КРИТЕРИЙ ОЦЕНКИ КАЧЕСТВА РЕШАЮЩЕЙ ФУНКЦИИ ПО ЭМПИРИЧЕСКОМУ РИСКУ В ЗАДАЧЕ КЛАССИФИКАЦИИ.

Неделько В. М.

Институт математики СО РАН, Россия
630090, Новосибирск, пр. Коптюга, 4,
nedelko@math.nsc.ru

ABSTRACT

The problem of statistical robustness of decision rules in the task of classification is considered. We will investigate whether Vapnik–Chervonenkis estimations of error probability may be improved. A new criterion estimating error probability by the empirical risk is proposed.

ВВЕДЕНИЕ

Широко распространенным подходом к оцениванию качества решающей функции классификации является подход Вапника–Червоненкиса [1], основанный на использовании емкостной характеристики класса решающих правил. Этот подход очень привлекателен, поскольку не требует принятия каких-либо априорных гипотез. Вместе с тем, опыт решения прикладных задач дает основания полагать оценки, получаемые данным подходом, весьма завышенными [2].

Эта завышенность оценок может проистекать из двух причин: во-первых, критерий Вапника–Червоненкиса не использует как событие тот факт, что получено определенное значение эмпирического риска; во-вторых, по-видимому, существуют априорные гипотезы [2], принятие которых при решении прикладных задач оправдано.

В данной работе исследуется вопрос, насколько в принципе возможно уточнение оценок Вапника–Червоненкиса доверительного интервала для вероятности ошибки. Для этого задача классификации рассматривается в простом частном случае.

ПОСТАНОВКА ЗАДАЧИ

Пусть имеется единственная дискретная переменная X с множеством значений $D_X = \{\beta_1,...,\beta_n\}$ и $Y \in D_Y = \{0,1\}$ – прогнозируемая переменная. Для упрощения изложения первоначально будем рассматривать случай двух классов. Заметим, что предлагаемая методика оценки вероятности ошибки применима для произвольного числа классов — достаточно в соответствующих формулах заменить биномиальное распределение полиномиальным.

Обозначим $p_j = P(x = \beta_j)$ – вероятность принятия переменной X значения β_j, и пусть $\xi_j = P\left(y=1 \middle/ x=\beta_j\right)$. Класс всех распределений на $D = D_X \times D_Y$ обозначим C. Параметр $c \in C$ будет идентифицировать конкретное распределение, задаваемое совокупностью пар (p_j, ξ_j), $j = \overline{1,n}$. Распределение c будем также называть стратегией природы.

Минимизирующей эмпирический риск решающей функцией, построенной на основе выборки $v = \left(m_j^\omega \;\middle|\; \omega = \overline{0,1},\ j = \overline{1,n} \right)$, $\sum_{j=1}^{n} \left(m_j^0 + m_j^1 \right) = N$, где m_j^ω – количество выборочных точек с $x = \beta_j$, $y = \omega$, будет отображение $\alpha : D_X \to D_Y$, такое что $y_\alpha(x) = \arg\min_\omega m_j^\omega$.

Для указанной решающей функции эмпирическим риском будет $\widetilde{R}(v)=\frac{1}{N}\sum_{j=1}^{n}\min_{\omega} m_j^{\omega}$ – доля ошибок распознавания на обучающей выборке, а $R(c,v)=\sum_{j=1}^{n} p_j r_j$, где $r_j=\begin{cases}\xi_j, & y_\alpha(\beta_j)=0\\ \xi_j, & y_\alpha(\beta_j)=1\end{cases}$, есть вероятность ошибки

Пусть при решении прикладной задачи получено значение эмпирического риска $\widetilde{R}_0$. Для проверки гипотезы о том, что для построенной при этом решающей функции вероятность ошибки равна R_0, в [1] предлагается критерий $K_1(R_0)=\sup_{c\in C}\mathrm{P}\left(\left|\widetilde{R}(v)-R(c,v)\right|\ge\varepsilon\right)$, где $\varepsilon=R_0-\widetilde{R}_0$. Если $K_1(R_0)<\eta$ – заданного уровня значимости, то считаем, что для построенной решающей функции вероятность ошибки не равна R_0, а значит, меньше этой величины. При этом ε задает ширину доверительного интервала для вероятности ошибки.

В [3] с помощью моделирования при малых n было показано, что критерий K_1 может быть улучшен. В данной работе исследуется асимптотическое поведение критериев при увеличении объема выборки и мощности признакового пространства: $N\to\infty$, $n\to\infty$, $\frac{N}{n}=const=M$.

АСИМПТОТИЧЕСКОЕ ПОВЕДЕНИЕ КРИТЕРИЕВ

При $N\to\infty$, $\widetilde{R}(v)$ и $R(c,v)$ сходятся к своим математическим ожиданиям, тогда

$$\varepsilon=\sup_{c}\left(\mathrm{E}\,R(c,v)-\mathrm{E}\,\widetilde{R}(v)\right) \qquad (1)$$

Согласно [1] справедлива асимптотическая оценка $\varepsilon\le\varepsilon_1=\sqrt{2\ln 2/4M}\approx\frac{1{,}18}{\sqrt{4M}}$. При заданном $\widetilde{R}_0$ можно оценить $\varepsilon_1=R_0-\widetilde{R}_0$, найдя R_0 из уравнения $\left(R_0-\widetilde{R}_0\right)^2=\frac{2\ln 2}{M}R_0(1-R_0)$.

Чтобы учесть факт получения определенного значения эмпирического риска, в [3] предложен комбинированный критерий $K_2(R_0)=\sup_{c\in C}\min\left(\mathrm{P}\left(R(c,v)\ge R_0 \,/\, \widetilde{R}(v)\le\widetilde{R}_0\right),\mathrm{P}\left(\widetilde{R}(v)\le\widetilde{R}_0\right)\right)$, который при $N\to\infty$ эквивалентен $\varepsilon_2=\sup_{c\in\widetilde{C}}\mathrm{E}\,R(c,v)-\widetilde{R}_0$, где $\widetilde{C}=\left\{c\in C\,\middle|\,\mathrm{E}\,\widetilde{R}(v)\le\widetilde{R}_0\right\}$.

Обозначим m_j – число объектов (выборочных точек), со значением $X=\beta_j$. Асимптотическое распределение данной величины есть $\mathrm{P}_{M_j}(m_j)=e^{-M_j}\frac{{M_j}^{m_j}}{m_j!}$ – распределение Пуассона с параметром $M_j=p_jN$.

Зафиксируем m_j. Тогда распределение на a_j – количестве объектов с $X=\beta_j$, $Y=1$, есть $\mathrm{P}(a_j,m_j)=C_{m_j}^{a_j}\,\xi_j^{a_j}(1-\xi_j)^{m_j-a_j}$. Средний эмпирический риск будет $\widetilde{R}_{m_j}=\frac{1}{m_j}\sum_{i=0}^{m'_j}\tau(i,m_j)\,i\,\mathrm{P}(i,m_j)$, где $m'_j=\left[\frac{m_j}{2}\right]$, а $\tau(i,m_j)$ равно 1, если $i=m'_j$ и m_j – четное, иначе $\tau(i,m_j)$ равно 2. Аналогично, средняя вероятность ошибки есть $R_{m_j}=\frac{\xi_j}{2}\sum_{i=0}^{m'_j}\tau(i,m_j)\mathrm{P}(i,m_j)+\frac{1-\xi_j}{2}\sum_{i=m'_j}^{m_j}\tau(i,m_j)\mathrm{P}(i,m_j)$.

Проведя усреднение по m_j, обозначим $\widetilde{R}_j(\xi_j,M_j)=\frac{1}{M_j}\sum_{m_j=0}^{\infty}m_j\widetilde{R}_{m_j}\mathrm{P}_{M_j}(m_j)$ и $R_j(\xi_j,M_j)=\sum_{m_j=0}^{\infty}R_{m_j}\mathrm{P}_{M_j}(m_j)$. Теперь можем найти $\mathrm{E}\,\widetilde{R}(v)=\sum_{j=1}^{n}p_j\widetilde{R}_j$ и $\mathrm{E}\,R(c,v)=\sum_{j=1}^{n}p_jR_j$.

Задачу нахождения ε_2 в асимптотическом случае можно свести к вариационной задаче

$$\varepsilon_2 = \sup_{\xi(\cdot),\varphi(\cdot)} \int_0^1 F\big(\xi(\theta),\varphi(\theta)\big)\,d\theta - \widetilde{R}_0, \quad \text{при} \quad \int_0^1 \widetilde{F}\big(\xi(\theta),\varphi(\theta)\big)\,d\theta \le \widetilde{R}_0, \quad \xi(\theta)\in[0,1], \quad \varphi(\theta)\ge 0, \quad \int_0^1 \varphi(\theta)\,d\theta = 1.$$

(2)

Где $F\big(\xi(\theta),\varphi(\theta)\big) = \varphi(\theta)\,R_j\big(\xi(\theta), M\varphi(\theta)\big)$, а $\widetilde{F}\big(\xi(\theta),\varphi(\theta)\big) = \varphi(\theta)\,\widetilde{R}_j\big(\xi(\theta), M\varphi(\theta)\big)$.

Решением подобной задачи при достаточно больших M получена для ε из (1) оценка $\varepsilon_1' \approx \frac{0{,}78}{\sqrt{4M}} \approx 0{,}68\varepsilon_1$, где $0{,}78$ – приближенное значение отношения среднего модуля отклонения к стандартному отклонению для нормального распределения. Таким образом, найдена поправка к оценке ε_1 для рассматриваемого дискретного случая.

Исследуем теперь возможность улучшения критерия Вапника–Червоненкиса (с учетом найденной поправки).

Приближенное вычисление ε_2 по данному критерию в общем случае дало результаты, несущественно отличающиеся от оценок Вапника–Червоненкиса (с учетом поправки), однако, если принять гипотезу о равномерном априорном распределении в D_X, то ε_2 оказывается существенно меньше ε_1', например, при $M = 5$, $\widetilde{R}_0 = 0{,}1$, получаем $\varepsilon_1' = 0{,}15$, а $\varepsilon_2 = 0{,}05$.

При равномерном априорном распределении в D_X в (2) плотность $\varphi(\theta)\equiv 1$, и задачу можно свести к

$$\varepsilon_2 = \sup_{\theta(\widetilde{F})} \int_{\widetilde{F}_1}^{\widetilde{F}_2} \hat{F}\big(\widetilde{F}\big)\,d\theta\big(\widetilde{F}\big) - \widetilde{R}_0, \text{ при } \int_{\widetilde{F}_1}^{\widetilde{F}_2} \widetilde{F}\,d\theta\big(\widetilde{F}\big) = \widetilde{R}_0, \quad \int_{\widetilde{F}_1}^{\widetilde{F}_2} d\theta\big(\widetilde{F}\big) = 1, \quad (3)$$

где $\hat{F}\big(\widetilde{F}\big) = \sup_{\xi:\,\widetilde{F}(\xi)=\widetilde{F}} F(\xi)$.

Задача (3) путем дискретизации может быть сведена к задаче линейного программирования частного вида, для которой алгоритм полного перебора вершин имеет трудоемкость, квадратичную по числу интервалов разбиения. Именно таким способом получены приводившиеся значения оценок доверительных интервалов.

Однако оказывается, что для $\hat{F}\big(\widetilde{F}\big)$ можно указать простую линейную оценку.

На рисунках 1–3 отображены пары $\big(\widetilde{F}(\xi), F(\xi)\big)$, соответствующие различным ξ. Для

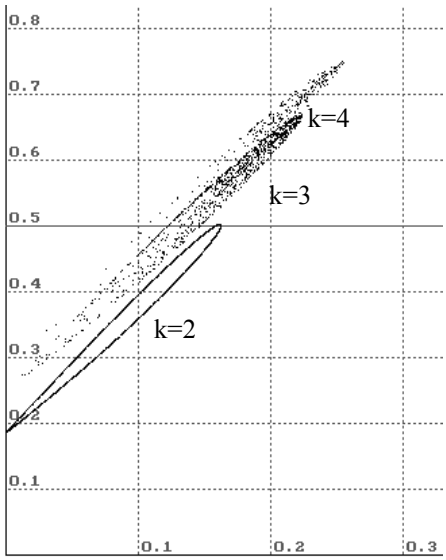

Рис. 1.

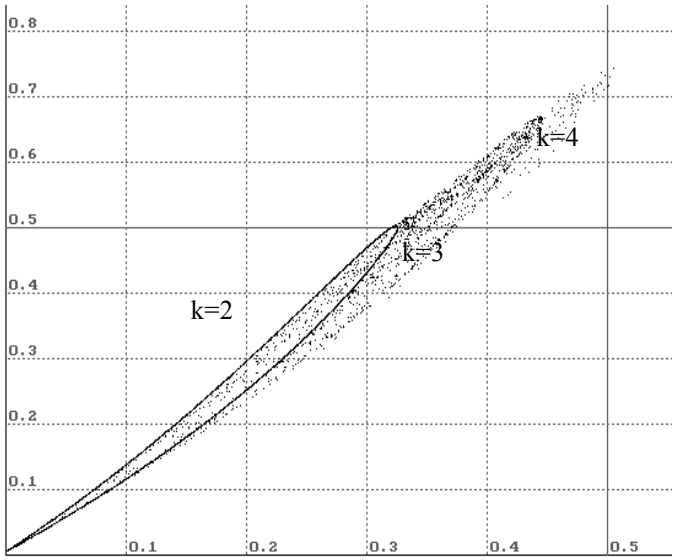

Рис. 2.

случая двух классов точки образуют сплошную линию в виде лепестка. При большем k – числе классов, ξ становится вектором, компоненты которого есть соответственно вероятности каждого образа. Чтобы изобразить множество $\big(\widetilde{F}(\xi), F(\xi)\big)$ в этом случае вектор ξ ра-

зыгрывался случайно в соответствии с равномерным распределением. Как видно на графике, форма лепестка сохраняется, но точки заполняют всю его площадь. На рисунке 1 приведены графики для $M=1$, на рис. 2 — для $M=5$; на рис. 3 графики приведены совместно, причем координаты точек нормированы делением на $F_{\max}=1-\frac{1}{k}$.

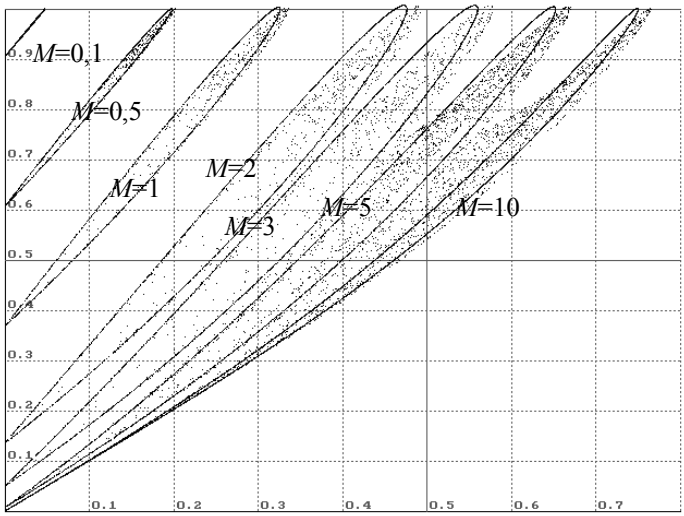

Рис. 3.

Как можно заметить, при всех k и M функция $\hat{F}\left(\widetilde{F}\right)$ может быть сверху весьма точно аппроксимирована некоторой прямой $\hat{F}\left(\widetilde{F}\right)=\hat{F}_1+\lambda(M)\widetilde{F}$.

Утверждение. Для вариационной задачи (3) справедлива следующая оценка: $\varepsilon_2+\widetilde{R}_0\le\hat{F}_1+\lambda(M)\widetilde{R}_0$

Сравним (при $k=2$) доверительные интервалы для вероятности ошибки, получаемые в соответствии с критериями Вапника–Червоненкиса и предложенным. Для этого оценку ε_1' будем находить из уравнения $\left(\varepsilon_1'\right)^2=4\delta_0^2\left(\varepsilon_1'+\widetilde{R}_0-\left(\varepsilon_1'+\widetilde{R}_0\right)^2\right)$, где δ_0 – максимальное значение доверительного интервала (достигается при $\xi(\theta)\equiv 0{,}5$, т. е. при максимальной вероятности ошибки).

На рис. 4 приведены графики (при $M=5$) для длины доверительного интервала в зависимости от $\widetilde{R}_0$.

Приведенные результаты показывают, что критерий Вапника–Червоненкиса в дискретном случае может быть существенно улучшен.

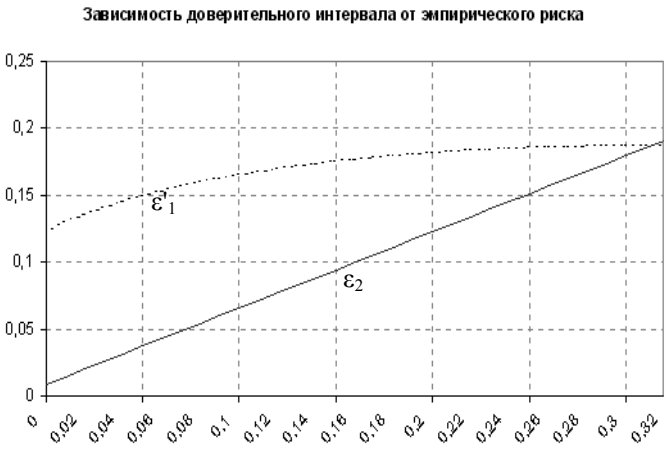

Рис. 4.

ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ ОЦЕНОК

Рассмотрим метод использования предложенных оценок в реальных задачах.

Пусть $\{X_1,\dots,X_n\}$ – разнотипный набор переменных и D_l – множество допустимых значений переменной X_l. Также пусть $Y\in D_Y=\{1,\dots,k\}$ – целевая переменная. Обозначим $D=\prod_{l=1}^{L}D_l$ – пространство значений переменных.

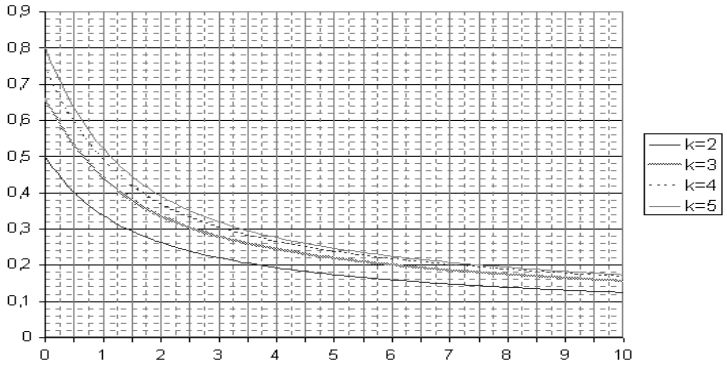

Рис. 5.

Метод оценивания вероятности ошибки основывается на гипотезе, что зависимость длины доверительного интервала от $\widetilde{R}_0$ в общем случае будет близкой к найденной ли-

нейной оценке, параметры $\hat{F}_1$ и λ которой однозначно определяются M. Примеры графиков зависимости δ_0 от M приведены на рис. 5.

Параметр M можно оценить как $\frac{N}{\log_2 \mathbf{C}}$, где $\mathbf{C}$ – емкость класса решающих правил.

Однако данная оценка будет в точности адекватна только для алгоритма выбора решающей функции, осуществляющего полный перебор. Для алгоритмов направленного поиска фактическая емкость может оказаться меньше.

Оценить фактическую емкость алгоритма построения решающей функции можно путем моделирования по следующей схеме.

Рис. 3.

Задаемся равномерным распределением в пространстве $D \times D_Y$.

В соответствии с равномерным распределением моделируем выборки того же объема, что и выборка в реальной задаче.

Исследуемым алгоритмом построения решающей функции строим решающее правило и находим для него значение эмпирического риска.

Многократным моделированием оцениваем $\widetilde{R}_U$ – среднее значение эмпирического риска.

Учитывая, что вероятность ошибки для любого правила при равномерном распределении есть $R_U = 1 - \frac{1}{k}$, получаем $\delta_0 = R_U - \widetilde{R}_U$.

Осталось заметить, что δ_0 однозначно связана с M, а значит, параметры линейной оценки для ε_2 полностью определены.

ИССЛЕДОВАНИЕ ПРИМЕНИМОСТИ

Проверим, насколько гипотеза об универсальной зависимости вероятности ошибки от эмпирического риска адекватна действительности.

Рассмотрим модельный пример непрерывной переменной $X \in [0,1]$, и $Y \in \{0,1\}$.

Будем использовать алгоритм построения решающей функции, минимизирующий эмпирический риск на классе решающих функций, разбивающих $[0,1]$ на два интервала.

Далее на $D_X \times D_Y$ случайно задаем кусочно-постоянные распределения, для которых моделируем по 1000 выборок, на которых находим средние R и $\widetilde{R}$.

Полученные точки $(\widetilde{R}, R)$ изображены на рис. 6 для $N = 6$ и на рис. 7 для $N = 25$.

Вычисляя δ_0, определяем, что случай $N = 6$ соответствует $M = 1$ ($\delta_0 = 0{,}33$), а $N = 25$ соответствует $M = 5$ ($\delta_0 = 0{,}18$).

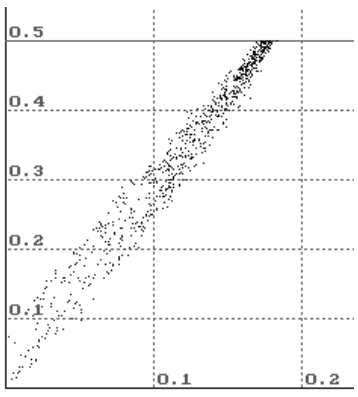

Рис. 6.

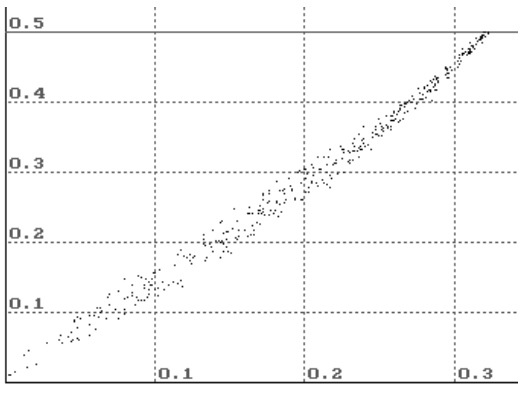

Рис. 7.

Как можно заметить, приведенные графики хорошо согласуются с соответствующими им для дискретного случая (рис 1–2). При этом, асимптотическая оценка хорошо приближает среднюю вероятность ошибки и для небольших N.

Рассмотренный метод оценивания вероятности ошибки также был опробован на реальных данных (задача определения культурной принадлежности наконечников стрел по характеристикам их формы).

Размерность признакового пространства составляла 10, объем обучающих выборок 100, объем контрольных выборок 50.

Задача классификации решалась 10 раз. При этом средний эмпирический риск составил 0,17, а средняя вероятность ошибки (на контрольной выборке) 0,22. Данный результат согласуется с оценкой для ε_2, которая при $\widetilde{R}_0 = 0{,}17$ составляет 0,07.

Работа выполнена при поддержке РФФИ, проект № 980100673.

ЛИТЕРАТУРА

1. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974. 415 с.
2. Лбов Г. С., Старцева Н. Г. Сложность распределений в задачах классификации. // Доклады РАН, 1994. Том 338 № 5.
3. Неделько В. М. "Оценивание доверительного интервала для вероятности ошибки решающей функции распознавания по эмпирическому риску". // Доклады Всероссийской конференции "Математические методы распознавания образов", Изд-во ВЦ РАН, Москва, 1999, с. 88–90.