Явления расслоения и сходства в семействах алгоритмов и их влияние на обобщающую способность

К. В. Воронцов
(vokov@forecsys.ru, http://www.ccas.ru/voron)

Вычислительный Центр им. А. А. Дородницына РАН

Распознавание образов и анализ изображений:
новые информационные технологии
14–20 сентября 2008
Нижний Новгород

Содержание

Определения и обозначения

Обучающая выборка: $X^\ell = \{x_i\}_{i=1}^\ell \subset \mathbb{X}$.

Метод обучения μ: $X^\ell \mapsto a$, где $a \in A$ — *алгоритм*.

Бинарная функция потерь

$I(a,x) = [\text{алгоритм } a \text{ допускает ошибку на объекте } x]$.

Бинарный вектор ошибок алгоритма a на выборке X^ℓ:

$\vec{a}(X^\ell) = \big(I(a,x_i)\big)_{i=1}^\ell$.

Частота ошибок алгоритма a на выборке X^ℓ

$$\nu(a,X^\ell) = \tfrac{1}{\ell}\sum_{i=1}^{\ell} I(a,x_i).$$

Тестовая выборка: $X^k = \{x_i\}_{i=1}^k \subset \mathbb{X}$.

Переобученность алгоритма $\mu(X^\ell)$ относительно X^ℓ, X^k:

$\delta(\mu, X^\ell, X^k) = \nu\big(\mu(X^\ell), X^k\big) - \nu\big(\mu(X^\ell), X^\ell\big)$.

Задача: поточнее оценить сверху *вероятность переобучения*

$\mathsf{P}_{X^\ell,X^k}\big\{\delta(\mu, X^\ell, X^k) > \varepsilon\big\} \leqslant \eta(\varepsilon), \quad \eta(\varepsilon) - ?$

Оценка по тестовой выборке (test set bound)

Теорема (вариант Закона Больших Чисел)

Для любого фиксированного алгоритма a, любой вероятностной меры P *на* $X^L = X^\ell \cup X^k$ *наблюдаемая частота* $\nu\big(a, X^\ell\big)$ *предсказывает неизвестную частоту* $\nu\big(a, X^k\big)$*:*

$$\mathsf{P}_n\big\{\delta(a, X_n^\ell, X_n^k) \geqslant \varepsilon\big\} \leqslant H_L^\ell(\varepsilon),$$

$H_L^\ell(\varepsilon) = \max\limits_{m=0,\ldots,L} \sum\limits_{t=s_0}^{s(\varepsilon)} \dfrac{C_m^t C_{L-m}^{\ell-t}}{C_L^\ell}$ *— верхняя оценка левого хвоста гипергеометрического распределения,*
$s_0 = (m-k)_+$, $s(\varepsilon) = \big\lfloor \frac{\ell}{L}(m - \varepsilon k) \big\rfloor$.

- ⊕ Оценка достаточно точна
- ⊖ Но не даёт рекомендаций для построения μ.

Оценки Вапника-Червоненкиса [1968–1971]

Для любого семейства A и любой вероятностной меры P

$$\mathsf{P}_{X^L}\{\delta(\mu, X^\ell, X^k) > \varepsilon\} \leqslant \mathsf{P}_{X^L}\Big\{\sup_{a\in A} \delta(a, X^\ell, X^k) > \varepsilon\Big\} \leqslant$$

$$\leqslant \sum_{\vec{a}\in A(X^L)} \mathsf{P}_{X^L}\{\delta(a, X^\ell, X^k) > \varepsilon\} \leqslant \Delta^A(L)\cdot H_L^\ell(\varepsilon) \leqslant$$

$$(\text{if } \ell = k) \leqslant \Delta^A(L)\cdot 1.5\, e^{-\varepsilon^2\ell},$$

$A(X^L) = \{\vec{a}(X^L) \mid a \in A\}$ — мн-во векторов ошибок A на X^L.

$\Delta^A(L) = \max\limits_{X^L} |A(X^L)|$ — коэффициент разнообразия (*shatter coefficient*) семейства A,

$\Delta^A(L) \leqslant 1.5\,\frac{L^h}{h!}$, где h — *ёмкость* (*VC-dimension*) семейства A.

⊕ Это приводит к методу структурной минимизации риска

⊖ Но оценка сильно завышена и на практике неприменима

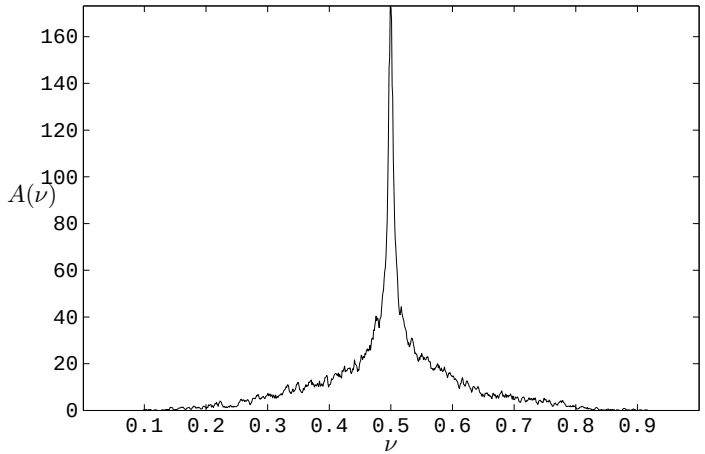
Причины завышенности

- Оценка «худшего случая» не учитывает:
  - особенности данных X^L
    (в том числе особенности целевой зависимости);
  - особенности метода обучения μ.
- *Эффект расслоения* (или *локализации*) семейства:
  чем хуже алгоритм, тем меньше шансы его получить.
  Семейство A *расслаивается* в каждой задаче по-разному.
- *Неравенство Буля* (union bound)
  $$P(S_1 \cup \cdots \cup S_\Delta) \leqslant P(S_1) + \cdots + P(S_\Delta),$$
  тем сильнее завышено, чем более схожи события
  $$S_d = \{\delta(a_d, X^\ell, X^k) > \varepsilon\}.$$
- Экспоненциальная аппроксимация $e^{-\varepsilon^2 \ell}$ завышена.

40 лет спустя проблема остаётся открытой

- Равномерная сходимость [Vapnik, Chervonenkis, 1968]
- Theory of learnable (PAC-learning) [Valiant, 1982]
- Data-dependent bounds [Haussler, 1992; Bartlett, 1998]
- Connected function classes [Sill, 1995]
- Similar classifiers VC bounds [Bax, 1997]
- Self-bounding learning algorithms [Freund, 1998]
- Microchoice bounds [Langford, Blum, 2001]
- Algorithmic stability [Bousquet, Elisseeff, 2002]
- Algorithmic luckiness [Herbrich, Williamson, 2002]
- Shell bounds [Langford, 2002]
- PAC-Bayes bounds [McAllester, 1999; Langford, 2005]

Оценки расслоения (shell bounds)



- Большинство алгоритмов $a \in A$ имеют $\nu(a, X^L) = 0.5$
- Только алгоритмы из левого хвоста гистограммы имеют шанс быть выбранными при $\nu(a, X_n^\ell) \to \min\limits_{a \in A}$
- Оценка сложно вычисляется, требует моделирования методом Монте-Карло, на практике не слишком точна.

John Langford. Quantitatively Tight Sample Complexity Bounds. PhD (Carnegie Mellon). 2002.

ВЧ-оценки, учитывающие сходство алгоритмов

Теорема

Пусть множество векторов ошибок $\{\vec{a}(X^L) \mid a \in A\}$ *кластеризуется по расстоянию Хэмминга на* $S(r)$ *кластеров радиуса* r *каждый. Тогда*

$$\mathsf{P}_{X^L}\{\delta(a, X^\ell, X^k) > \varepsilon + r/\ell\} \leqslant S(r) \cdot H_L^\ell(\varepsilon).$$

- Если A линейно, то $S(r) = \Delta^A(L)/(2r+1)$.
- Возможность оптимизации по r (открытая проблема: как r зависит от размерности пространства X?)
- Оценка не точна, даже после оптимизации по r.

Bax E. Similar Classifiers and VC Error Bounds. CalTech-CS-TR97-14, June 1997. `citeseer.ist.psu.edu/bax97similar.html`

Связные семейства алгоритмов

Определение

Семейство A связно, если $\forall\, \vec{a} \in A(X^L)$ *с вероятностью 1* $\exists\, \vec{a}_1 \in A(X^L)$ *на хэмминговом расстоянии* $\|\vec{a} - \vec{a}_1\| = 1$.

- SVM, двухслойная ИНС, RBF, и т. д. — связны.
- **Теорема:** *Если A связно, то*

$$\mathsf{P}_{X^L}\left\{\delta(a, X^\ell, X^k) > \varepsilon\right\} \leqslant \frac{1}{\sqrt{\pi L}} \Delta^A(L) \cdot H_L^\ell(\varepsilon).$$

- Оценка не точна, лишь немногим отличаясь от ВЧ-оценки.

Sill J. Generalization Bounds for Connected Function Classes. 1995. `http://citeseer.ist.psu.edu/127284.html`

Sill J. Monotonicity and Connectedness in Learning Systems. PhD thesis, CalTech, 1998.

Зачем нужно измерение факторов завышенности?

- **Конечная цель (ОТКРЫТАЯ ПРОБЛЕМА)**
  получить точные (практичные) оценки.
- **Промежуточная цель (ВЫПОЛНЕНА — см. ниже)**
  понять причины завышенности, оценив их вклад количественно в экспериментах на реальных данных
- **Проблема:**
  Стандартная вероятностная техника приводит к цепочке оценок, завышенность которых не поддаётся анализу
- **Почему так происходит?**
  слишком много промежуточных оценок;
  вводятся вспомогательные трудно оцениваемые величины
- **Что предлагается:**
  не использовать асимптотики и грубые неравенства;
  использовать только легко измеряемые величины

Слабая (комбинаторная) вероятностная аксиоматика

1. $X^L = \{x_i\}_{i=1}^L$ — заданная конечная выборка.
2. Все разбиения $X^L = X_n^\ell \cup X_n^k$ *равновероятны*,
   где $n = 1, \ldots, N$, $N = C_L^k$, $L = \ell + k$;
   X_n^ℓ — наблюдаемая обучающая выборка;
   X_n^k — скрытая тестовая выборка.

Переобученность при n-м разбиении: $\delta_n(\mu) \equiv \delta(\mu, X_n^\ell, X_n^k)$.

Вероятность переобучения определяется через долю разбиений:

$$\mathsf{P}_n\{\delta_n(\mu) > \varepsilon\} = \frac{1}{N}\sum_{n=1}^{N}[\delta_n(\mu) > \varepsilon].$$

Замечание. Понятие «вероятность» вводится без теории меры и без предельного перехода $L \to \infty$.

Преимущества слабой аксиоматики

- Не избыточна.
- Может давать точные не асимптотические оценки.
- Вероятности легко измеряются эмпирически:

$$\hat{\mathsf{P}}_n\{\delta_n > \varepsilon\} = \frac{1}{|N'|}\sum_{n\in N'}[\delta_n > \varepsilon] \overset{N'\to\infty}{\to} \mathsf{P}_n\{\delta_n > \varepsilon\}.$$

- Возврат в колмогоровскую аксиоматику тривиален:

  если $\mathsf{P}_n\{\delta(X_n^\ell, X_n^k) > \varepsilon\} \leqslant \eta(\varepsilon, X^L)$,

  то $\mathsf{P}_{X^L}\{\delta(X^\ell, X^k) > \varepsilon\} \leqslant \mathsf{E}_{X^L}\eta(\varepsilon, X^L)$.

- Достаточна для вывода:
  - закона больших чисел (точная оценка);
  - критерия Колмогорова-Смирнова (точная оценка);
  - многих непараметрических (порядковых) критериев;
  - оценок Вапника-Червоненкиса (см. далее);

Оценка по тестовой выборке (test set bound)

Рассмотрим фиксированный алгоритм a, $\nu(a, X^L) = m/L$.

Теорема (точная оценка)

Наблюдаемая частота $\nu\left(a, X^\ell\right)$ *предсказывает скрытую частоту* $\nu\left(a, X^k\right)$*:*

$$\mathsf{P}_n\left\{\delta(a, X_n^\ell, X_n^k) \geqslant \varepsilon\right\} = H_L^{\ell,m}\left(s(\varepsilon)\right),$$

где $H_L^{\ell,m}(s) = \sum\limits_{t=s_0}^{s} \dfrac{C_m^t C_{L-m}^{\ell-t}}{C_L^\ell}$ *— левый хвост ГГР (гипергеометрического распределения);*

$s(\varepsilon) = \left\lfloor \frac{\ell}{L}(m - \varepsilon k) \right\rfloor$; $s_0 = \max\{0, m - k\}$.

Оценка Вапника-Червоненкиса

Для любого метода обучения μ и любой выборки X^L:

$$Q_\varepsilon = \mathsf{P}_n\big\{\delta(a_n, X_n^\ell, X_n^k) > \varepsilon\big\} \leqslant$$
$$\leqslant \sum_{m=1}^{L} D_m \cdot H_L^{\ell,m}(s(\varepsilon)) \leqslant$$
$$\leqslant \Delta_L^\ell \cdot H_L^\ell(\varepsilon) \overset{\ell=k}{\leqslant} \Delta^A(L) \cdot 1.5\, e^{-\varepsilon^2 \ell};$$

$\Delta_L^\ell(\mu, X^L)$ — *локальный коэфф. разнообразия* (local shatter coefficient) множества алгоритмов $\big\{a_n = \mu(X_n^\ell) \;\big|\; n = 1, \ldots, N\big\}$;

$D_m(\mu, X^L)$, $m = 0, \ldots, L$ — *профиль разнообразия* — последовательность коэффициентов разнообразия множеств алгоритмов, допускающих m ошибок на X^L:
$\big\{a_n = \mu(X_n^\ell) \;\big|\; \nu(a_n, X^L) = \frac{m}{L},\ n = 1, \ldots, N\big\}$.

Эффективный локальный коэффициент разнообразия

Расслоение семейства A (частично) учтено, но

- не ясно, как оценивать D_m;
- не ясно, даст ли это существенный выигрыш в точности.

Идея: оценить причины завышенности эмпирически

Определение

Эффективный локальный коэффициент разнообразия (ЭЛКР):

$$\hat{\Delta}_L^\ell(\varepsilon) = \frac{\hat{\mathsf{P}}_n\{\delta(a_n, X_n^\ell, X_n^k) > \varepsilon\}}{H_L^{\ell,m}(s(\varepsilon))} = \frac{\hat{\mathsf{P}}_n\{\delta(a_n, X_n^\ell, X_n^k) > \varepsilon\}}{\hat{\mathsf{P}}_n\{\delta(a, X_n^\ell, X_n^k) > \varepsilon\}}.$$

Факторы завышенности оценок Вапника-Червоненкиса

Степень завышенности раскладывается в произведение четырёх факторов:

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

где $H = \max\limits_m H^{\ell,m}_L\left(s(\varepsilon)\right)$.

Причины завышенности:

- $r_1 \geqslant 1$: пренебрегли расслоением
- $r_2 \geqslant 1$: пренебрегли сходством (из-за union bound)
- $r_3 \geqslant 1$: оценили профиль разнообразия сверху константой
- $r_4 \geqslant 1$: взяли экспоненциальную аппроксимацию ГГР

Алгоритм поиска логических закономерностей

- *Закономерность* — это предикат $\phi_y : X \to \{0, 1\}$, который выделяет преимущественно объекты класса y.
- *Взвешенное голосование* закономерностей:

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

где $\phi_y^t(x)$ — t-я закономерность класса y, w_y^t — её вес.

- *Метод обучения закономерностей* класса y:
$\mu_y : X^\ell \mapsto \left\{ \phi_y^t(x) \;\middle|\; t = 1, \ldots, T_y \right\}$.

- Почему логические алгоритмы удобны для оценивания завышенности ВЧ-оценок:
  - КР (коэффициент разнообразия) $\Delta^A(L)$ известен;
  - ЛКР $\Delta_L^\ell(\mu, X^L)$ легко оценивается снизу;
  - ЭЛКР $\hat{\Delta}_L^\ell(\varepsilon)$ легко оценивается.

Эксперимент

- 7 задач классификации на два класса (из репозитория UCI)
- 20×2-кратный скользящий контроль, $\ell = k$
- Логический алгоритм Forecsys LogicPro®
  [Воронцов, Кочедыков, Ивахненко]

Задача	L	n	средняя ошибка не тестовых данных				
			C4.5	C5.0	RIPPER	SLIPPER	LogicPro
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 — объём полной выборки; n — число признаков.

Результаты эксперимента

Причины завышенности ВЧ-оценок

(пороги $\varepsilon_0, \varepsilon_1, \varepsilon_2$ соответствуют надёжности $\hat{Q}_\varepsilon = 0.05,\ 0.1,\ 0.01$).

Задача	y	r_1	$r_2(\varepsilon_0)$	$r_3(\varepsilon_0)$	$r_4(\varepsilon_0)$	$\hat{\Delta}^\ell_L[\varepsilon_1, \varepsilon_2]$	$\hat{\Delta}^\ell_L(\varepsilon_0)$
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

Выводы

- Коэффициент разнообразия $\hat{\Delta}_L^\ell$ должен был бы принимать значения порядка 10^2 и меньше, чтобы оценка не была завышена. Известные теории не дают таких оценок.
- *Эффективная локальная ёмкость* (если бы мы её определяли) вырождается в 1.
  Открытая проблема №1:
  Как ввести «правильную» характеристику размерности?
- Не имеет смысла оценивать профиль разнообразия D_m.
- **Открытая проблема №2 (towards tighter bounds):**
  Как учесть одновременно и *расслоение* и *сходство*?

Vorontsov K. V. Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. — 2008. — Vol. 18, no. 2. — Pp. 243–259.

«Игрушечный пример»: пара алгоритмов

Пусть алгоритмы a_1, a_2 допускают m_1, m_2 ошибок на X^L:

$$\vec{a}_1(X^L) = (\overbrace{1\,1\,1\;1\,1\,1\,1\,1}^{m_1}\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0);$$
$$\vec{a}_2(X^L) = (0\,0\,0\,\underbrace{1\,1\,1\,1\,1\;1\,1\,1\,1\,1}_{m_2}\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0).$$

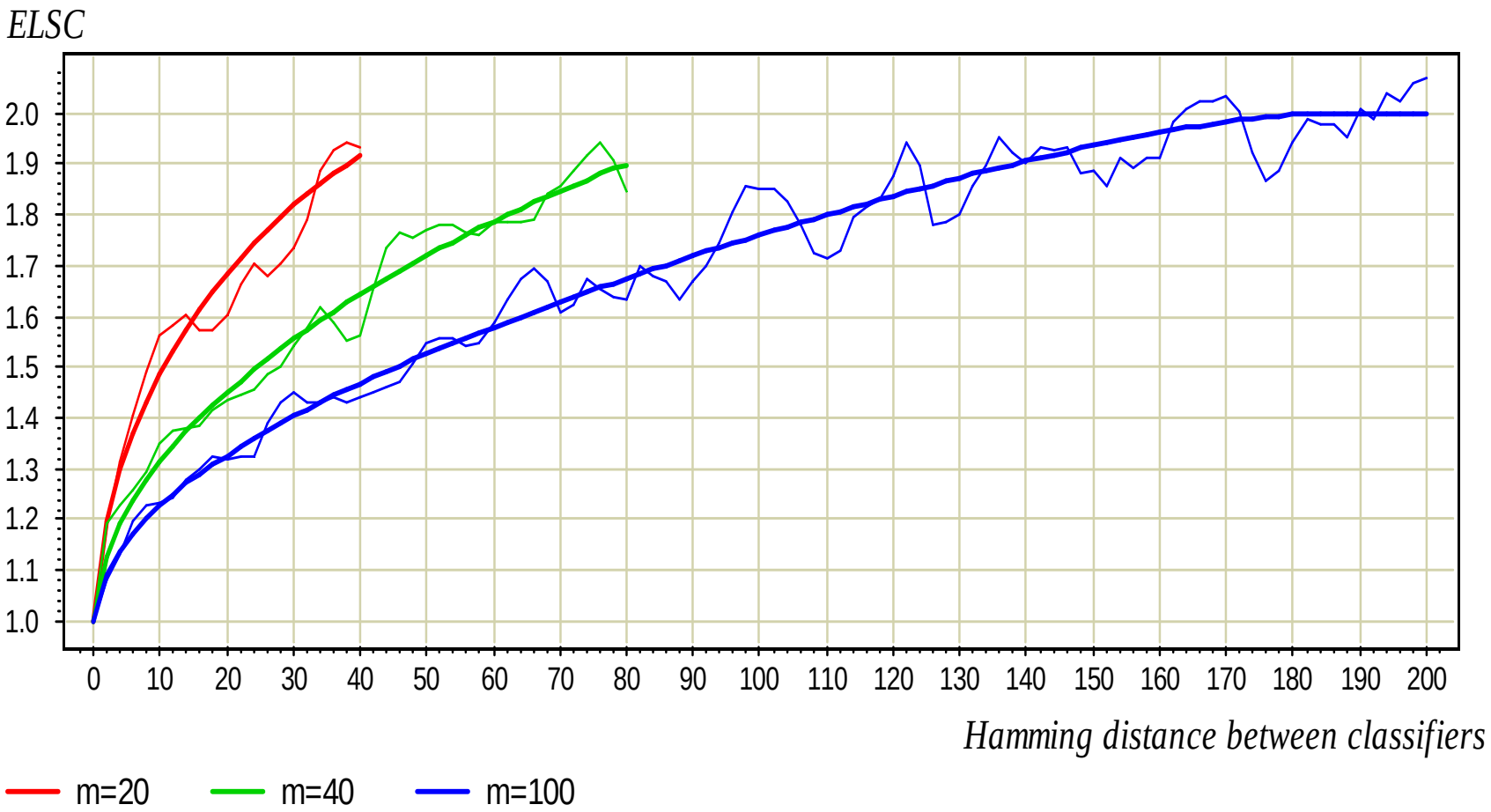
m_0

Теорема (точная оценка вероятности переобучения)

$$\mathsf{P}_n\{\delta(\mu, X_n^\ell, X_n^k) \geqslant \varepsilon\} = \sum_{s_0=0}^{m_0}\sum_{s_1=0}^{m_1}\sum_{s_2=0}^{m_2} \frac{C_{m_0}^{s_0} C_{m_1}^{s_1} C_{m_2}^{s_2} C_{L-m_0-m_1-m_2}^{\ell-s_0-s_1-s_2}}{C_L^\ell} \times$$
$$\times \big[m_0+m_1+m_2-k \leqslant s_0+s_1+s_2 \leqslant \ell\big] \times$$
$$\times \Big(\big[s_1 < s_2\big]\big[s_0+s_1 \leqslant \tfrac{\ell}{L}(m_0+m_1-\varepsilon k)\big] + $$
$$+ \big[s_1 \geqslant s_2\big]\big[s_0+s_2 \leqslant \tfrac{\ell}{L}(m_0+m_2-\varepsilon k)\big]\Big).$$

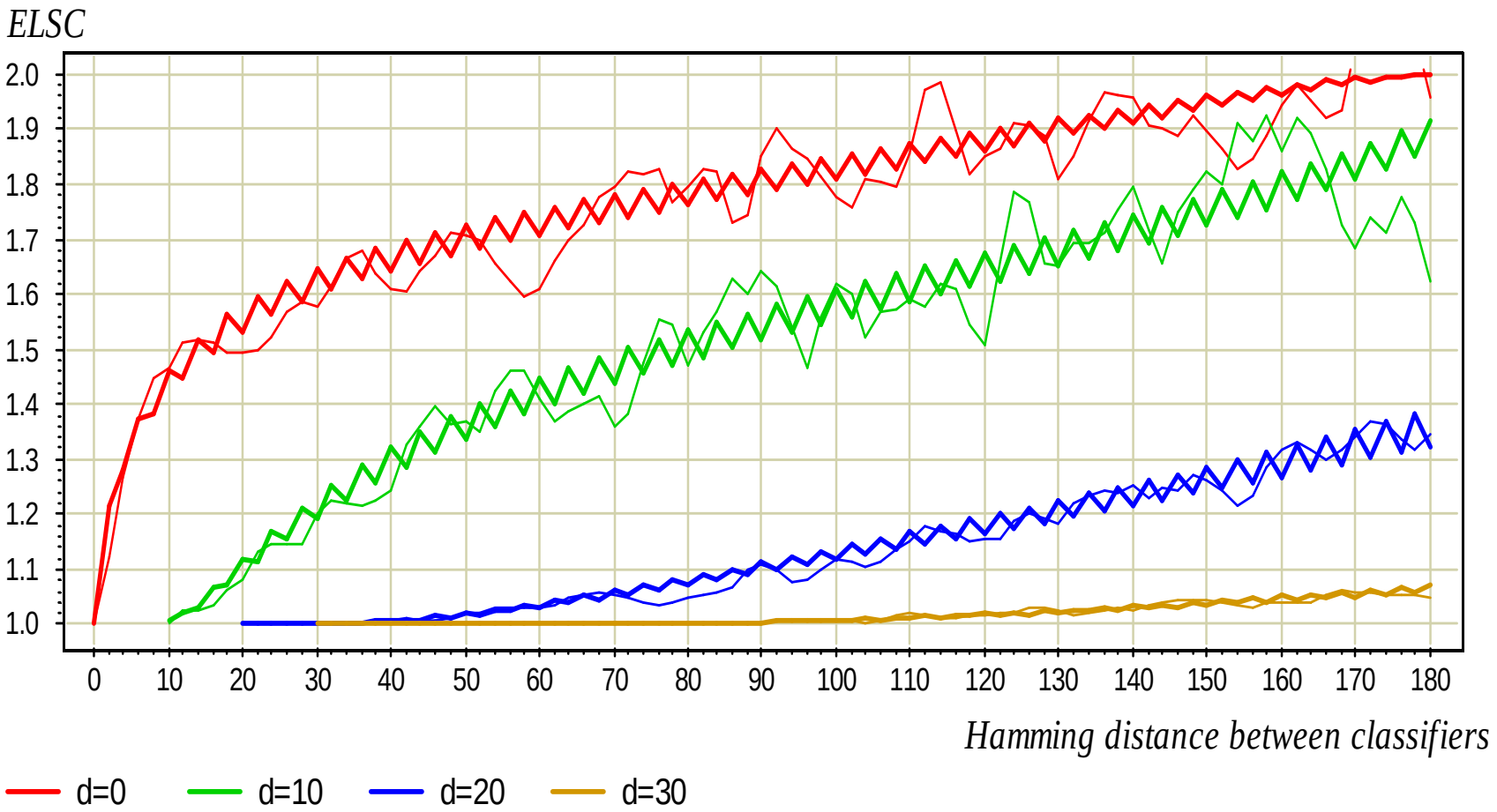
Эксперимент №1. Два алгоритма одинакового качества

$\ell = k = 100$; $\varepsilon = 0.05$; $m_1 = m_2$; $m = 20, 40, 100$



Эксперимент №2. Два алгоритма разного качества (расслоение)

$\ell = k = 100;\ \varepsilon = 0.05;\ \underline{m_0 = 20};\ \underline{d \equiv m_2 - m_1} = 0,\ 10,\ 20,\ 30$



Эксперимент №3. Цепочка 1000 алгоритмов

$D = 1000$ алгоритмов задаются своими векторами ошибок;
$\ell = k = 100$ — длина обучающей и тестовой выборок ($L = 200$);
$m/L = 0.05, 0.25$ — качество лучшего алгоритма в цепочке;
$\varepsilon = 0.05$ — порог переобученности;
$N' = 1000$ случайных разбиений по методу Монте-Карло.

Бинарная $L \times D$-матрица векторов ошибок:

Example:

1	1→0	0	0→1	1	1	1	1	1	...
0	0	0	0→1	1	1	1	1	1→0	...
0	0	0	0	0	0	0	0	0	0	...
0	0	0→1	1	1	1	1→0	0	0	...
0	0	0	0	0	0	0→1	1	1	1	...
0→1	1	1	1	1→0	0	0→1	1	...

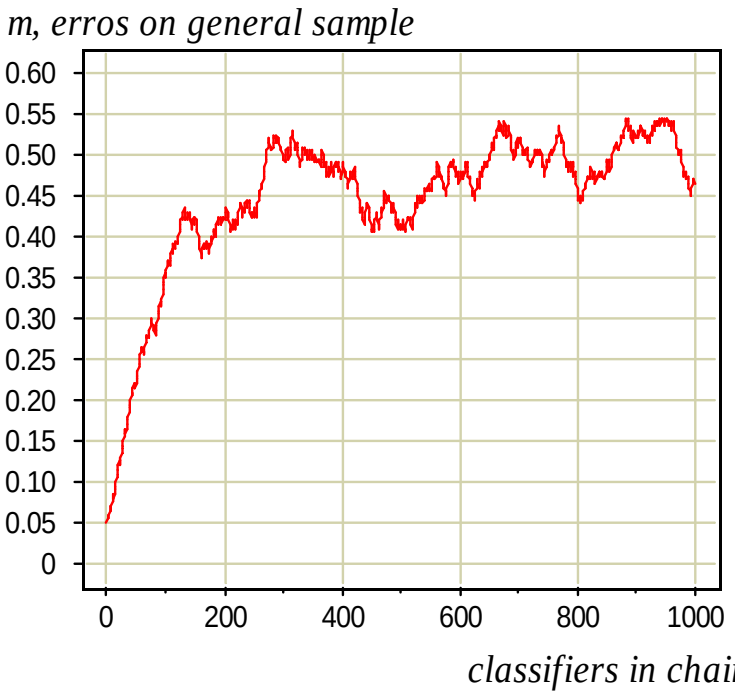
Цепочка алгоритмов — последовательность векторов ошибок, в которой каждые два соседних вектора отличаются только в одном разряде.

Цепочки с расслоением и без

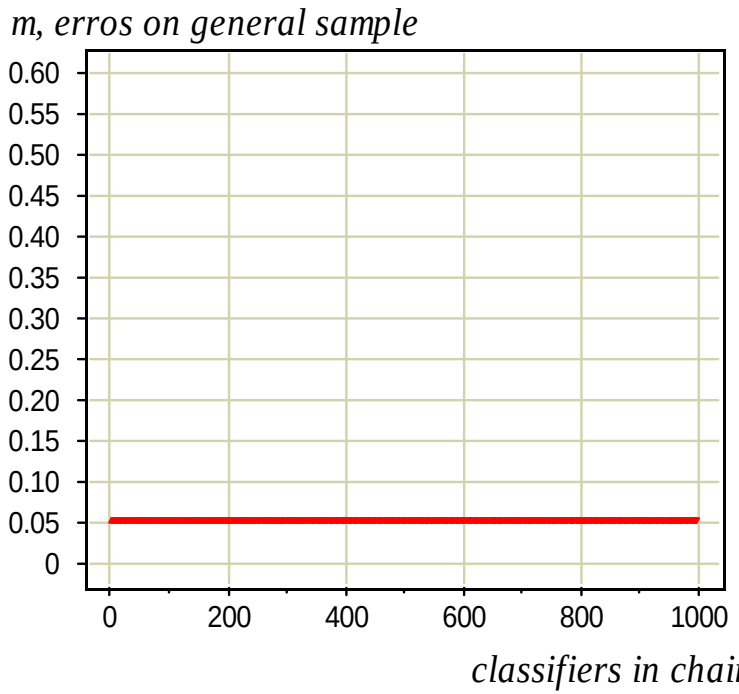
Цепочка алгоритмов — последовательность векторов ошибок, в которой каждые два соседних вектора отличаются только в одном разряде.

Два крайних случая цепочек:

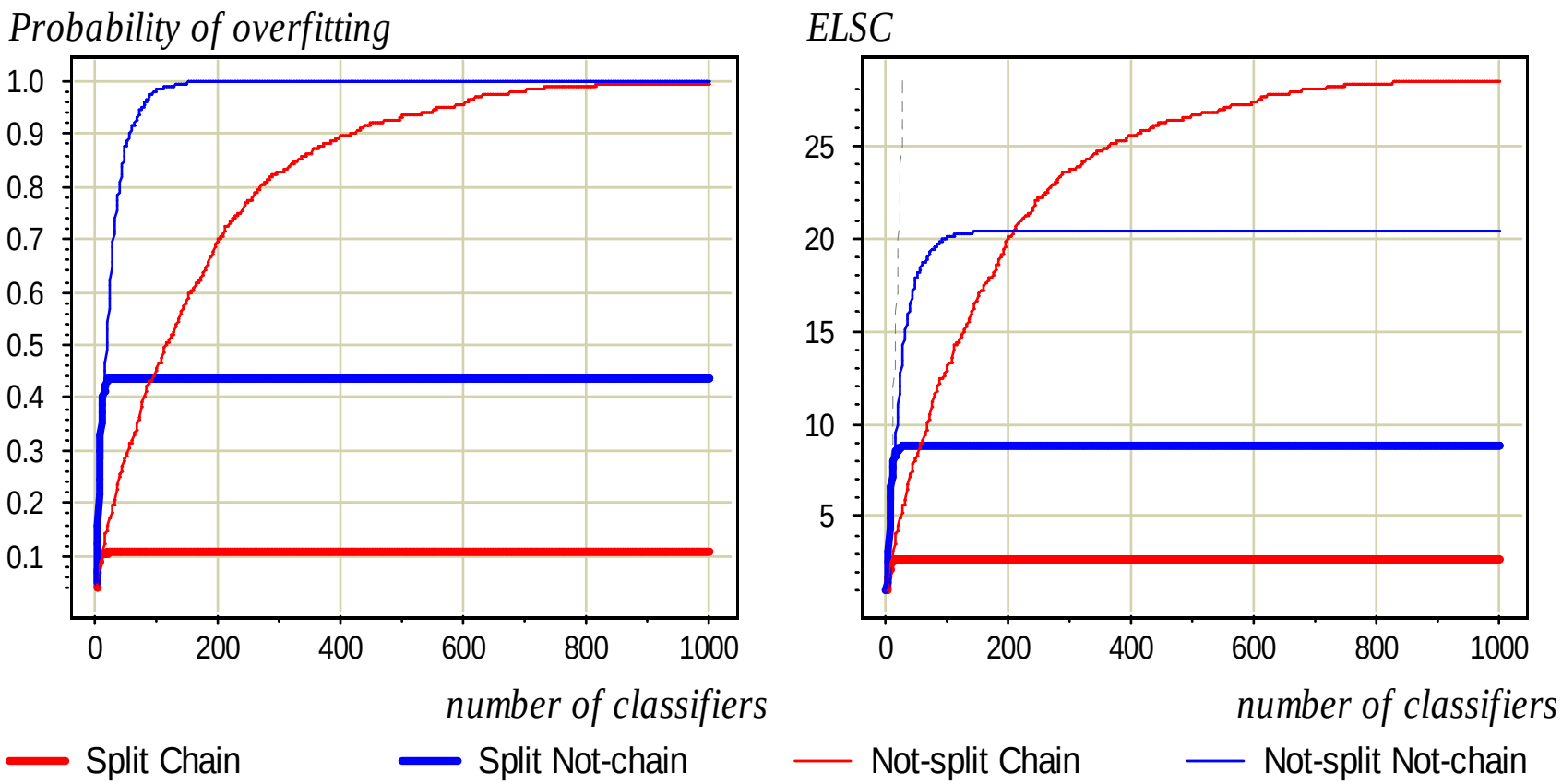
(1) split chain



(2) not-split chain

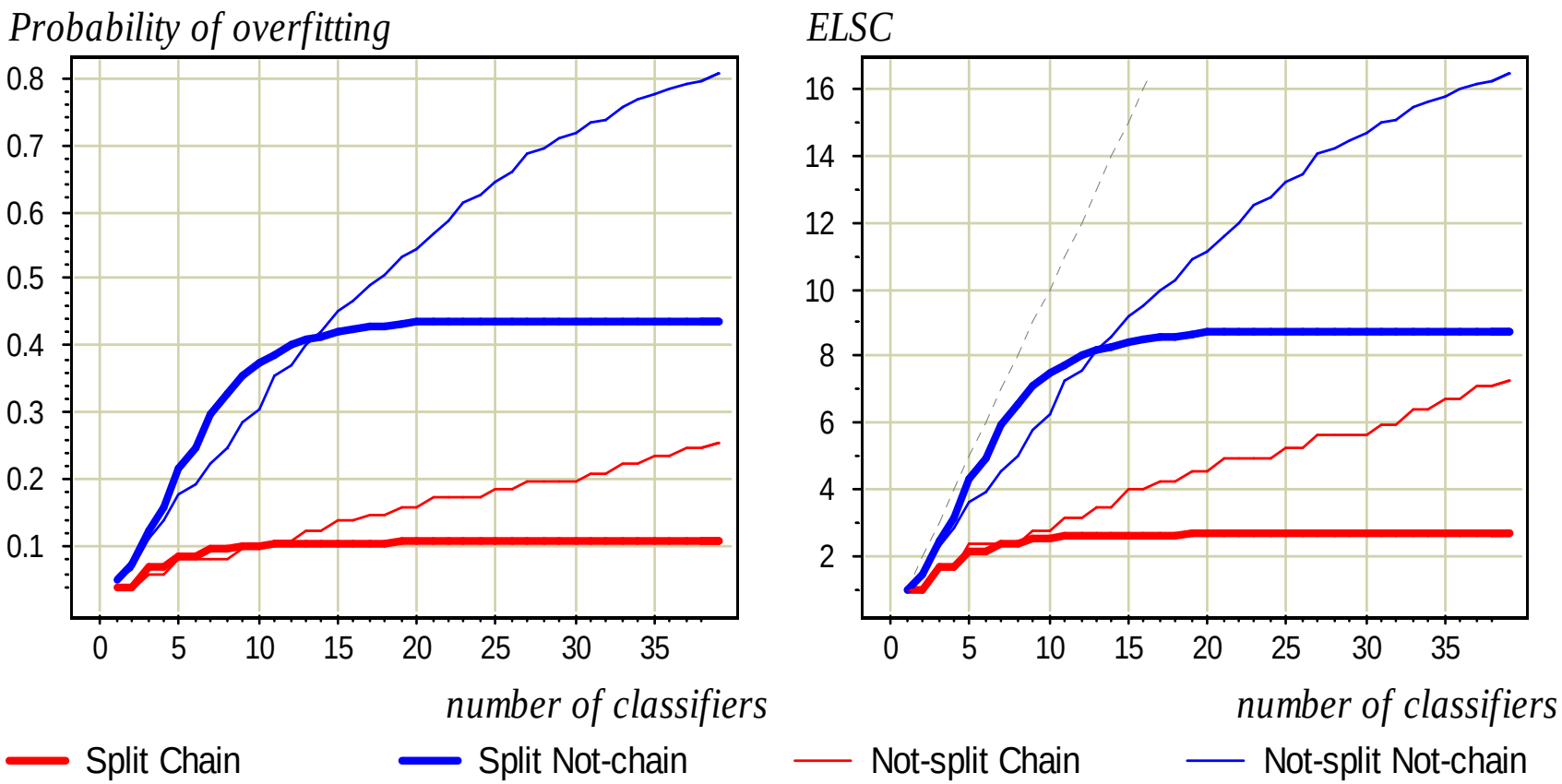


Цепочки и не-цепочки; с расслоением и без ($m/L = 0.05$)



В случае расслоения и низкого уровня ошибок ($m/L = 0.05$), вероятность переобучения не достигает 1 при $D \to \infty$.

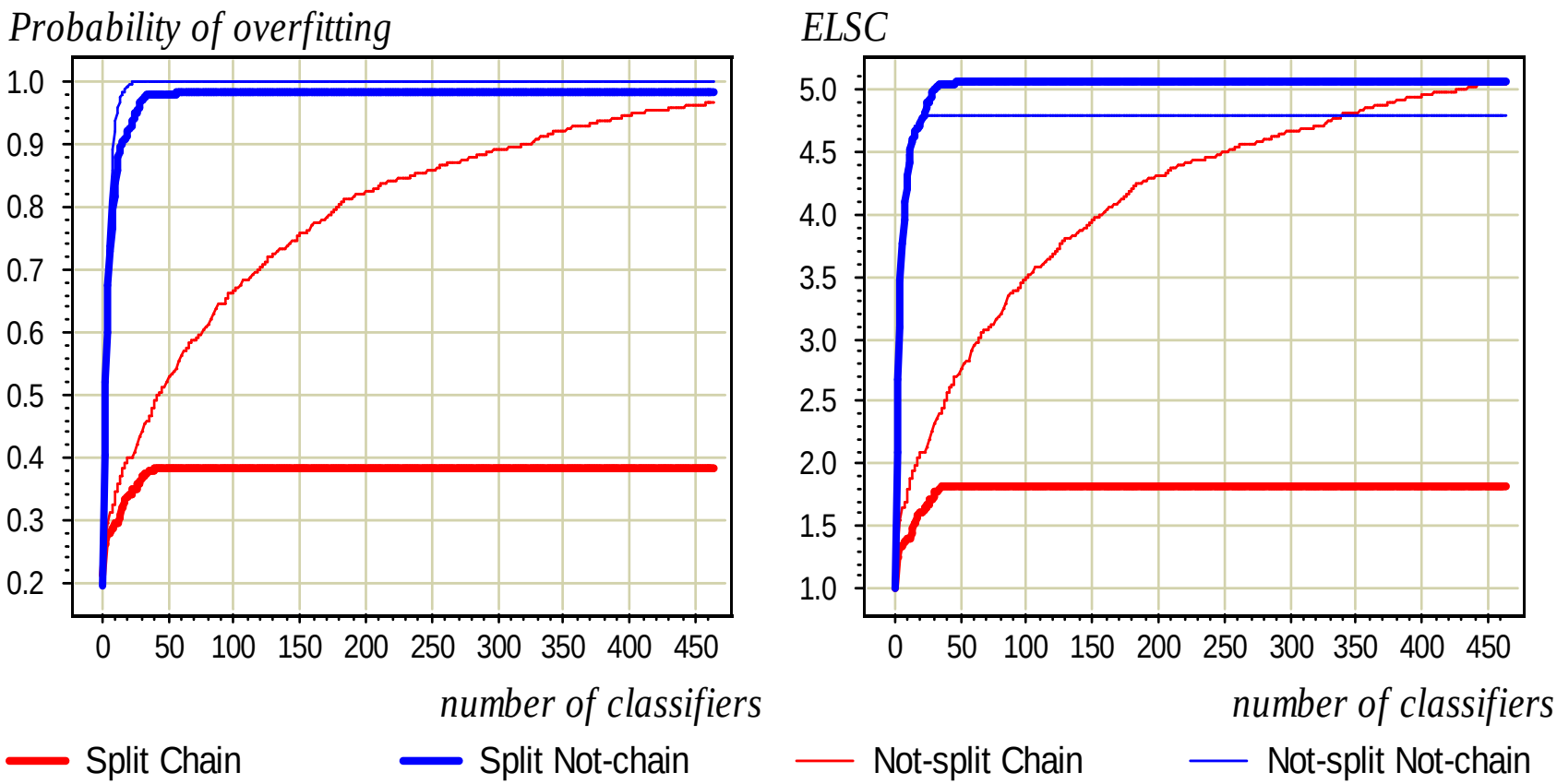
Цепочки и не-цепочки; с расслоением и без ($m/L = 0.05$, увеличено)



Согласно теории Вапника-Червоненкиса $\hat{\Delta}(D) = D$.

Это реализуется только для не-цепочек при малых D.

Цепочки и не-цепочки; с расслоением и без ($m/L = 0.25$)



В случае высокого уровня ошибок ($m/L = 0.25$) только *цепочки с расслоением* имеют низкую вероятность переобучения.

Выводы

- ЭЛКР $\hat{\Delta}(D)$ имеет горизонтальную асимптоту, тогда как согласно теории Вапника-Червоненкиса $\hat{\Delta} = D$.
- Наличие *цепочки* существенно замедляет рост $\Delta(D)$.
- Наличие *расслоения* опускает вниз горизонтальную асимптоту.
- **О природе переобучения:** переобучение возникает даже в случае двух алгоритмов.
- **Источник оптимизма:** *Цепочки с расслоением* — самый частый на практике случай; и именно в этом случае вероятность переобучения мала.
- **Мотивация для дальнейших исследований:** Пока ни одна теория не позволяет учесть расслоение и сходство алгоритмов одновременно.