УДК 519.2

Е. Е. ЖУК

СТАТИСТИЧЕСКОЕ ОТНЕСЕНИЕ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ К КЛАССАМ, ЗАДАННЫМ ОБУЧАЮЩИМИ ВЫБОРКАМИ

Белорусский государственный университет

(Поступила в редакцию 13.12.2013)

1. Математическая модель и постановка задачи. Пусть в пространстве наблюдений R^N ($N \geq 1$) зафиксировано $L \geq 2$ классов $\{\Omega_i\}_{i \in S}$ ($S = \{1, \ldots, L\}$ – множество номеров классов), заданных своими обучающими выборками [1, 2]:

$$X^{(i)} = \{x_t^{(i)}\}_{t=1}^{n_i},\ i \in S. \tag{1}$$

Выборка $X^{(i)}$ объема $n_i \geq 1$ образована независимыми в совокупности одинаково распределенными случайными N-векторами-наблюдениями $x_t^{(i)} \in R^N$, $t = \overline{1, n_i}$, имеющими одну и ту же плотность распределения:

$$p_i(x) \geq 0,\ x \in R^N : \int_{R^N} p_i(x)dx = 1,\ i \in S. \tag{2}$$

Плотности распределения $\{p_i(\cdot)\}_{i \in S}$ из (2), определяющие классы $\{\Omega_i\}_{i \in S}$, неизвестны. Задача заключается в отнесении вновь поступившего наблюдения $x \in R^N$ к одному из классов $\{\Omega_i\}_{i \in S}$ [3–5]. Причем распределение вероятностей наблюдения x, вообще говоря, не совпадает ни с одним из распределений (2), задающих классы $\{\Omega_i\}_{i \in S}$, и также неизвестно.

В случае известных плотностей $\{p_i(\cdot)\}_{i \in S}$ из (2) для решения поставленной выше задачи в [3] было предложено использовать решающее правило (РП) максимального правдоподобия:

$$d = d(x) = \arg\max_{i \in S} p_i(x),\ x \in R^N, \tag{3}$$

которое относит наблюдение x к тому классу с номером $d(x) \in S$, появление в котором для него «наиболее правдоподобно» ($p_d(x) \geq p_i(x)$, $i \in S$).

Эффективность РП (3) в [3] было предложено характеризовать риском:

$$r = r(d) = 1 - \min_{i \in I} P^{(i)},\ I = \{k : R_k = \min_{j \in S} R_j\}; \tag{4}$$

$$P^{(i)} = \mathrm{P}\{d(x) = i\} = \int_{R^N} \prod_{\substack{l \in S \\ l \neq i}} U\big(p_i(x) - p_l(x)\big) p(x)dx,\ i \in S,$$

где $U(z) = \{1,$ если $z \geq 0;\ 0,$ если $z < 0\}$ – единичная функция Хэвисайда,

$$p(x) \geq 0,\ x \in R^N : \int_{R^N} p(x)dx = 1, \tag{5}$$

– плотность распределения вероятностей наблюдения, подлежащего отнесению, а через $\{R_j\}_{j\in S}$ в (4) обозначены межклассовые расстояния [1–5], характеризующие «степень близости» наблюдения x с плотностью (5) к классам $\{\Omega_i\}_{i\in S}$ с плотностями (2). Для определения межклассовых расстояний в работе [5] было предложено использовать направленную дивергенцию Кульбака:

$$R_j = \int_{R^N} \ln\left(\frac{p(x)}{p_j(x)}\right) p(x)dx,\ j \in S. \tag{6}$$

Риск *r* из (4), (6) – это максимальная вероятность не отнести при помощи РП (3) наблюдение *x* к тому классу из $\{\Omega_i\}_{i\in S}$, к которому оно ближе в смысле межклассовых расстояний (6). Чем меньше значение риска *r*, тем эффективнее решение, принимаемое РП (3).

2. Решающие правила на основе гистограммных оценок плотностей. При неизвестных плотностях $\{p_i(\cdot)\}_{i\in S}$ непосредственно применить РП (3) для решения задачи отнесения не представляется возможным. Воспользуемся в данном случае наличием обучающих выборок (1) и построим РП по подстановочному принципу [1, 2]:

$$\hat{d} = \hat{d}(x) = \arg\max_{i\in S} \hat{p}_i(x),\ x \in R^N, \tag{7}$$

где $\hat{p}_i(\cdot)$ – статистическая оценка для плотности $p_i(\cdot)$ по выборке $X^{(i)}$ объема n_i ($i \in S$).

Поскольку здесь плотности $\{p_i(\cdot)\}_{i\in S}$, определяющие классы, считаются полностью неизвестными и неизвестен их вид, то в качестве статистических оценок $\{\hat{p}_i(\cdot)\}_{i\in S}$ в подстановочном РП (ПРП) (7) будем использовать гистограммные оценки [6, 7], разбив пространство наблюдений R^N на ячейки прямоугольной формы:

$$\hat{p}_i(x) = \frac{1}{n_i \prod_{j=1}^{N} h_j} \sum_{t=1}^{n_i} \mathrm{I}_{\Gamma(x)}\left(x_t^{(i)}\right),\ x \in R^N,\ i \in S, \tag{8}$$

где $\mathrm{I}_A(x) = \{1,$ если $x \in A;\ 0,$ если $x \notin A\}$ – индикатор множества $A \subset R^N$, $\Gamma(x)$ – ячейка, в которую попала точка-наблюдение $x = \left(\tilde{x}_j\right)_{j=1}^N \in R^N$:

$$\Gamma(x) = \bigtimes_{j=1}^{N} \left[\left[\frac{\tilde{x}_j}{h_j}\right] h_j, \left(\left[\frac{\tilde{x}_j}{h_j}\right] + 1\right) h_j\right), \tag{9}$$

$[z]$ – целая часть числа $z \in R$, «×» – символ декартова произведения, а h_j – коэффициент «размытости» [6, 7] по *j*-й компоненте *N*-вектора $x = \left(\tilde{x}_j\right)_{j=1}^N$.

Отметим, что все оценки-гистограммы (8) построены по одному и тому же разбиению пространства R^N на ячейки, причем все ячейки (9) имеют один и тот же «объем»: $\mu(\Gamma(x)) = \prod_{j=1}^{N} h_j$ $\forall x \in R^N$, где $\mu(\cdot)$ – мера Лебега в R^N.

Введем обозначение ($i \in S$, $x \in R^N$):

$$n^{(i)}(x) = \sum_{t=1}^{n_i} \mathrm{I}_{\Gamma(x)}\left(x_t^{(i)}\right) \geq 0 \tag{10}$$

– число наблюдений из выборки $X^{(i)}$ объема n_i, попавших в ячейку $\Gamma(x)$. Тогда запись оценок (8) упрощается:

$$\hat{p}_i(x) = \frac{n^{(i)}(x)}{n_i \mu(\Gamma(x))},\ x \in R^N,\ i \in S, \tag{11}$$

а ПРП (7) с учетом (11) и того факта, что среди $\{\hat{p}_i(x)\}_{i\in S}$, по построению (11), (10), могут быть совпадающие по значению, примет вид ($x \in R^N$):

$$\hat{d}(x) \in \hat{D}(x),\ \hat{D}(x) = \left\{ k : \frac{n^{(k)}(x)}{n_k} = \max_{i\in S} \frac{n^{(i)}(x)}{n_i} \right\}. \tag{12}$$

Если среди $\{\hat{p}_i(x)\}_{i\in S}$ нет совпадающих по значению (множество $\hat{D}(x)$ в (12) содержит один элемент: $\left|\hat{D}(x)\right| = 1$), то ПРП (12) упрощается и выносит решение относительно x однозначно:

$$\hat{d}(x) = \arg\max_{i\in S} \frac{n^{(i)}(x)}{n_i}. \tag{13}$$

Если вдобавок ко всему объемы выборок из (1) одинаковы: $n_i = n_j$, $i \neq j \in S$, то

$$\hat{d}(x) = \arg\max_{i\in S} n^{(i)}(x), \tag{14}$$

и ПРП (14) имеет простой содержательный смысл: оно относит наблюдение x к тому классу, наблюдений из которого больше в ячейке $\Gamma(x)$.

Отметим также, что возможна ситуация, когда ячейка $\Gamma(x)$ вообще не содержит наблюдений из обучающих выборок (1): $n^{(i)}(x) = 0$, $i \in S$. В этом случае: $\left|\hat{D}(x)\right| = L$, и решение «полностью неопределено». Если в (12): $1 < \left|\hat{D}(x)\right| \le L$, $\sum_{i\in S} n^{(i)}(x) \neq 0$, то можно говорить, что наблюдение x «в одинаковой мере» относится к одному из классов с номерами $\hat{D}(x) \subseteq S$.

3. Асимптотическое исследование эффективности. Установим асимптотические свойства ПРП (12) при увеличении объемов обучающих выборок (1): $n_i \to +\infty$, $i \in S$. Сначала, по аналогии с [6], получим асимптотический результат для оценок плотностей из (8), (9).

Л е м м а. *Пусть плотности* $\{p_i(x)\}_{i\in S}$ *из* (2) *непрерывны и ограничены* ($\forall x \in R^N$), *а наблюдения в выборках* (1) *независимы в совокупности. Тогда в условиях асимптотики:*

$$h_j \to 0,\ j = \overline{1,N};\ n_i \prod_{j=1}^{N} h_j \to +\infty,\ n_i \to +\infty,\ i \in S, \tag{15}$$

оценки $\{\hat{p}_i(\cdot)\}_{i\in S}$ *из* (8), (9) *состоятельны по вероятности:*

$$\hat{p}_i(x) \xrightarrow{P} p_i(x),\ n_i \to +\infty,\ x \in R^N,\ i \in S. \tag{16}$$

Д о к а з а т е л ь с т в о. Зафиксируем $i \in S$ и $x \in R^N$. Введем в рассмотрение случайные величины ($t = \overline{1,n_i}$):

$$\xi^{(i)}_{t,n_i} = \frac{\mathrm{I}_{\Gamma(x)}\left(x_t^{(i)}\right) - \gamma^{(i)}(x)}{n_i \prod_{j=1}^{N} h_j}, \tag{17}$$

где $\gamma^{(i)}(x) = \int_{\Gamma(x)} p_i(y)dy$ – вероятность попадания наблюдений из выборки $X^{(i)} = \{x_t^{(i)}\}_{t=1}^{n_i}$ в ячейку $\Gamma(x)$ из (9) ($\gamma^{(i)}(x) = \mathrm{P}\{x_t^{(i)} \in \Gamma(x)\}$).

По аналогии с [6] устанавливаем, что $\{\xi_{t,n_i}^{(i)}\}_{t=1}^{n_i}$ из (17) в условиях леммы удовлетворяют закону больших чисел для серий [6, 8]:

$$\eta_{n_i}^{(i)} = \sum_{t=1}^{n_i} \xi_{t,n_i}^{(i)} \xrightarrow{P} 0\,,\ n_i \to +\infty\,. \tag{18}$$

Доказательство соотношения (18) полностью копирует аналогичное доказательство из [6] и здесь не приводится.

Далее, преобразовав $\eta_{n_i}^{(i)}$ из (18), имеем:

$$\eta_{n_i}^{(i)} = \hat{p}_i(x) - \frac{\gamma^{(i)}(x)}{\prod_{j=1}^{N} h_j}\,,$$

где $\hat{p}_i(x)$ – гистограммная оценка (8), (9) плотности $p_i(x)$. По теореме о среднем в асимптотике (15) ($h_j \to 0$, $j = \overline{1,N}$) получаем:

$$\frac{\gamma^{(i)}(x)}{\prod_{j=1}^{N} h_j} = \frac{\int_{\Gamma(x)} p_i(y)dy}{\mu(\Gamma(x))} \to p_i(x),$$

что, с учетом (18), и доказывает лемму.

Т е о р е м а. *Пусть в условиях леммы плотности* $\{p_i(x)\}_{i\in S}$ *из* (2) *совпадают между собой по значению разве что на множестве аргумента меры Лебега нуль:*

$$\mu\{x : p_i(x) = p_j(x)\} = 0,\ i \neq j \in S\,, \tag{19}$$

тогда в условиях асимптотики (15) ($x \in R^N$)*:*

$$\hat{d}(x) \xrightarrow{P} d(x)\,,\ \mathrm{P}\{\left|\hat{D}(x)\right| = 1\} \to 1\,, \tag{20}$$

где $\hat{d}(x) \in \hat{D}(x)$ *– ПРП* (12), *а* $d(x)$ *– РП* (3).

Д о к а з а т е л ь с т в о. Воспользуемся ПРП (12) в виде ($x \in R^N$):

$$\hat{d}(x) \in \hat{D}(x)\,,\ \hat{D}(x) = \{k : \hat{p}_k(x) = \max_{i\in S} \hat{p}_i(x)\}\,,$$

где $\{\hat{p}_i(x)\}_{i\in S}$ – оценки плотностей из (8), (9). Справедливы очевидные соотношения:

$$\hat{p}_{\hat{d}(x)}(x) = \max_{i\in S} \hat{p}_i(x)\,,\ \hat{d}(x) \in \hat{D}(x)\,;\ p_{d(x)}(x) = \max_{j\in S} p_j(x)\,,$$

с учетом которых в асимптотике (15) получаем (использован результат (16) леммы):

$$\left|\hat{p}_{\hat{d}(x)}(x) - p_{d(x)}(x)\right| = \left|\max_{i\in S} \hat{p}_i(x) - \max_{j\in S} p_j(x)\right| \le \max_{i\in S} \left|\hat{p}_i(x) - p_i(x)\right| \xrightarrow{P} 0\,,$$

откуда устанавливаем, что

$$\hat{p}_{\hat{d}(x)}(x) \xrightarrow{P} p_{d(x)}(x)\,. \tag{21}$$

Справедливость (20) следует из (21), условия (19) на плотности $\{p_i(x)\}_{i\in S}$ и известных теорем непрерывности [7].

Результат (20) теоремы говорит о том, что с ростом объемов $\{n_i\}_{i\in S}$ обучающих выборок (1) решение, выносимое ПРП (12), сходится по вероятности к соответствующему решению РП (3), и эффективность принимаемых ПРП (12) решений в асимптотике (15) можно по-прежнему характеризовать риском (4), (6). Сама асимптотика (15) накладывает ограничения на «скорость» стремления коэффициентов «размытости» $\{h_j\}_{j\in S}$ к нулю с ростом объемов выборок $\{n_i\}_{i\in S}$.

Однако непосредственно воспользоваться риском r из (4), (6) в качестве меры эффективности не представляется возможным, поскольку плотности $\{p_i(\cdot)\}_{i\in S}$ из (2) и плотность $p(\cdot)$ из (5) неизвестны.

4. Статистическое оценивание риска по результатам экспериментов. Пусть имеется выборка $X=\{x_t\}_{t=1}^{n}$ объема n, составленная из наблюдений $x_t \in R^N$, $t=\overline{1,n}$, относительно которых при помощи ПРП (12) решалась задача отнесения к классам $\{\Omega_i\}_{i\in S}$ на основе обучающих выборок $X^{(i)}=\{x_t^{(i)}\}_{t=1}^{n_i}$, $i\in S$, из (1). И пусть все наблюдения из выборки X имеют одну и ту же плотность (5).

Воспользуемся принятыми ПРП (12) решениями: $\hat{d}(x_t)\in\hat{D}(x_t)$, $t=\overline{1,n}$, и оценим риск r из (4), (6).

Преобразуем межклассовые расстояния $\{R_j\}_{j\in S}$ из (6):

$$R_j=\int_{R^N}\ln\bigl(p(x)\bigr)p(x)dx-\int_{R^N}\ln\bigl(p_j(x)\bigr)p(x)dx=\int_{R^N}\ln\bigl(p(x)\bigr)p(x)dx-\mathrm{E}\bigl\{\ln\bigl(p_j(x_t)\bigr)\bigr\},$$

заменим математическое ожидание, зависящее от номера класса «*j*», на его статистическую оценку типа арифметического среднего [2, 7] по выборке $X=\{x_t\}_{t=1}^{n}$ объема *n*, а неизвестную плотность $p_j(\cdot)$ – на оценку $\hat{p}_j(\cdot)$ из (11), и для I из (4) получим статистическую оценку:

$$\hat{I}=\arg\max_{j\in S}\frac{1}{n}\sum_{\substack{t=1\\ n^{(j)}(x_t)\neq 0}}^{n}\ln\left(\frac{n^{(j)}(x_t)}{n_j}\right)=\arg\max_{j\in S}\sum_{\substack{t=1\\ n^{(j)}(x_t)\neq 0}}^{n}\ln\left(\frac{n^{(j)}(x_t)}{n_j}\right). \qquad (22)$$

В (22) учтено, что в некоторые ячейки (9) могут не попасть наблюдения из соответствующих классов, а также использован тот факт, что вероятность совпадения по значению непрерывно распределенных случайных величин равна нулю [2, 8].

Далее оценим непосредственно риск *r*, подставив в (4) вместо *I* его оценку-число $\hat{I}$ из (22). Но сначала преобразуем (4), считая I одноточечным множеством (числом):

$$r=1-P^{(I)}=1-\mathrm{P}\{d(x_t)=I\}=\mathrm{E}\{1-\delta_{d(x_t)\,I}\},$$

где $\delta_{k,l}=\{1,$ если $k=l;$ $0,$ если $k\neq l\}$ – символ Кронекера, откуда будем иметь следующую оценку типа арифметического среднего:

$$\hat{r}=\frac{1}{n}\sum_{t=1}^{n}\Bigl(1-\mathrm{I}_{\hat{D}(x_t)}\bigl(\hat{I}\bigr)\Bigr). \qquad (23)$$

Оценка риска $\hat{r}$ из (23) на практике имеет простой содержательный смысл, являясь долей решений относительно наблюдений из выборки $X=\{x_t\}_{t=1}^{n}$ объема *n*, принимаемых ПРП (12) не в пользу класса с номером $\hat{I}$ из (22), который по результатам проведенного эксперимента считается ближайшим к наблюдениям из *X*.

Литература

1. *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Классификация и снижение размерности. М., 1989.

2. *Харин Ю. С., Зуев Н. М., Жук Е. Е.* Теория вероятностей, математическая и прикладная статистика. Минск, 2011.
3. *Жук Е. Е.* // Весці НАН Беларусі. Сер. фіз.-мат. навук. 2012. № 4. С. 37–41.
4. *Жук Е. Е.* // Весці НАН Беларусі. Сер. фіз.-мат. навук. 2013. № 3. С. 38–42.
5. *Жук Е. Е.* // Весці НАН Беларусі. Сер. фіз.-мат. навук. 2013. № 4. С. 101–106.
6. *Жук Е. Е., Храмова Е. В.* // Вестн. Белорус. ун-та. Сер. 1: Физика. Математика. Информатика. 2001. № 2. С. 80–86.
7. *Боровков А. А.* Математическая статистика. М., 1984.
8. *Боровков А. А.* Теория вероятностей. М., 1986.

E. E. ZHUK

STATISTICAL ASSIGNMENT OF MULTIVARIATE OBSERVATIONS TO THE CLASSES DETERMINED BY TRAINING SAMPLES

Summary

The problem of statistical assignment of arbitrarily distributed multivariate observations to the classes determined by training samples is considered. The decision rule based on the histogrammic estimators of probability densities is proposed and its efficiency is analytically investigated.