Вероятность переобучения плотных и разреженных семейств алгоритмов

И. О. Толстихин
iliya.tolstikhin@gmail.com

ВЦ РАН

«Интеллектуализация обработки информации»
октябрь 2010

Содержание

Постановка задачи

Объекты $\mathbb{X} = \{x_1, \dots, x_L\}$; алгоритмы $A = \{a_1, \dots, a_D\}$;

$I(a, x) = \bigl[$алгоритм a ошибается на объекте $x\bigr]$;

$n(a, X)$ — *число ошибок* a на выборке X;

$\nu(a, X) = n(a, X)/|X|$ — *частота ошибок* a на выборке X.

Статистическое обучение: $a(X) = \arg\min\limits_{a \in A} n(a, X)$.

Переобучение: $\nu(a, X') - \nu(a, X) \geqslant \varepsilon$.

Пример. Бинарная $L \times D$-матрица ошибок, $L = \ell + k$:

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_D	
x_1	1	1	0	0	0	1	1	1	1	
x_2	0	0	0	0	1	1	1	1	1	X, обучение
x_ℓ	0	0	1	0	0	0	0	0	0	
x'_1	0	0	0	1	1	1	1	1	0	
x'_2	0	0	0	1	0	0	0	1	1	X', контроль
x'_k	0	1	1	1	1	1	0	0	0	

Задача: оценить вероятность переобучения.

Основная задача — оценить вероятность переобучения

$$Q_\varepsilon = \mathsf{P}\Big[\nu\big(a(X), X'\big) - \nu\big(a(X), X\big) \geqslant \varepsilon\Big].$$

Теорема (Вапник и Червоненкис, 1971)

Для любой выборки, метода обучения и числа $\varepsilon \in (0,1)$

$$Q_\varepsilon \leqslant \mathsf{P}\Big[\sup_{a\in A}\big(\nu(a,X') - \nu(a,X)\big) \geqslant \varepsilon\Big] \leqslant |A| \max_m H_L^{\ell,m}\left(\tfrac{\ell}{L}(m-\varepsilon k)\right),$$

где $H_L^{\ell,m}(z) = \sum\limits_{s=0}^{\lfloor z\rfloor} \dfrac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ *— гипергеометрическая функция распределения.*

Основная проблема — завышенность оценки:

в реальных задачах $|A| \sim 10^6 - 10^{12}$

Выявление причин завышенности

Основные причины завышенности:

- не учитывается *расслоение* семейства алгоритмов: чем выше уровень ошибок m, тем меньше вероятность получить алгоритм в результате обучения (завышенность в 10^2–10^5 раз);
- не учитывается *связность* семейства алгоритмов: чем больше схожих алгоритмов, тем сильнее завышенность (завышенность в 10^3–10^4 раз).

Реальные семейства, как правило, расслоены и связны.

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

Постановка задачи

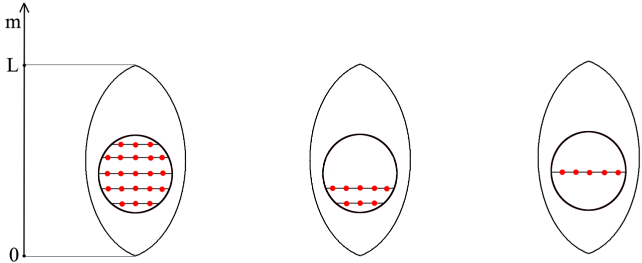
Гипотеза

Вероятность переобучения расслоенного и связного множества алгоритмов может быть аппроксимирована вероятностью переобучения его подмножества, состоящего из существенно различных алгоритмов нижних слоев.

Задача: на примере модельных семейств показать,
что возможна замена исходного семейства его подмножеством малой мощности.

Модельные семейства

A_m — m-й слой булева куба $\{0,1\}^L$ — множество алгоритмов, допускающих на полной выборке m ошибок.



Шар алгоритмов   Нижние слои шара   Слой шара

Точная оценка для шара алгоритмов

A — хэммингов шар радиуса r_0 с центром в некотором элементе m-го слоя.

Теорема (Оценка для шара)

Точная оценка вероятности переобучения этого семейства есть

$$Q_\varepsilon = \sum_{i=0}^{r_0} h_L^{\ell,m}(i) \frac{\sum_{r_1=0}^{r_0} \sum_{n_1=0}^{r_1} S\left(n_1, r_1, i\right) \left[m + r_1 - 2n_1 \geqslant \varepsilon k\right]}{\sum_{r_2=0}^{r_0} \sum_{n_2=0}^{r_2} S\left(n_2, r_2, i\right)} + \\ + \sum_{i=r_0+1}^{\lfloor s_d + \frac{rk}{L} \rfloor} h_L^{\ell,m}(i).$$

где $h_L^{\ell,m}(i) = \frac{C_m^i C_{L-m}^{\ell-i}}{C_L^\ell}$, $S\left(n, r, i\right) = C_{m-i}^{n-i} C_{k-m+i}^{r-n}$, $s_d = \frac{\ell}{L}(m - \varepsilon k)$.

Вклады слоёв шара

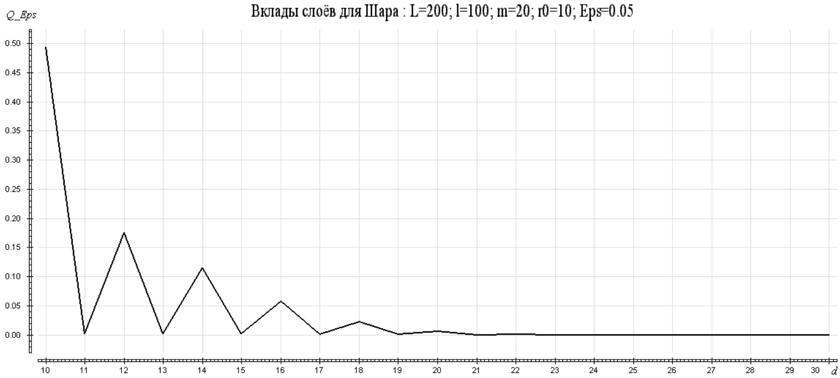


Рис.: Вклады слоёв шара в вероятность переобучения

Точная оценка для нижних слоев шара

A — хэммингов шар радиуса r_0 с центром в некотором элементе m-го слоя в пересечении с d его нижними слоями.

Теорема (Оценка для нижних слоев шара)

Точная оценка вероятности переобучения этого семейства есть

$$Q_\varepsilon = \sum_{i=0}^{r_0} h_L^{\ell,m}(i) \frac{\sum_{r=0}^{r_0} \sum_{n=0}^{r} S'(n,r,i)\big[m + r - 2n \geqslant \varepsilon k\big]}{\sum_{r=0}^{r_0} \sum_{n=0}^{r} S'(n,r,i)} + \\ + \sum_{i=r_0+1}^{\lfloor s_d + \frac{rk}{L} \rfloor} h_L^{\ell,m}(i),$$

где

$h_L^{\ell,m}(i) = \frac{C_m^i C_{L-m}^{\ell-i}}{C_L^\ell}$, $S'(n,r,i) = C_{m-i}^{n-i} C_{k-m+i}^{r-n} \big[r + r_0 + 1 \leqslant 2n + d\big]$, $s_d = \frac{\ell}{L}(m - \varepsilon k)$.

Оценки для нижних слоёв шара

Во всем шаре 66 018 452 алгоритмов. В трех нижних слоях шара 23 176 алгоритмов.

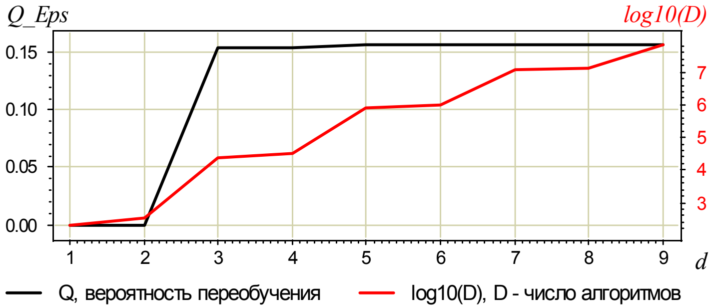


Рис.: Зависимость Q_ε и $\log_{10}|A|$ от числа d нижних слоев шара, при $\ell = k = 100,\ \ m = 10,\ \ r_0 = 4,\ \ \varepsilon = 0.07.$

Зависимость Q_ε от числа алгоритмов в семейства

Во всем шаре 66 018 452 алгоритмов.

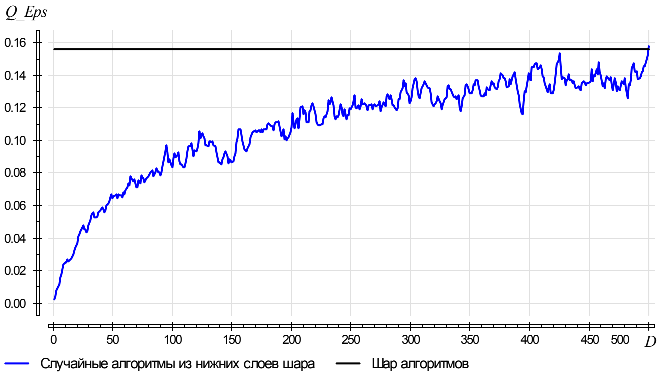
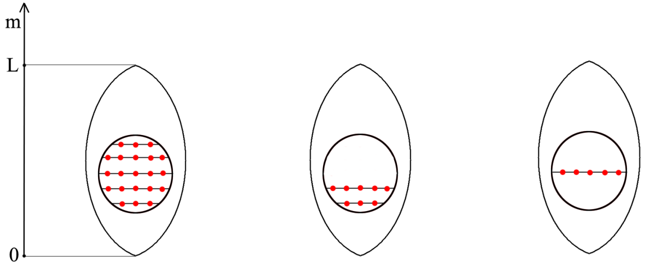


Рис.: Оценки Q_ε для шара алгоритмов и D случайных алгоритмов из трех его нижних слоев, при $\ell = k = 100$, $m = 10$, $r_0 = 4$, $\varepsilon = 0.07$.

Модельные семейства

A_m — m-й слой булева куба $\{0,1\}^L$ — множество алгоритмов, допускающих на полной выборке m ошибок.



Шар алгоритмов   Нижние слои шара   Слой шара

Точная оценка для слоя шара

$B(m, r_0)$ — пересечение m-го слоя булева куба $\{0,1\}^L$ с хэмминговым шаром радиуса r_0 с центром в некотором элементе m-го слоя.

Теорема (Оценка для слоя шара)

Если μ минимизирует частоту ошибок на обучающей выборке, то достижимая верхняя оценка вероятности переобучения этого семейства есть

$$Q_\varepsilon = H_L^{\ell,m}\left(s_d + \lfloor r_0/2 \rfloor\right),$$

где $s_d = \frac{\ell}{L}(m - \varepsilon k)$.

Зависимость Q_ε от числа алгоритмов в семейства

$$|B(m, r_0)| = 809\,876.$$

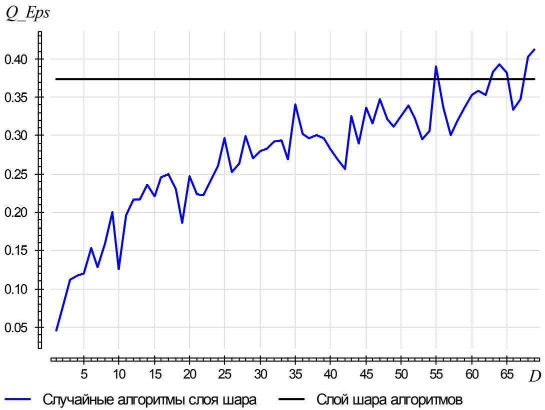


Рис.: Оценки Q_ε для слоя шара и его случайного подмножества из D алгоритмов, при $l = k = 100$, $m = 10$, $r_0 = 4$, $\varepsilon = 0{,}05$.

Вероятность переобучения слоя шара сосредоточена на его «внешней окружности»

Лемма

Пусть $A \subset A_m$ и $a \in A_m \setminus A$.
Тогда $Q_\varepsilon(A) \leqslant Q_\varepsilon(A \cup \{a\})$.

$S(m, r_0)$ — пересечение m-го слоя булева куба $\{0,1\}^L$ с хэмминговой сферой радиуса r_0 с центром в некотором элементе m-го слоя.

Теорема

Пусть $m = n(a, \mathbb{X})$ и $k \geqslant m + \lfloor r_0/2 \rfloor$.
Тогда вероятности переобучения множеств $B(m, r_0)$ и $S(m, 2\lfloor r_0/2 \rfloor)$ совпадают.

Подмножества B' и B''

Пусть a_0 — центр сферы $S(m, 2\lfloor r_0/2\rfloor)$;
$n(a_0, X^m) = m$, $\bar{X}^m = \mathbb{X} \setminus X^m$, $\delta = \lfloor r_0/2\rfloor$.
Подмножество $B'(m, r_0) \subset S(m, 2\delta)$ *образовано всеми различными алгоритмами, которые допускают $m - \delta$ ошибок на X^m и δ ошибок на фиксированных объектах подвыборки $\bar{X}^m$.*
$|B'(m, r_0)| = C_m^\delta$;

Пусть $L - m$ кратно δ.
Подмножество $B''(m, r_0) \subset S(m, 2\delta)$ *образовано всеми различными алгоритмами, допускающими $m - \delta$ ошибок на X^m и δ ошибок на $\bar{X}^m$, так что для любого $x \in \bar{X}^m$ и любого подмножества $X' \subset X^m$: $|X'| = m - \delta$ существует единственный $a \in B''$: $I(a, x) = 1$, $n(a, X') = m - \delta$.*
$|B''(m, r_0)| = C_m^\delta \frac{L-m}{\lfloor r_0/2\rfloor}$.

Оценки вероятности переобучения для подмножеств B' и B''

Теорема (Подмножество B')

Точная оценка вероятности переобучения семейства $B'(m, r_0)$

$$Q_\varepsilon = \sum_{\substack{i=0 \\ i+j+p=\ell}}^{m} \sum_{j=0}^{h} \sum_{p=0}^{\delta} \frac{C_m^i C_h^j C_\delta^p}{C_L^\ell} \times$$

$$\times \Big([i < \delta]\,[p \leqslant s_d(\varepsilon)] + [i \geqslant \delta]\,[i + p \leqslant \delta + s_d(\varepsilon)] \Big),$$

где $h = L - m - \delta$, $s_d(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$, $\delta = \lfloor r_0/2 \rfloor$.

Теорема (Подмножество B'')

Пусть $\ell < \frac{L-m}{\lfloor r_0/2 \rfloor}$. Тогда вероятности переобучения множеств $B''(m, r_0)$ и $B(m, r_0)$ совпадают.

Приближение оценки слоя шара подмножествами B' и B''

$$|B(m, r_0)| = 809\,876, \ \ |B'| = 45, \ \ |B''| = 4275.$$

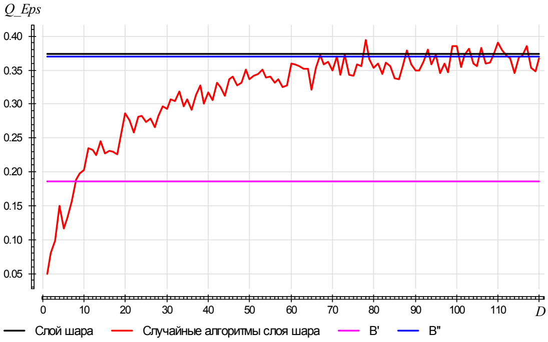


Рис.: Оценки Q_ε для слоя шара, множеств $B'(m, r_0)$, $B''(m, r_0)$ и случайного подмножества D алгоритмов из слоя шара, при $l = k = 100$, $m = 10$, $r_0 = 4$, $\varepsilon = 0{,}05$.

Результаты и открытые вопросы

Полученные результаты.

- Показана возможность аппроксимации вероятности переобучения расслоенных и связных семейств с помощью их разреженных подмножеств.

Открытые вопросы.

- Как строить разреженные подсемейства в практических задачах?
- Как минимизировать мощность этих разреженных подсемейств?