СТАТИСТИЧЕСКИЕ ПРОБЛЕМЫ УПРАВЛЕНИЯ
1984 ВЫПУСК 66

УДК 519.281

ВЫЧИСЛЕНИЕ БАЙЕСОВСКОГО РИСКА ПРИ ПОМОЩИ КЕС ОЦЕНОК ПЛОТНОСТИ ВЕРОЯТНОСТИ

Софья ГАЛФАЯН, Ашот ЧИЛИНГАРЯН

Рассматриваются методы оценивания байесовского риска при помощи КЕС оценок функции плотности вероятности. Приводятся адаптивные КЕС оценки, позволяющие повысить точность восстановления неизвестной плотности.

I. Введение. Одной из самых важных задач в распознавании образов и многомерном статистическом анализе является вычисление вероятности ошибочной классификации (ВОК). ВОК используется как мера различия многомерных распределений, представленных обучающими выборками (ОВ), как мера оптимальности выбранной процедуры классификации и т.п. Кроме того, в задачах оценивания доли $p(\omega_1)$ при классификации смеси распределений

$$p(\vec{x}) = p(\omega_1)\cdot p(\vec{x}/\omega_1) + (1-p(\omega_1))\cdot p(\vec{x}/\omega_2) \qquad (I.1)$$

по обучающей выборке необходимо вычислить вероятности "перекачки" $P_{\omega_1\to\omega_2}$ и $P_{\omega_2\to\omega_1}$, т.е. вероятность того, что вектор $\vec{x}$, принадлежащий классу ω_1 с условной плотностью $p(\vec{x}/\omega_1)$ будет отнесен к классу ω_2, и наоборот.

Если альтернативные распределения заданы явно, т.е. условные плотности $p(\vec{x}/\omega_i)$ и априорные вероятности $p(\omega_i)$ известны, то

при классификации можно использовать оптимальное (байесовское) решающее правило (РП) [1].

$$d\vec{X} = \begin{cases} \omega_1 & \text{, если } p(\omega_1/\vec{X}) > p(\omega_2/\vec{X}) \\ \omega_2 & \text{в противном случае.} \end{cases} \qquad (I.2)$$

Апостериорные плотности определяются согласно формуле Байеса

$$p(\omega_i/\vec{X}) = \frac{p(\omega_i)\cdot p(\vec{X}/\omega_i)}{p(\vec{X})} \qquad (I.3)$$

Байесовское РП обеспечивает минимум ошибок среди всех возможных РП, и, следовательно, байесовская ошибка (равная байесовскому риску при использовании простой функции потерь) является минимальной и может служить для выбора наилучшего комплекса признаков.

В дальнейшем без потери общности ограничимся обучающими выборками с двумя классами и равными априорными вероятностями. Так как будет использоваться простая функция потерь, то не будет разницы между употреблениями выражений вероятности ошибочной классификации и риска процедуры.

Ошибка, совершаемая при классификации вектора $\vec{X}$ байесовским РП, равна

$$\tau^{B}(\vec{X}) = \min\left\{ p(\omega_1/\vec{X}),\ p(\omega_2/\vec{X}) \right\} \qquad (I.4)$$

и, следовательно, байесовский риск равен

$$R^{B} = E\left\{\tau^{B}(\vec{X})\right\} = \int_{-\infty}^{\infty} \tau^{B}(\vec{X})\cdot p(\vec{X})\, d\vec{X}. \qquad (I.5)$$

В практических задачах вид условных плотностей редко бывает известным, и вычисление риска осуществляется с помощью непараметрического оценивания плотности вероятности.

В настоящее время предложено несколько таких процедур [2], использующих КБС ($\vec{R}$-ближайших соседей) и парзеновские оценки. Процедуры вычисления $\vec{R}^{B}$ отличаются способом использования ОВ и методом оценивания плотности. Точность оценок, кроме того, зависит от размерности пространства, объема ОВ и неизвестного значения R^{B}.

Взаимосвязь этих факторов была рассмотрена Ш. Раудисом в ряде работ (см.,напр.,[3]).

Преимущество U метода скользящего экзамена (leave one out) над R и H методами (обучение и экзамен проводится по одной и той же выборке; выборка делится пополам, по одной половине проводится обучение, а другая используется для экзамена) также было показано в ряде работ (см., напр.,[4]).

Нашей целью будет исследование зависимости оценок от параметров и модификаций КВС метода и выбор метода, наилучшего при ограниченных ОВ. Способом исследования будут вычислительные эксперименты с использованием выборок из нормального распределения. Качество методов будет определяться среднеквадратичным отклонением от R^*, вычисленным согласно

$$R^* = \Phi(-\mathcal{D}_M/2)$$

где Φ - функция кумулятивного нормального распределения, а $\mathcal{D}_M$ - расстояние Махаланобиса.

2. Метод оценивания байесовского риска. Одним из первых предложенных методов оценивания R^* был метод эмпирического подсчета ошибок [5], в котором непосредственно оценивается значение интеграла (I.5).

Введем случайную величину

$$\eta(\vec{X}_i) = \begin{cases} 0, & \text{если } \vec{X}_i \text{ классифицируется правильно,} \\ 1, & \text{в противном случае} \end{cases} \qquad (2.1)$$

Оценку риска получим в виде

$$R_э = \frac{1}{M} \sum_{i=1}^{M} \eta(\vec{X}_i), \qquad \vec{X}_i \in p(\vec{X}) \qquad (2.2)$$

а дисперсию оценки, равной

$$\sigma^2(R_э) = \frac{1}{M} R^*(1-R^*) \qquad (2.3)$$

Другой метод основан на оценивании подынтегрального выражения (I.4) [6]

$$\hat{r}(\vec{X}_i) = \min\left\{\hat{p}(\omega_1/\vec{X}),\ \hat{p}(\omega_2/\vec{X})\right\} \qquad (2.4)$$

и

$$R_p = \frac{1}{M}\sum_{i=1}^{M} \hat{r}(\vec{X}_i), \quad \vec{X}_i \in p(\vec{X}) \qquad (2.5)$$

Дисперсия оценки равна

$$\sigma^2(R_p) = \frac{1}{M} R^B(1-R^B) - \frac{1}{2}\frac{R^B}{M} \qquad (2.6)$$

Этот результат кажется парадоксальным, так как во втором случае не используется информация об истинной принадлежности вектора $\vec{X}$. Уменьшение дисперсии связано с тем, что R_p может принимать любые рациональные значения, тогда как значения $R_э$ ограничены рядом отношений целых чисел $m_{изм}/M$, что обуславливает их больший разброс. Однако, как мы увидим далее, одна лишь кучность оценок не является серьезным преимуществом.

Интеграл (I.5) можно представить в виде

$$R^B = \int_{-\infty}^{\infty} r^B(\vec{X})\cdot p(\vec{X})\, d\vec{X} = \int_{\Gamma} r^B(\vec{X}) \qquad (2.7)$$

где Γ – область признакового пространства, в которой обе условные плотности отличны от нуля – то есть область пересечения

Точность оценок определяется тем, насколько хорошо оцениваются условные плотности в области Γ. Ввиду того, что при малых объемах ОВ и/или малых значениях R^B в область Γ попадает небольшое число векторов $\vec{X}$, оценки типа 2.5 подвержены сильным флуктуациям, поэтому нами предложена модификация этого метода, позволяющая несколько сгладить ошибки, вносимые флуктуациями в ОВ. Используя допущение $p(\omega_1) = p(\omega_2) = 0.5$, можно представить интеграл (I.4) в виде

$$R^B = 0.5\int_{\Gamma} \min\left\{p(\vec{X}/\omega_1),\ p(\vec{X}/\omega_2)\right\} d\vec{X} \qquad (2.8)$$

и непосредственно вычислить значение интеграла:

$$R_э = \Delta X \cdot \sum_{i=1}^{a} \min \left\{ p(\vec{X}_i/\omega_1),\ p(\vec{X}_i/\omega_2) \right\}, \qquad X_i = X_{i-1} + \Delta X \tag{2.9}$$

3. ИБС методы оценивания функции плотности вероятности. Локальную оценку плотности в точке $\vec{X}$ признакового пространства получим в виде

$$\hat{p}(\vec{X}) = U(\vec{X})/V(\vec{X}), \tag{3.1}$$

где $U(\vec{X})$ - покрытие области Γ, содержащей точку $\vec{X}$

$$U(\vec{X}) = \int_{\Gamma} p(\vec{Y})\, d\vec{Y} \tag{3.2}$$

$V(\vec{X})$ - объем области Γ. Несмещенной оценкой покрытия является K/M, где K - число представителей ОВ, попавших в область Г, поэтому

$$\hat{P}_{K,M}(\vec{X}) \approx \frac{K}{M \cdot V_k(\vec{X})} \tag{3.3}$$

Основным фактором в выборе метода оценивания является способ определения области Γ. Если разбить признаковое пространство на ячейки одинаковой величины, придем к простой гистограмме, если поместить точку, в которой оценивается плотность, в центр ячейки, получим гистограмму Розенблатта [7]. Для состоятельности этих методов должны выполняться условия типа:

$$V \to 0 \qquad MV \to \infty, \tag{3.4}$$

чего невозможно добиться при сколь-нибудь значительных размерностях признакового пространства.

В ИБС методах фиксируется не объем ячейки V, а величина покрытия U. Независимо от истинной плотности требуется, чтобы оценки проводились по области Γ, содержащей K представителей ОВ [8].

$$U(\vec{X}) \cdot M = K, \quad \vec{X} \in p(\vec{X}) \tag{3.5}$$

Области Γ, удовлетворяющие этому соотношению, называются толерантными.

Успешному применению непараметрических методов препятствует наличие в них неизвестных параметров. В методе гистограммы – это размер ячейки, в парзеновском методе – тип и ширина ядра, в КБС методе – значение параметра K. Эти параметры зависят от неизвестной функции плотности, объема ОВ и размерности пространства, поэтому практические рекомендации часто противоречивы. Для КБС метода значение K варьирует от $K=1$ до $K=\sqrt{N}$ или $N/2$. Исследование зависимости среднеквадратичной ошибки (MSE) КБС оценок плотности от объема ОВ и истинной плотности (рис. I) показало, что значение параметра K, соответствующее минимуму MSE, очень сильно зависит от точки, в которой плотность оценивается. Если для точек вблизи моды распределения ($x=0,1$) MSE минимальна при $x \approx N/2$, то для удаленных от моды "периферийных" точек ($x \geq 2$) наилучшее K равно 2-5.

Для того, чтобы оценки не так сильно зависели от параметра, были предложены методы, связанные с использованием не одного, а нескольких значений параметра K. Вычисляется ряд оценок $\{p_{i,N}(\vec{X})\}$, $i=1,Q$ и с его помощью конструируется усредненная КБСЗ оценка [9]

$$p_{Q,N}(\vec{X}) = \frac{1}{Q} \cdot \sum_{i=1}^{Q} p_{i,N}(\vec{X}) \qquad (3.6)$$

Эта оценка более детально использует информацию о КБС окрестности точки $\vec{X}$ и обеспечивает более устойчивые к выбору параметра (в случае КБСЗ – параметр Q) оценки. Однако из-за флуктуаций в ОВ в ряде оценок плотности могут быть значительные отклонения, поэтому вместо среднеарифметического усреднения нами использовалась линейная комбинация порядковых статистик вариационного ряда оценок плотности

$$p_{[Q],N}(\vec{X}) = \sum_{i=1}^{Q} a_i \, p_{[i],N}(\vec{X}), \quad \sum_{i=1}^{Q} a_i = 1, \qquad (3.7)$$

где коэффициенты при порядковых статистиках выбираются так, чтобы крайние члены вариационного ряда не вносили вклад в оценку, например,

$$a_i = \begin{cases} 1, \text{ если } i = [Q/2]+1 \\ 0, \text{ в противном случае.} \end{cases} \qquad (3.8)$$

Медианные оценки (3.7), (3.8) хорошо восстанавливают плотность в районе моды распределения, однако значительно завышают плотность в периферийных областях, поэтому в этих областях следует оценивать плотность иначе. Окончательное адаптивное КБС оценивание принимает вид

$$p_{AA,M}(\vec{X}) = \begin{cases} p_{j,M}(\vec{X}), \text{ если } \varsigma(\vec{X})/\bar{\varsigma}_M > \alpha \quad j=3,\ \alpha=3 \\ p_{[Q/2]+1,M}(\vec{X}) \quad \text{в противном случае, } Q = M/2, \end{cases}$$

где $\varsigma(\vec{X})$ – размер локальной области вокруг точки $\vec{X}$. Чем больше этот размер в сравнении со средним по ОВ, тем в более "разряженной" области находится $\vec{X}$.

Таким образом снимается неопределенность в выборе параметров K и Q в КБС2 и КБС3 методах. Адаптивное КБС правило, в котором способ определения параметров i, Q, α не зависит от неизвестных плотностей и объема ОВ, а плотность определяется информацией, заключенной в ОВ, позволяет (рис. 2) строить оценки, которые лучше, чем полученные с любым фиксированным параметром K или Q (в метрике взвешенного среднеквадратичного отклонения IMSE).

4. Обсуждение результатов. Сравнительный анализ методов оценивания байесовского риска проводился путем вычисления $E\{R_{n,p,s}\}$ с использованием выборок из многомерного нормального распределения. Размерность признакового пространства изменялась от 1 до 12, объем ОВ – от 25 до 500. Значение R^* принималось равным 0.1, 0.234, 0.3085 и 0.5. Вычислялись одновременно ошибки, соответствующие оценкам КБС2 с K от 1 до $M/2$, КБС3 и $КБС_{АД}$ оценкам. Усред-

нение проводилось по 10 расчетам с использованием независимых выборок.

На рис. 3 и 4 приведены результаты вычисления для $M=2$, $M=100$, $R^{*}=0.234$, где N – размерность пространства, M – объем ОВ.

Крайние правые точки на рисунках соответствуют адаптивным КБС оценкам плотности. Интересно отметить, что для малых K метод эмпирического подсчета ошибок значительно завышает, а метод усреднения апостериорной плотности занижает R^{*}. Это происходит оттого, что минимальная плотность в (2.4) соответствует оценкам плотности в периферийных точках и, как видно из рис. 1, КБС2 оценки для таких точек сначала занижают плотность (до $K=5$), а затем завышают ее. Это связано с тем, что с ростом K (числитель формулы (3.3)), а знаменатель, равный объему области, практически не изменяется. Отклонение $E\{R_3\}$ от R^{*} связано с ошибками в оценивании плотности, при которых меняется соотношение плотностей, т.е. если истинные плотности связаны соотношением $p(\vec{x}/\omega_1) > p(\vec{x}/\omega_2)$, то ошибки в оценивании приводят к $\hat{p}(\vec{x}/\omega_1) < \hat{p}(\vec{x}/\omega_2)$. Из рис. 1 видно, что вероятность такой ошибки при использовании КБС2 метода оценивания зависит от K и от значения истинной плотности в точке $\vec{x}$. И так как при малых K плотность оценивается с большими ошибками для областей, в которых находится основная часть "вероятностной массы" распределения, то можно показать, что использование КБС2 оценок плотности для вычисления байесовского риска будет приводить к завышению значения R_3 при малых K.

Тот факт, что в методе R_3 используется относительная величина оценок плотности, а в R_p – абсолютная, приводит к тому, что среднеквадратичное отклонение последнего метода значительно выше, чем у R_3. Это особенно заметно при больших размерностях признакового пространства, где $MSE(R_p)$ превышает $MSE(R_3)$ на порядок. Использование модификации R_p метода – R_6 позволяет уменьшить MSE,

Таблица I

Сравнение методов оценивания байесовского риска с использованием адаптивных КБС оценок плотности

N	M	R^B	$\sigma(R_э)$	$\sigma(R_р)$	$\hat{\sigma}(R_э)$	$\hat{\sigma}(R_р)$	$\widehat{MSE}(R_э)$	$\widehat{MSE}(R_р)$	$MSE(R_э)$
I	I00	0.3085	0.046	0.024	0.039	0.026	0.0005	0.00I3	0.00I0
I	200	0.3085	0.033	0.0I7	0.026	0.0I5	0.0003	0.0008	0.0005
2	50	0.I340	0.048	0.03I	0.044	0.027	0.0030	0.0039	0.0032
8	50	0.5	–	–	0.05	0.0I2	0.005	0.044	–
I0	50	0.5	–	–	0.047	0.0I7	0.004	0.046	–

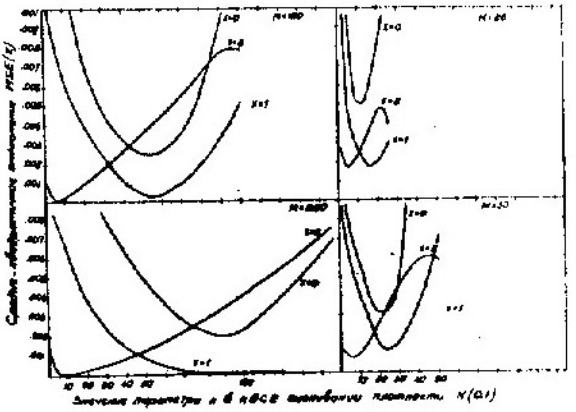

Рис. I. Зависимость среднеквадратичной ошибки (MSE(x)) от значения параметра K и точки X к КБС оценивании функции плотности вероятности стандартного нормального распределения N(0,1).

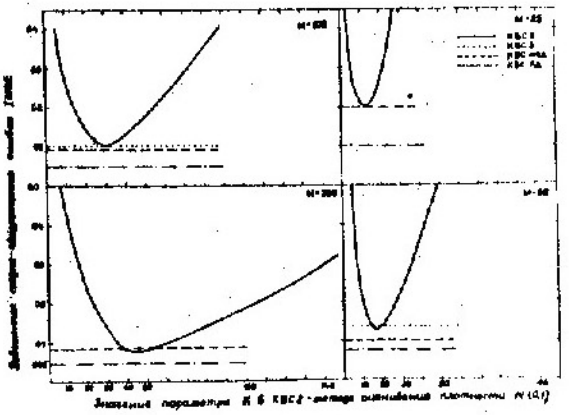

Рис. 2. Взвешенная среднеквадратичная ошибка (IMSE) различных модификаций КБС методов оценивания стандартного нормального распределения N(0,1).

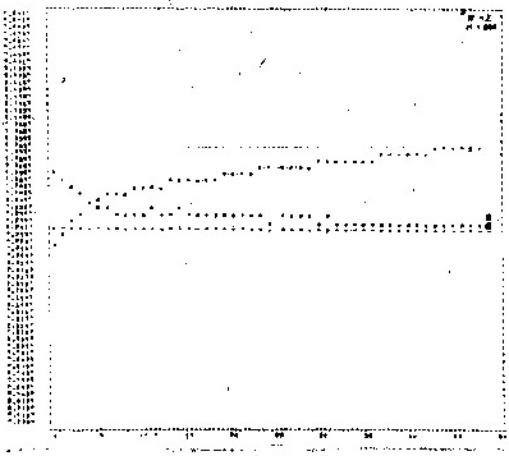

Рис. 3. Зависимость оценок байесовского риска R_p – знак и R_p – знак + от значения параметра K в КБС$_2$ методе оценивания. Знаком "–" показано значение R^B=0.23; знаками X и + – значения, соответствующие адаптивным оценкам функции плотности вероятности. Размерность признакового пространства N =2, объем ОВ N =100

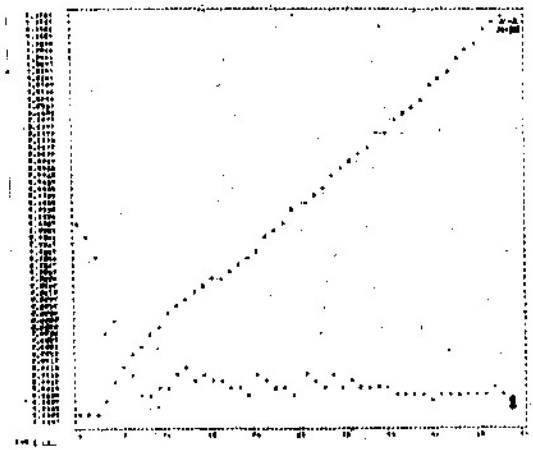

Рис. 4. Среднеквадратичная ошибка оценивания байесовского риска R_p – знак X и R_p –знак +; X и + – значения, соответствующие адаптивным оценкам функции плотности вероятности.Размерность признакового пространства N =2, объем ОВ N =100

но следует отметить, что этот метод применим только для не очень больших размерностей, так как в противном случае объем вычислений резко возрастет (для N =5 и числа разбиения каждой оси на 10 интервалов требуется оценить плотность в 10^5 точках).

5. Выводы. Исследование методов вычисления байесовского риска показало преимущество метода эмпирического подсчета ошибок с использованием адаптивных КЕС оценок плотности вероятности. Указанный метод был реализован в рамках библиотеки стандартных модулей прикладного статистического анализа центра данных. Использование метода для определения химического состава первичного космического излучения и выбора модельно зависимых комбинаций первичных физических признаков при обработке данных экспериментов с рентген-эмульсионными камерами позволит значительно повысить точность определения характеристик космического излучения [10].

ЛИТЕРАТУРА

1. РАЙФА Г., ШЛЕЙФЕР Р. Прикладная теория статистических решений. - М.:Статистика, 1977. - 306 с.

2. TOUSSAINT G.T. Bibliography of misclassification. - IEEE Trans. Inf.Theory, 1974, v. IT-20, p.472-479.

3. РАУДИС Ш. Оценка вероятности ошибки классификации. - В сб.: Статистические проблемы управления. Вильнюс, 1973, вып. 5, с.9-44.

4. КОТЛОВ В., БУТОРИН А. Экспериментальное сравнение различных оценок ошибки классификации. - В сб.:Статистические проблемы управления, Вильнюс, 1981, вып. 50, с. 78-86.

5. COVER J.M., HART P.E. Nearest neighbor pattern classification.- IEEE Trans.Inf.Theory,1967, v.IT-13, p. 21-27.

6. FUKUNAGA K., HOSTETLER L.D. K-nearest neighbor Bayes risk estimation. - IEEE Trans.Inf.Theory, 1975, v. IT-21, p. 285-293.

7. ROSENBLATT M. Remarks on some nonparametric estimates of a density function. - Ann.Math.Stat., 1956, v.42, p. 1815-1842.

8. ФУКУНАГА К. Введение в статистическую теорию распознавания образов. - М.:Наука, 1979. - 367 с.

9. ПАТРИК Э.А. Основы теории распознавания образов. – М.: Сов.радио, 1980. – 407 с.

IO. CHILINGARIAN A.A., GALFAYAN S.Kh., ZAZYAN M.Z., DUNAEVSKY A. Selection of model-dependent feature in photon-hadron families, registered by roentgen-emulsion chambers. – Proc. on 18th ICRC, Bangalor, v. 5, p. 487-490.

BAJESINĖS RIZIKOS SKAIČIAVIMAS ĮVERTINANT TIKIMYBINĘ TANKIO FUNKCIJĄ KAK (K-ARTIMIAUSIŲ KAIMYNŲ) METODU

Sofija GALFAJAN, Ašot ČILINGARJAN

Išnagrinėti bajesinės rizikos įvertinimo metodai, įvertinant tikimybinio tankio funkciją KAK metodu. Pasiūlyti adaptyvūs KAK įvertinimai, įgalinantys padidinti nežinomo tankio atstatymo tikslumą.

CALCULATION OF THE BAYES RISK BY ESTIMATING A PROBABILITY DENSITY FUNCTION BY A KNN(K-NRAREST NEIGHBORS) METHOD

Sofija GALFAYAN, Ašot CHILINGARAYAN

The Bayes risk estimation methods are described by estimating a probability density function by a KNN method. Adaptive KNN estimations are also presented which increase the accuracy of restoration of an unknown density.