_{\ell,k}\big(\tfrac{1}{2}\big)\Big).$$

Оценка расслоения–связности для ошибок I и II рода

Теорема

Для монотонного метода обучения μ и любых $\mathbb{X}$, R, $\varepsilon \in (0,1)$

$$Q'_\varepsilon \leqslant \sum_{r\in R} \frac{C^{\ell-u}_{L-u-q}}{C^\ell_L} \mathcal{H}^{\ell-u,\,m'-q'}_{L-u-q}\left(\tfrac{\ell}{L}(m'-\varepsilon k)\right);$$

$$Q''_\varepsilon \leqslant \sum_{r\in R} \frac{C^{\ell-u}_{L-u-q}}{C^\ell_L} \mathcal{H}^{\ell-u,\,m''-q''}_{L-u-q}\left(\tfrac{\ell}{L}(m''-\varepsilon k)\right);$$

где $u = |X_r|$ — верхняя связность правила r,

$q = |X'_r|$, $q' = |X'_r \cap \mathbb{X}_y|$, $q'' = |X'_r \cap \mathbb{X}_{\bar{y}}|$ — неполноценность правила r относительно индикаторов ошибки I, I', I'',

$m' = m'(r,\mathbb{X})$, $m'' = m''(r,\mathbb{X})$ — число ошибок правила r относительно индикаторов ошибки I', I''.

Наблюдаемые и ненаблюдаемые оценки

Мы построили *критерий предсказанной информативности:*

$$\widehat{\mathscr{H}} = \mathscr{H}\left(\tfrac{k}{\ell}p(r,X) - k\varepsilon'_{\ell,k}\left(\tfrac{1}{2}\right),\ \ \tfrac{k}{\ell}n(r,X) + k\varepsilon''_{\ell,k}\left(\tfrac{1}{2}\right)\right).$$

Но нам нужен критерий, предсказывающий информативность на скрытой выборке $\bar{\mathbb{X}}$ длины K

$$\widehat{\mathscr{H}}_{\text{ненабл}} = \mathscr{H}\left(\tfrac{K}{L}p(r,\mathbb{X}) - K\varepsilon'_{L,K}\left(\tfrac{1}{2}\right),\ \ \tfrac{K}{L}n(r,\mathbb{X}) + K\varepsilon''_{L,K}\left(\tfrac{1}{2}\right)\right).$$

Однако мы не можем вычислить $\varepsilon'_{L,K}$ и $\varepsilon''_{L,K}$, т.к. $\bar{\mathbb{X}}$ скрыта.

Эвристика: заменив $\varepsilon'_{L,K}$ на $\varepsilon'_{\ell,k}$ и $\varepsilon''_{L,K}$ на $\varepsilon''_{\ell,k}$, получим *наблюдаемый критерий предсказанной информативности:*

$$\widehat{\mathscr{H}}_{\text{набл}} = \mathscr{H}\left(\tfrac{K}{L}p(r,\mathbb{X}) - K\varepsilon'_{\ell,k}\left(\tfrac{1}{2}\right),\ \ \tfrac{K}{L}n(r,\mathbb{X}) + K\varepsilon''_{\ell,k}\left(\tfrac{1}{2}\right)\right).$$

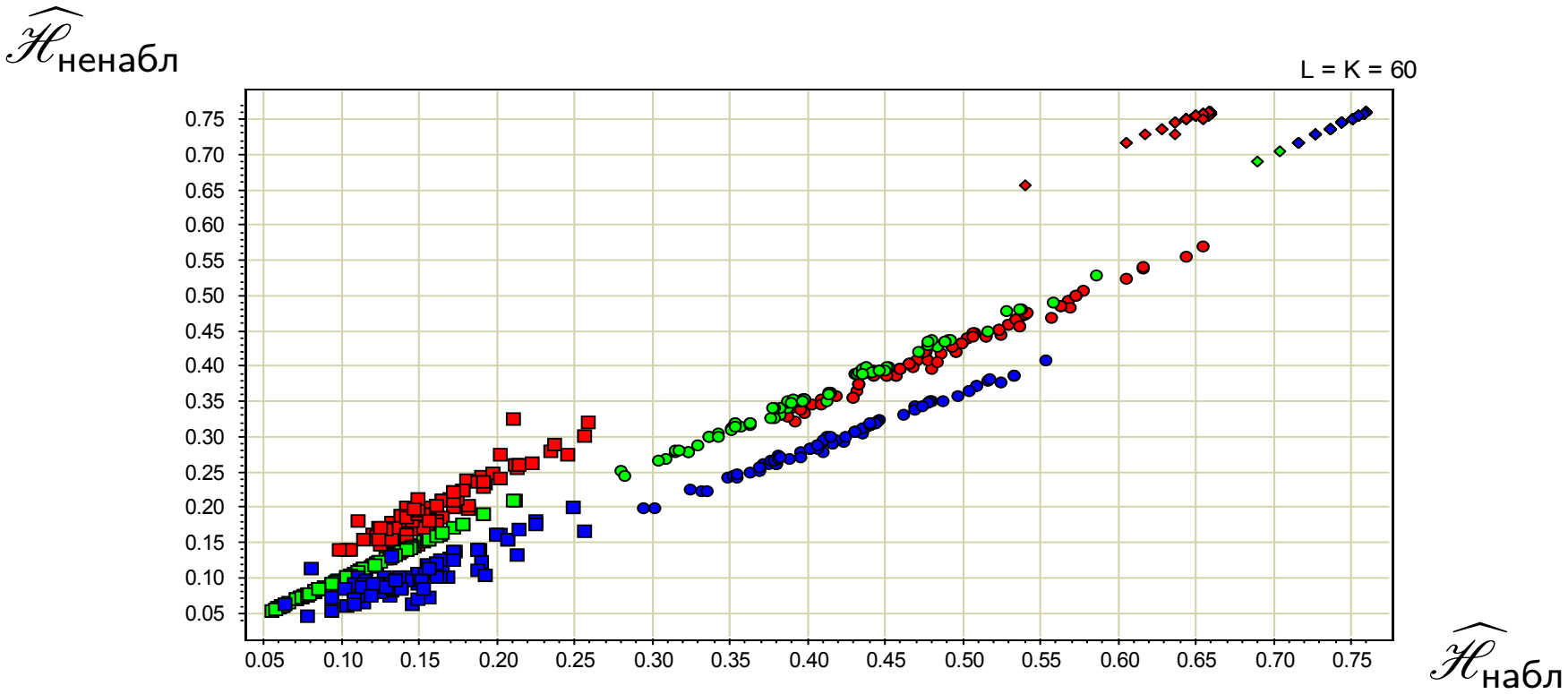
Почему мы имеем право на такую замену?

Эксперимент 1: зависимость $\widehat{\mathscr{H}}_{\text{ненабл}}$ от $\widehat{\mathscr{H}}_{\text{набл}}$

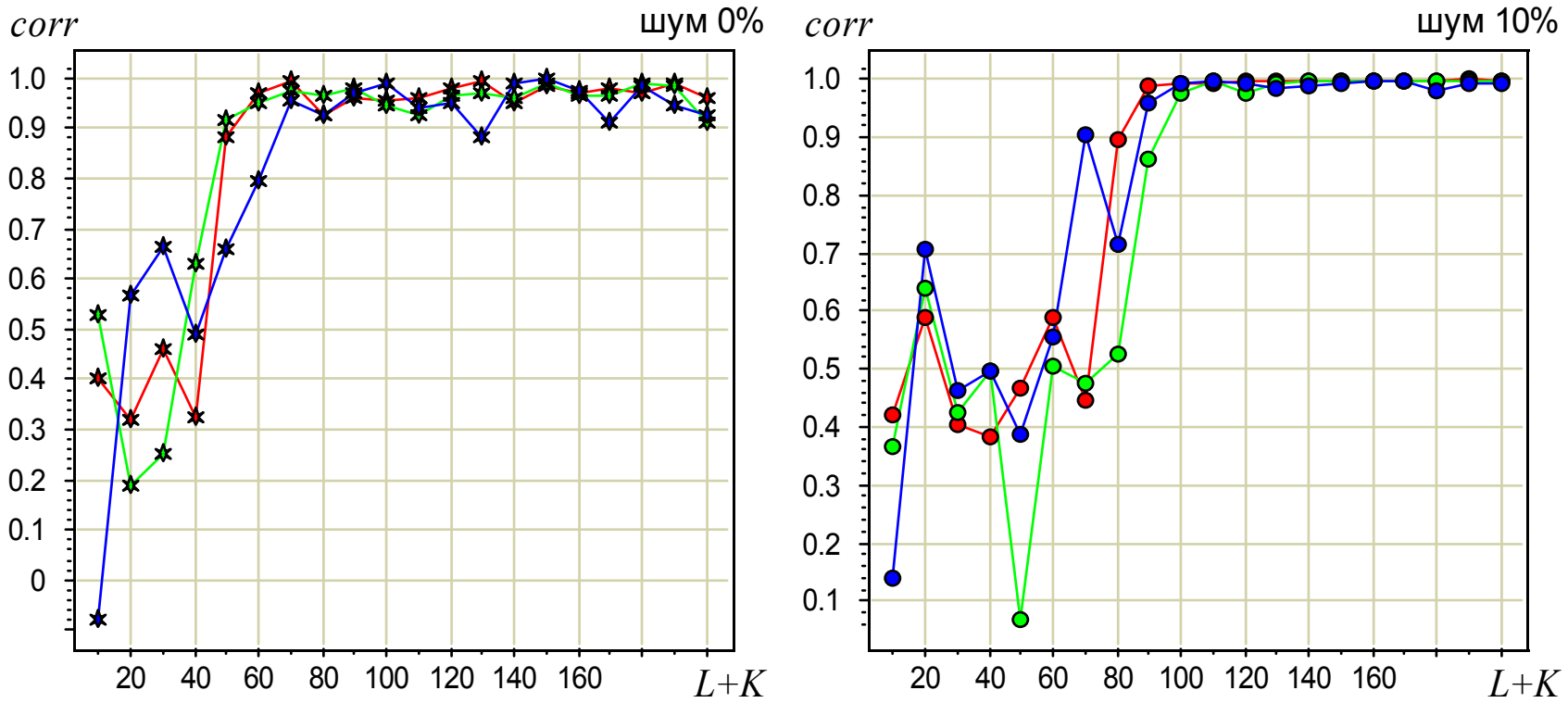
наблюдаемая выборка $\mathbb{X}$ | скрытая выборка $\bar{\mathbb{X}}$

наблюдаемые $\varepsilon'_{\ell,k}$, $\varepsilon''_{\ell,k}$

ненаблюдаемые $\varepsilon'_{L,K}$, $\varepsilon''_{L,K}$



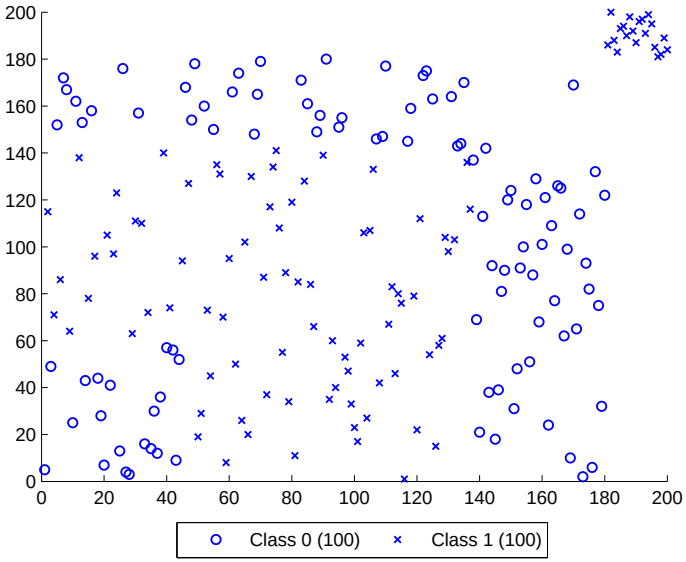
Эксперимент 1: зависимость корреляции $\widehat{\mathscr{H}}_{\text{ненабл}}$ и $\widehat{\mathscr{H}}_{\text{набл}}$ от длины выборки и зашумлённости данных



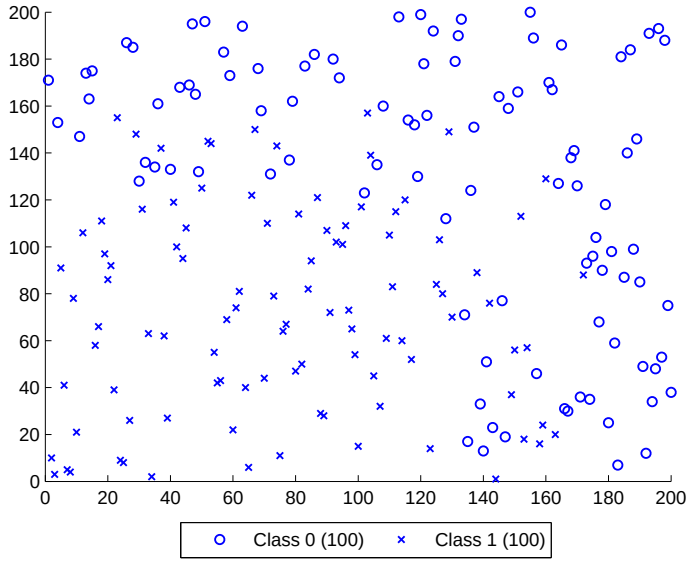
Экспериментальные результаты получены М. Дударенко на модельных данных с энтропийным критерием $\mathscr{H}$.

Эксперимент 2: в каких случаях предсказанная информативность даёт выигрыш?

выборка 1: $L = 200$,
шум вдали от границы классов



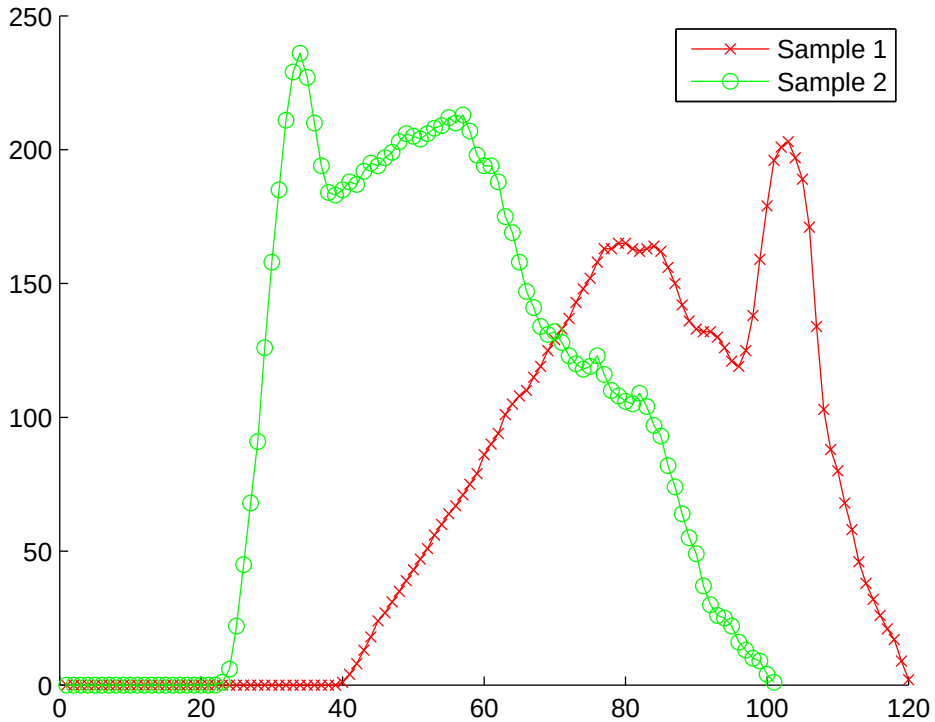
выборка 2: $L = 200$,
шум вдоль границы классов



	выборка 1	выборка 2
ошибок лучшего алгоритма, %	20,0	11,5
$\mathscr{H}$ обучение	40,3	40,5
$\mathscr{H}$ контроль	39,1	34,5
наблюдаемый $\widehat{\mathscr{H}}$	34,9	31,2

Эксперимент 2: Профили расслоения

Профили расслоения $D(m) = |A_m|$ выборок 1 и 2.



Выборка 2 содержит почти вдвое лучший алгоритм, однако найти его по случайной подвыборке практически невозможно из-за мощных нижних слоёв и переобучения.

Эксперимент 2: выводы

Выводы:

- Избавиться от переобучения порогов нельзя.
- Но можно предсказать, когда (для каких наборов признаков) оно будет меньше, проанализировав структуру графа расслоения–связности:
- когда профиль расслоение растёт медленнее,
- то есть когда граница классов наименее зашумлена.

Эксперимент 3: на реальных данных

Задачи из репозитория UCI:

задачи	объектов	признаков
australian	690	14
echo cardiogram	74	10
heart disease	294	13
hepatitis	155	19
labor relations	40	16
liver	345	6

Методы обучения:

- WV — weighted voting (boosting);
- DL — decision list;
- LR — logistic regression.

Методика тестирования: 10-кратный скользящий контроль.

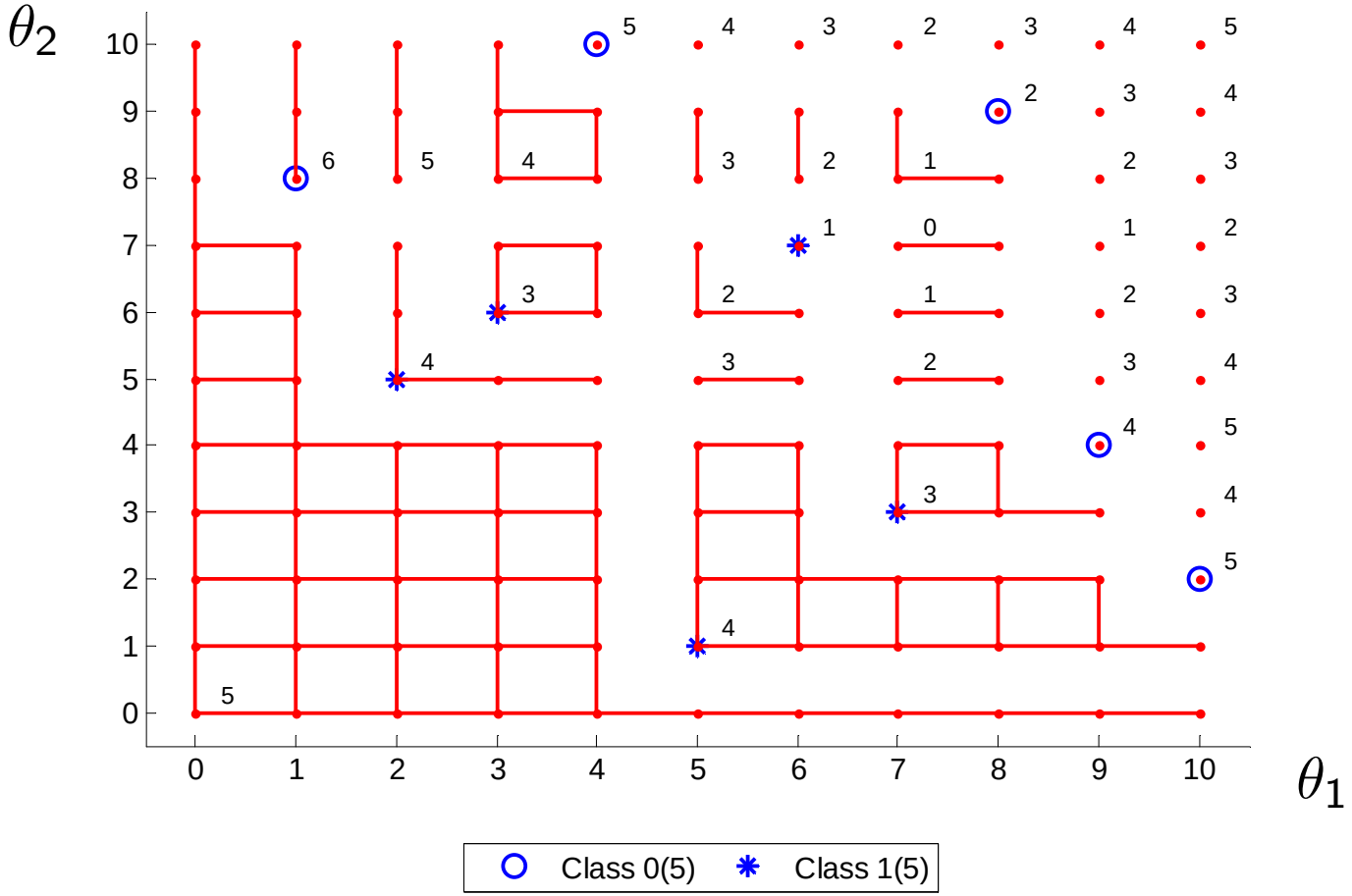
Результаты эксперимента на реальных данных

методы	задачи					
	austr	echo	heart	hepa	labor	liver
RIPPER-opt	15.5	2.97	19.7	20.7	18.0	32.7
RIPPER+opt	15.2	5.53	20.1	23.2	18.0	31.3
C4.5(Tree)	14.2	5.51	20.8	18.8	14.7	37.7
C4.5(Rules)	15.5	6.87	20.0	18.8	14.7	37.5
C5.0	14.0	4.30	21.8	20.1	18.4	31.9
SLIPPER	15.7	4.34	19.4	17.4	12.3	32.2
LR	14.8	4.30	19.9	18.8	14.2	32.0
WV	14.9	4.37	20.1	19.0	14.0	32.3
DL	15.1	4.51	20.5	19.5	14.7	35.8
WV+CS	14.1	3.2	19.3	18.1	13.4	30.2
DL+CS	14.4	3.6	19.5	18.6	13.6	32.3

По каждой задаче выделено два лучших результата.

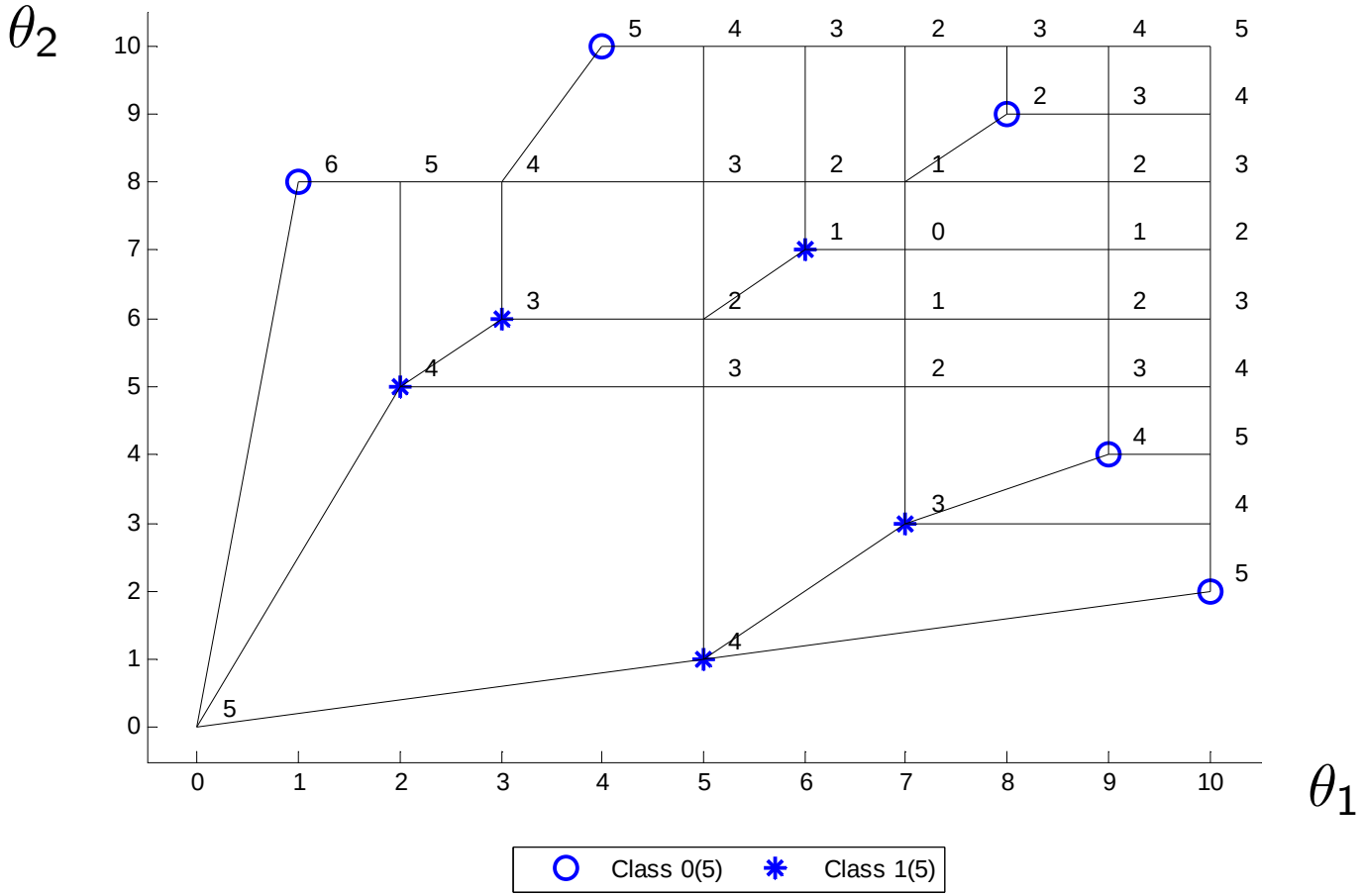
Классы эквивалентности правил (точки — правила)

Пример: разделимая 2-мерная выборка, $L = 10$, два класса.
правила: $r(x) = \left[f_1(x) \leqslant \theta_1\right] \cdot \left[f_2(x) \leqslant \theta_2\right]$.



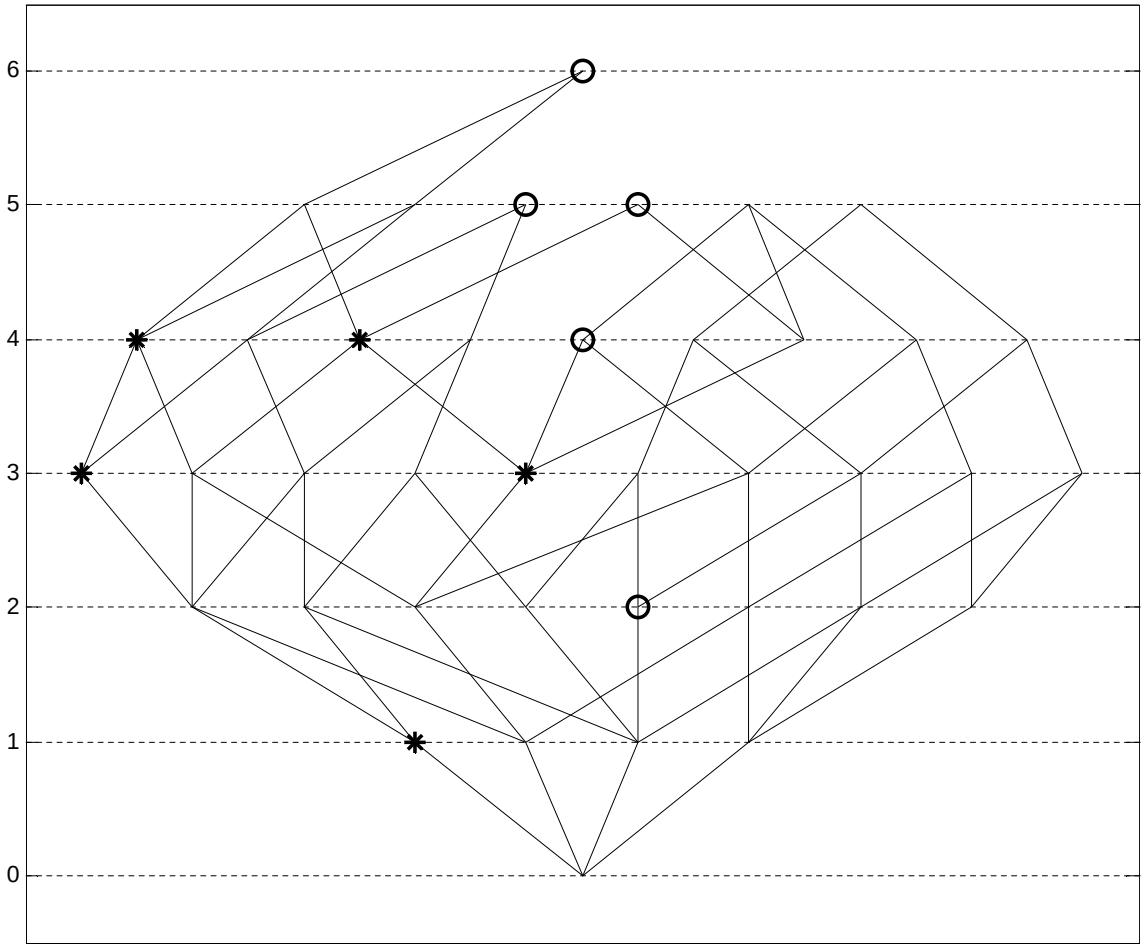
Классы эквивалентности правил (точки — классы)

Пример: разделимая 2-мерная выборка, $L = 10$, два класса.
правила: $r(x) = \left[f_1(x) \leqslant \theta_1 \text{ and } f_2(x) \leqslant \theta_2\right]$.



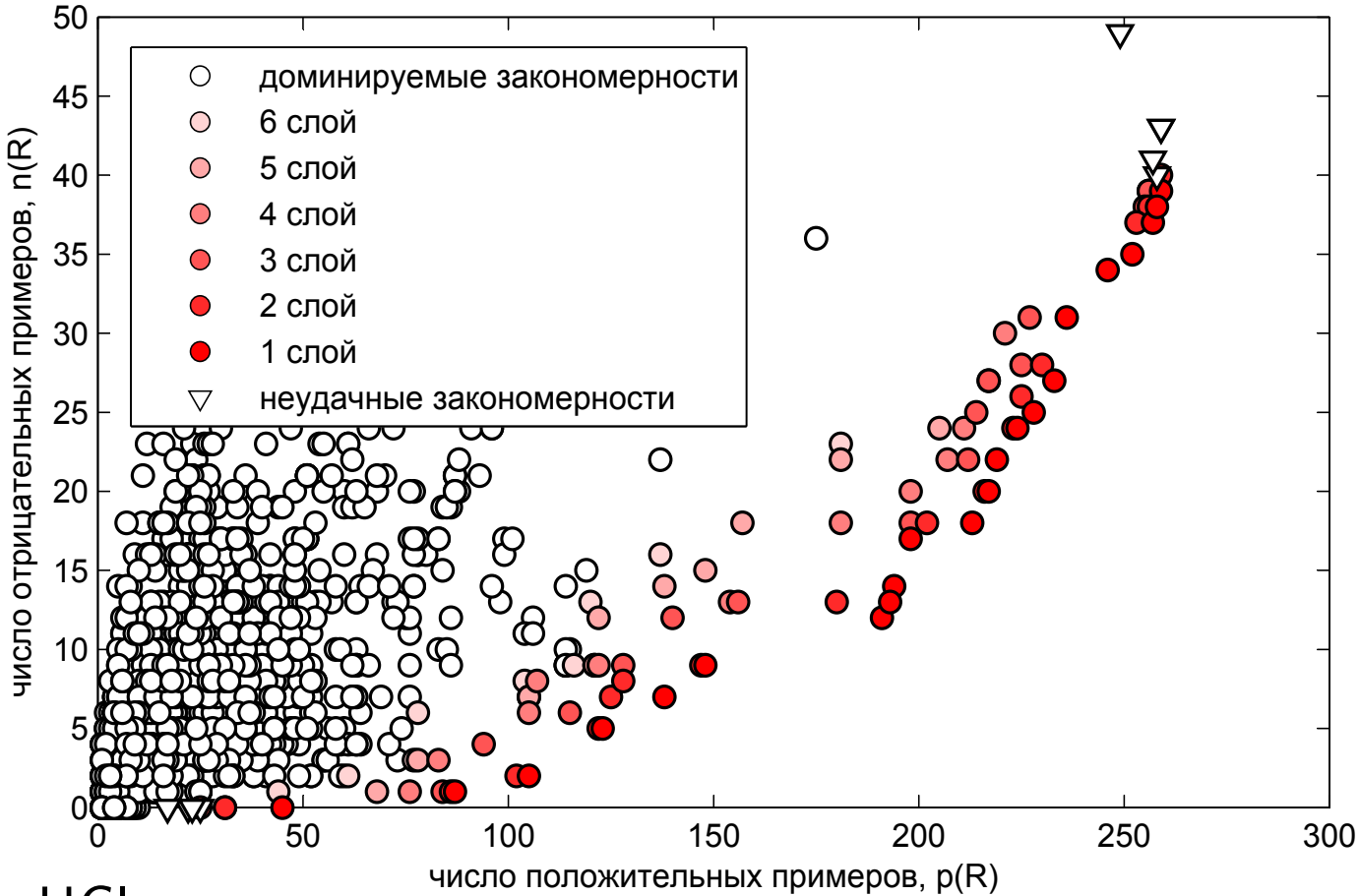
Классы эквивалентности правил (граф P-C)

Пример: Граф расслоения–связности, изоморфный графу классов эквивалентности правил с предыдущего слайда.



Парето-расслоение множества правил в плоскости p, n

Парето-фронт — множество недоминируемых закономерностей (точка (p, n) недоминируема, если правее и ниже точек нет)



задача UCI:german

Выводы

1. Комбинаторные оценки вероятности переобучения учитывают расслоение и связность семейства алгоритмов.
2. Для многомерных сетей алгоритмов эти оценки точные.
3. Встраивание этих оценок в критерий информативности уменьшает частоту ошибок логических алгоритмов классификации на 1–2%.

Новые результаты:

1. Оценка вероятности переобучения обобщена на класс *монотонных* методов обучения.
2. Получены раздельные оценки для ошибок I и II рода и предложен критерий предсказанной информативности.
3. Обоснована (пока только экспериментально) замена ненаблюдаемой оценки наблюдаемой.

Спасибо за внимание!

Воронцов Константин Вячеславович
vokov@forecsys.ru

Страницы на www.MachineLearning.ru:

- Участник:Vokov
- Слабая вероятностная аксиоматика
- Расслоение и сходство алгоритмов (виртуальный семинар)