УДК 519.6

ОЦЕНИВАНИЕ СМЕЩЕНИЯ ЭМПИРИЧЕСКОГО РИСКА ДЛЯ ЛИНЕЙНЫХ КЛАССИФИКАТОРОВ

В.М. Неделько

Институт математики СО РАН,
Новосибирск, пр-т Коптюга, 4, Россия, 630090
e-mail: *nedelko@math.nsc.ru*

Abstract

An empirical risk bias in classification problem is researched. Statistical modeling performed shows that the risk bias dependence on decision class capacity appears to be the same both for multinomial (discrete) case and for linear classifier. This result ensures that universal scaling of Vapnik-Chervonenkis bias estimations may be available since such scaling was obtained for a discrete case. To prove using an empirical risk estimator a comparison of it's volatility versus volatility of leave-one-out estimator is also performed.

Введение

Рассматривается задача оценивания качества решающей функции распознавания по обучающей выборке. Используемые для ее решения подходы обладают различными недостатками, в частности, для емкостного подхода [1] на данный момент в общем случае известны только приближенные оценки риска, байесовский — требует предположений о классе распределений [2]. Для оценок скользящего экзамена, как правило, не известна дисперсия, более того девиация этих оценок часто уменьшается медленнее, чем $N^{-1/2}$. Кроме того их вычисление обычно требует существенных вычислительных затрат, хотя в некоторых случаях такую оценку можно получить комбинаторно [4].

Для задачи распознавания образов в дискретном пространстве получены [3] точные оценки вероятности ошибочной классификации, что позволило определить погрешность оценок Вапника-Червоненкиса. В настоящей работе исследовалась возможность переноса полученного результата на непрерывный случай, для чего проведено статистическое моделирование построения оптимальной выборочной линейной разделяющей функции при наихудшем распределении в пространстве переменных. Полученные в результате моделирования значения смещения эмпирического риска оказываются практически совпадающими с решающих функций. Это говорит о том, что степень завышенности оценок [1] может быть одинаковой для различных задач.

1. Постановка задачи

Пусть X — пространство значений переменных, используемых для прогноза, а Y — пространство значений прогнозируемых переменных, и пусть C — множество всех (с некоторыми формальными ограничениями) вероятностных мер на $D = X \times Y$.

Тогда элементом $c \in C$ будет $P_c[D]$. Здесь и далее квадратные скобки используются для указания множества, на σ-алгебре подмножеств которого задана мера.

Решающей функцией назовем $f : X \to Y$ соответствие и введем для нее функцию потерь: $L : Y^2 \to [0, \infty)$.

Под риском будем понимать средние потери:

$$R(c, f) = \int L(y, f(x))dP_c[D] = \int R_X(c, f)dP_c[X],$$

где $R_x(c, f) = \int L(y, f(x))dP_c[Y/x]$.

Пусть $v_c = \{(x^i, y^i) \in D \mid i = \overline{1, N}\}$ – случайная независимая выборка из распределения $P_c[D]$. Эмпирический риск определим как средние потери на выборке:

$$\widetilde{R}(v, f) = \frac{1}{N} \sum_{i=1}^{N} L(y^i, f(x^i)).$$

Все практически используемые алгоритмы построения решающих функций так или иначе минимизируют эмпирический риск, поэтому последний оказывается смещенной оценкой риска. *Целью работы* является оценка данного смещения.

Обозначим: $F(c, Q) = ER(c, f_{Q,v}), \widetilde{F}(c, Q) = E\widetilde{R}(c, f_{Q,v})$, где $Q : \{v\} \to \{f\}$ — алгоритм построения решающих функций, а $f_{Q,v}$ — функция, построенная по выборке v алгоритмом Q. Математическое ожидание берется по всем выборкам объема N.

Введем функцию максимального смещения:

$$S_Q(\widetilde{F}_0) = \hat{F}_Q(\widetilde{F}_0) - \widetilde{F}_0, \tag{1}$$

где $\hat{F}_Q(\widetilde{F}_0) = \sup\limits_{c:\widetilde{F}(c,Q)=\widetilde{F}_0} F(c, Q)$.

2. Дискретный случай

Пусть X дискретно, т. е. $X = \{1, \ldots, n\}$, и пусть решающая функция минимизирует эмпирический риск независимо в каждой точке:

$f^*_v(x) = \arg\min\limits_{y \in Y} \widetilde{R}(v_x, y)$, где $v_x = \{(x^i, y^i) \in v | x^i = x\}$.

Для данной дискретной постановки ранее (см. напр. [3]) был разработан метод нахождения точной зависимости $S_Q(\widetilde{F}_0)$ при заданных n и N. Метод применим для любого типа пространства Y, но наиболее иллюстративные результаты получаются для задачи классификации при двух образах, поскольку в этом случае результат легко сравним с оценками Вапника-Червоненкиса.

Рассмотрим задачу классификации двух образов с функцией потерь:

$$L(y, y') = \begin{cases} 0, y = y' \\ 1, y \neq y' \end{cases}$$

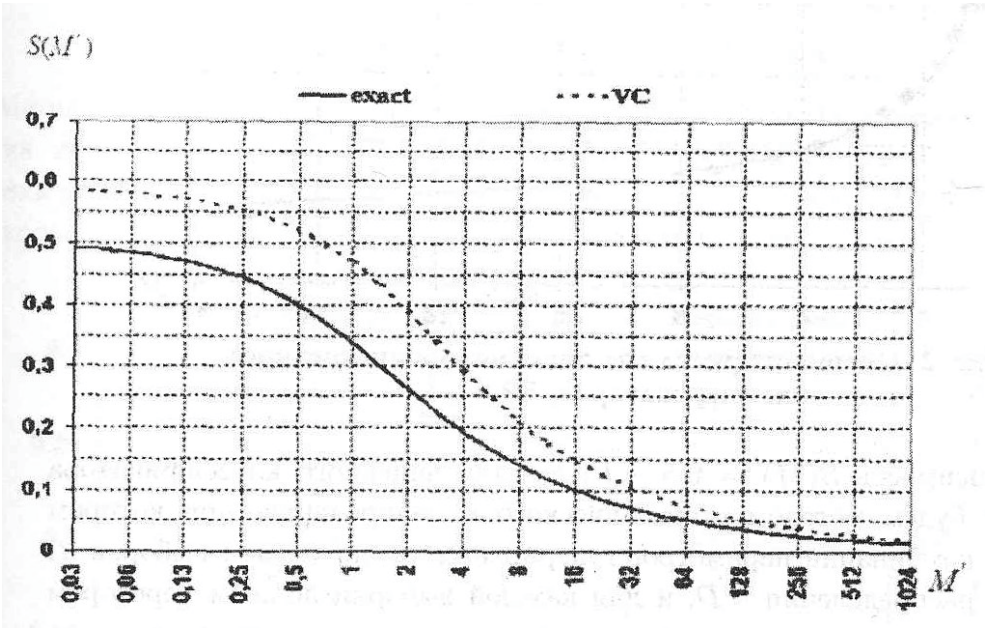

Рис. 1. Погрешность оценок В-Ч, $ER = 0.5$.

На рис. 1 приведена зависимость $S(M) = 0.5 - \widetilde{F}(C_U)$ и ее $S_V(M) = \sqrt{\frac{\ln 2}{2M'}}$, предложенная Вапником и Червоненкисом. Здесь $\widetilde{F}(C_U)$ – ожидаемый эмпирический риск при равномерном распределении; $\frac{N}{n} = M = const, N \to \infty, n \to \infty; M' = M/(1-e^{-M})$ – объем выборки, отнесенный к емкости класса решающих функций (двойный логарифм числа способов разбиения выборки).

Заметим, что равномерное распределение в D дает максимальное смещение эмпирического риска, если нет ограничений на $\widetilde{F}_0$. Если же зафиксировать значение $\widetilde{F}_0$, то «наихудшее» распределение (при котором риск максимален) не будет равномерным при $M > 1$.

3. Линейный классификатор

Сравним теперь значения смещения эмпирического риска, полученные для дискретного случая, со смещениям для линейных разделяющих функций. Для простоты предположим, что распределение в D равномерно. При такой c риск (вероятность ошибки) для любой решающей функции равна 0.5, однако эмпирический риск будет в среднем значительно меньше.

Искать зависимость $S(M) = 0.5 - \widetilde{F}(C_U)$ для линейного классификатора в $X = [0,1]^d$ будем методом статического моделирования, при котором для заданных комбинаций параметров генерируется около тысячи выборок из равномерного распределения в D, и для каждой выборки полным перебором находится гиперплоскость, наилучшим образом разделяющая классы на выборке.

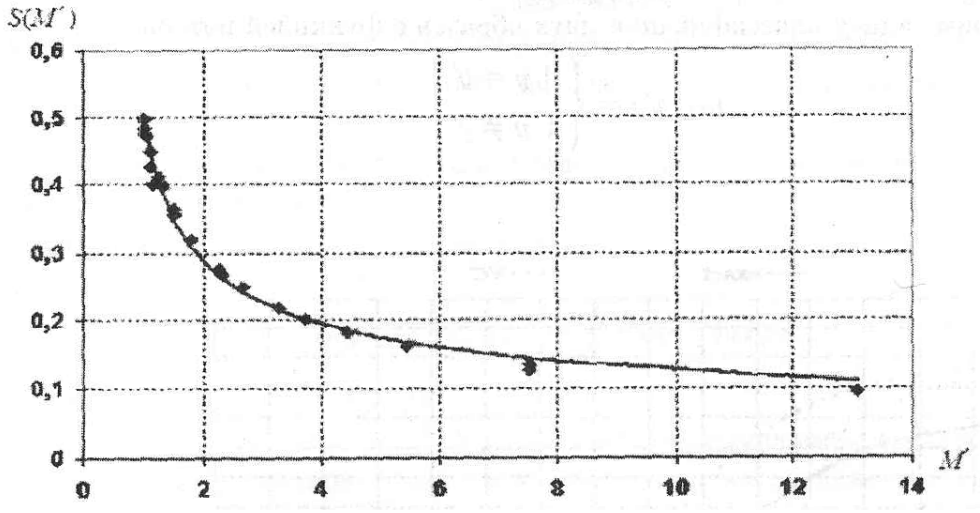

Рис. 2. Смещения риска для линейного и дискретного классификаторов, $ER = 0.5$.

Результаты приведены в таблице.

d	N	M'	S	d	N	M'	S
1	3	1.16	0.4	1	5	1.5	0.36
1	10	2.31	0.27	1	20	3.75	0.20
1	50	7.53	0.13	1	100	13.08	0.095
2	4	1.05	0.47	2	10	1.53	0.357
2	20	2.33	0.27	2	50	4.44	0.18
2	100	7.53	0.13	3	5	1.019	0.48
3	10	1.247	0.41	3	20	1.79	0.32
3	50	3.28	0.22	3	100	5.46	0.162
4	6	1.008	0.498	4	10	1.111	0.45
4	20	1.5	0.355	4	50	2.66	0.249
5	7	1.003	0.499	5	10	1.044	0.476
5	20	1.33	0.398	5	50	2.27	0.275

Здесь d размерность пространства X, N – объем выборки, $M' = \frac{N}{\log_2 C}$ – объем выборки, отнесенный к емкости класса линейных классификаторов (C – максимальное число способов разделить гиперплоскостью выборку на два подмножества), S – смещение риска.

Этот же результат (точки отмечены маркерами) показан на рис. 2 в сравнении с зависимостью $S(M')$ для дискретного случая (сплошная линия). На рисунке видно, что зависимость смещения эмпирического риска от M' для линейного классификатора близка в зависимости для дискретного случая.

4. Оценки скользящего экзамена

Метод скользящего экзамена привлекателен тем, что получаемая оценка риска является несмещенной. Однако также известно, что дисперсия этой оценки существенно выше, чем дисперсия эмпирического риска. Чтобы понять, насколько этот факт существенен, рассмотрим два примера.

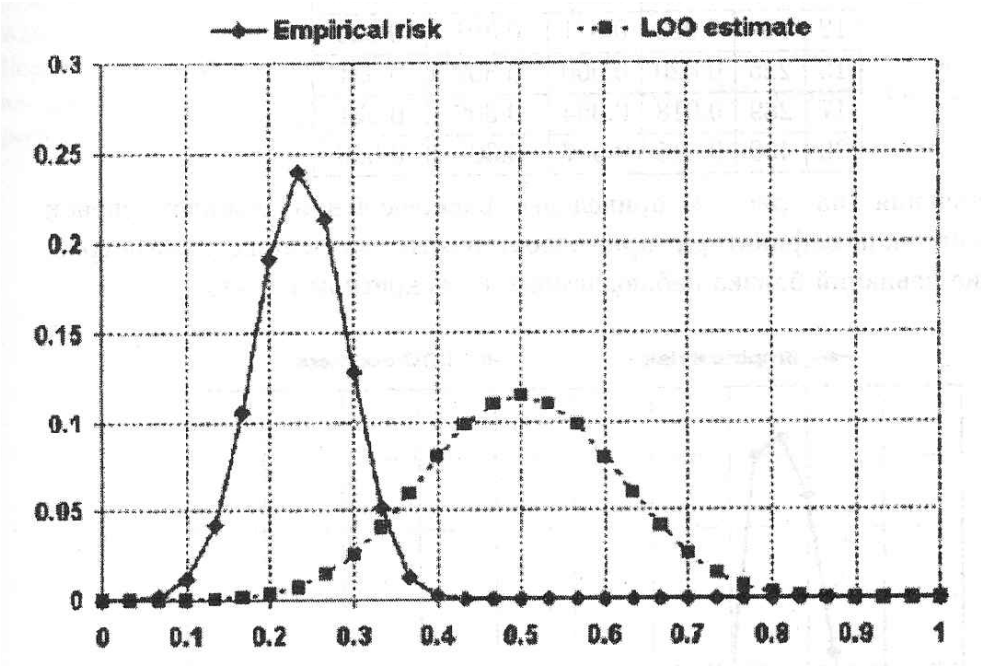

Рис. 3. Распределения для эмпирического риска и оценки скользящего экзамена при $N = 30, n = 15, ER = 0.5$.

На рис. 3 приведены распределения вероятностей на значениях эмпирического риска и на значениях оценки скользящего экзамена для дискретного случая двух классов при $N = 30, n = 15$ и равномерном распределении в D ($ER = 0.5$).

В таблице приведены значения девиаций оценок для типичного случая, когда параметр n, выступающий в роли сложности, растет как $N^{-1/2}$. Видно, что девиация эмпирического риска $\sigma_{\widetilde{R}}$ уменьшается пропорционально $N^{-1/2}$, а девиация σ_L оценки скользящего экзамена для приведенных значений параметров аппроксимируется только как $N^{-0.4}$.

n	N	$\sigma_{\widetilde{R}}$	σ_L	$\sigma_{\widetilde{R}} N^{1/2}$	$\sigma_L N^{0.4}$
2	4	0.151	0.300	0.303	0.522
3	9	0.100	0.220	0.300	0.529
4	16	0.075	0.174	0.301	0.529
5	25	0.060	0.146	0.301	0.528
6	36	0.050	0.126	0.301	0.527
7	49	0.043	0.111	0.301	0.525
8	64	0.038	0.099	0.301	0.525
9	81	0.033	0.090	0.301	0.524
10	100	0.030	0.083	0.301	0.523
12	144	0.025	0.071	0.301	0.522
15	225	0.020	0.060	0.302	0.521
17	289	0.018	0.054	0.302	0.521
20	400	0.015	0.047	0.301	0.520

Для сравнения на рис. 4 приведены распределения данных оценок для линейного классификатора при сопоставимых значениях параметров. Соотношение девиаций близко наблюдаемому в дискретном случае.

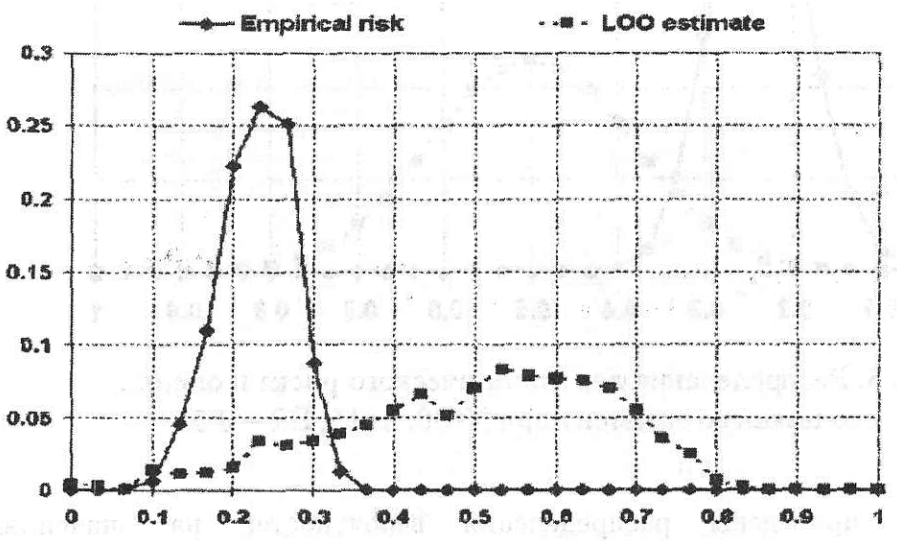

Рис. 4. Распределения оценок для случая линейного классификатора при $N = 30, d = 3, ER = 0.5$.

Заключение

Проведенное исследование позволяет выдвинуть гипотезу о том, что смещение эмпирического риска зависит от емкости класса решающих функций, но не от самого

класса, и не зависит от типа переменных. Поэтому точные оценки смещения, полученные для дискретного случая, могут быть перенесены и на реальные задачи.

После того, как получена точная оценка смещения эмпирического риска, использование последнего становится зачастую более оправданным, чем оценки скользящего экзамена, ввиду меньшей дисперсии.

Список литературы

1. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М., 1974. 415 с.
2. Лбов Г. С., Старцева Н. Г. Сложность распределений в задачах классификации // Доклады РАН, 1994. Том 338 № 5.
3. Nedel'ko V. M. Estimating a Quality of Decision Function by Empirical Risk // LNAI 2734. Macine Learning and Data Mining, MLDM 2003, Leipzig. Proceedings. Springer-Verlag. 2003. pp. 182-187.
4. Воронцов К. В. О комбинаторном подходе к оценке качества обучения алгоритмов // Доклады 11-й Всероссийской конф. «Математические методы распознавания образов». М.: ВЦ РАН, 2003.