УДК 519.2

Е. Е. ЖУК

СТАТИСТИЧЕСКОЕ СООТНЕСЕНИЕ СЛУЧАЙНЫХ ВЫБОРОК С ЗАДАННЫМ ВЕРОЯТНОСТНЫМ РАСПРЕДЕЛЕНИЕМ МЕТОДОМ МАКСИМУМА ПРАВДОПОДОБИЯ

Белорусский государственный университет

(Поступила в редакцию 20.02.2014)

1. Математическая модель и постановка задачи. Пусть в пространстве наблюдений R^N ($N \geq 1$) зарегистрировано $m \geq 2$ случайных выборок $X^{(1)},...,X^{(m)}$, удовлетворяющих следующим двум условиям.

У1. Выборка $X^{(i)} = \{x_t^{(i)}\}_{t=1}^{n_i}$ состоит из независимых в совокупности одинаково распределенных N-векторов-наблюдений $x_t^{(i)} \in R^N$, $t = \overline{1, n_i}$, где n_i – ее объем, а сами наблюдения имеют плотность

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

У2. Выборки $\{X^{(i)}\}_{i=1}^m$ независимы между собой в совокупности.

Плотности $\{p_i(\cdot)\}_{i=1}^m$ из (1), определяющие выборки $\{X^{(i)}\}_{i=1}^m$, неизвестны. Задана некоторая фиксированная (гипотетическая [1, 2]) плотность распределения вероятностей

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

и среди выборок $\{X^{(i)}\}_{i=1}^m$ необходимо выделить ту, которая по своему распределению вероятностей «ближе» к распределению с гипотетической плотностью (2) (соотнести с гипотетическим распределением (2) ту выборку из $\{X^{(i)}\}_{i=1}^m$, которая с ним «лучше всего согласуется»).

Отметим, что поставленная выше задача отличается от задачи проверки так называемых гипотез согласия с заданным вероятностным распределением [1, 2], поскольку у соотносимых с гипотетическим распределением (2) выборок распределения вероятностей (1), вообще говоря, заведомо не совпадают с (2). Отличается данная задача и от задачи статистической классификации [3, 4]. Выше зафиксирован (задан) только один класс с плотностью $p(\cdot)$ из (2), и к нему необходимо отнести одну из выборок $\{X^{(i)}\}_{i=1}^m$, построив решающее правило (РП) вида

$$d = d(X^{(1)},...,X^{(m)}) \in M,\ M = \{1,\ldots,m\}, \tag{3}$$

где $d \in M$ – номер соотнесенной с гипотетическим распределением (2) выборки.

2. Решающее правило по принципу максимума правдоподобия и его риск. Для построения РП (3) по аналогии с [4] воспользуемся принципом максимума правдоподобия [1–4]:

$$d = d(X^{(1)}, ..., X^{(m)}) = \arg\max_{i \in M} P(X^{(i)}); \tag{4}$$

$$P(X^{(i)}) = \prod_{t=1}^{n_i} p(x_t^{(i)}),\ i \in M,$$

где $P(X^{(i)})$ – гипотетическая плотность выборки $X^{(i)}$, записанная в предположении, что $p_i(\cdot) \equiv p(\cdot)$ ($P(X^{(i)})$ – гипотетическая функция правдоподобия [1, 2], вычисленная для $X^{(i)}$).

Содержательный смысл РП (4) состоит в следующем: оно соотносит с гипотетическим распределением (2) ту выборку $X^{(d)}$ ($d \in M$) из предложенных выборок $\{X^{(i)}\}_{i\in M}$, появление которой в условиях заданного гипотетического распределения (2) «наиболее правдоподобно» («наиболее вероятно»).

Т е о р е м а 1. *Пусть в условиях* У1, У2 *конечны интегралы:*

$$\int_{R^N} \left|\ln\left(p(x)\right)\right| p_i(x)dx < +\infty,\ i \in M, \tag{5}$$

где $\{p_i(\cdot)\}_{i\in M}$, $p(\cdot)$ *– плотности из* (1), (2), *а среди величин*

$$H_i = H(p_i(\cdot), p(\cdot)) = \int_{R^N} \ln\left(p(x)\right) p_i(x)dx,\ i \in M, \tag{6}$$

лишь одна по значению больше остальных ($\exists d^o \in M:\ H_{d^o} > H_i,\ \forall i \neq d^o,\ i \in M$). *Тогда при совпадении объемов выборок:*

$$n_i = n,\ i \in M, \tag{7}$$

РП (4) *сходится в смысле почти наверное:*

$$d = d(X^{(1)}, ..., X^{(m)}) \xrightarrow{\text{п.н.}} d^o,\ n \to +\infty; \tag{8}$$

$$d^o = \arg\max_{i\in M} H_i.$$

Д о к а з а т е л ь с т в о. При $n_i = n$, $i \in M$, РП (4) можно записать в эквивалентном виде:

$$d = d(X^{(1)},...,X^{(m)}) = \arg\max_{i\in M} l(X^{(i)}); \tag{9}$$

$$l(X^{(i)}) = \frac{1}{n}\ln\left(P(X^{(i)})\right) = \frac{1}{n}\sum_{t=1}^{n}\ln\left(p(x_t^{(i)})\right),\ i \in M.$$

Далее с учетом условия (5) и того факта, что $\{x_t^{(i)}\}_{t=1}^n$ независимы в совокупности и одинаково распределены с плотностью $p_i(\cdot)$, применяем к $l(X^{(i)})$ усиленный закон больших чисел Колмогорова [1] ($i \in M$):

$$l(X^{(i)}) \xrightarrow{\text{п.н.}} H_i,\ n \to +\infty; \tag{10}$$

$$H_i = \mathrm{E}\left\{\ln\left(p(x_t^{(i)})\right)\right\} = \int_{R^N} \ln\left(p(x)\right) p_i(x)dx.$$

Из соотношений (9), (10) и известных теорем непрерывности [2] и заключаем справедливость (8).

Отметим, что функционалы $\{H(p_i(\cdot), p(\cdot))\}_{i\in M}$ из (6) традиционно связаны с классическим методом максимума правдоподобия [1, 2] и для большинства встречающихся на практике семейств Π допустимых плотностей ($p_i(\cdot) \in \Pi$, $i \in M$; $p(\cdot) \in \Pi$) обладают так называемым свойством идентифицируемости [1, 2]:

$$\forall\, p(\cdot) \in \Pi,\ \forall\, p_i(\cdot) \in \Pi:\ H(p_i(\cdot), p_i(\cdot)) > H(p_i(\cdot), p(\cdot)),\ p(\cdot) \not\equiv p_i(\cdot),\ i \in M. \tag{11}$$

Свойство (11) проясняет смысл величины $d^o \in M$ в (8). Величины $\{H_i\}_{i\in M}$ из (6) определяют степень «схожести» («близости») распределений (1) выборок $\{X^{(i)}\}_{i\in M}$ с гипотетическим распределением с плотностью $p(\cdot)$ из (2), а величина d^o – номер выборки, наиболее близкой по распределению к гипотетическому распределению (в смысле величин (6)).

Приведенные выше аналитические результаты позволяют, подобно [4], предложить в качестве меры эффективности РП (4) так называемый риск:

$$r = r\left(d\left(X^{(1)}, ..., X^{(m)}\right)\right) = \mathrm{P}\left\{d\left(X^{(1)}, ..., X^{(m)}\right) \notin D^o\right\}; \tag{12}$$

$$D^o = \left\{ k : H_k = \max_{j \in M} H_j \right\},$$

где учтено, что среди $\{H_i\}_{i \in M}$ из (6), в отличие от условий теоремы 1, на практике могут быть совпадающие по значению, а само РП (4) используется в общем случае (условие (7) о совпадении объемов $\{n_i\}_{i \in M}$ выборок $\{X^{(i)}\}_{i \in M}$, вообще говоря, не предполагается).

Содержательный смысл риска (12) очевиден: r – это вероятность не соотнести при помощи РП (4) с гипотетическим распределением (2) те выборки из $\{X^{(i)}\}_{i \in M}$, которые к нему близки по распределению в смысле величин (6). Чем меньше значение r из (12) ($0 \le r \le 1$), тем эффективнее РП (4) решает задачу соотнесения с гипотетическим распределением (2) выборок с плотностями (1).

Отметим, что если среди $\{H_i\}_{i \in M}$ из (6) нет совпадающих по значению (множество D^o в (12) содержит один элемент: $|D^o| = 1$), то риск (12) упрощается:

$$r = r\left(d\left(X^{(1)}, ..., X^{(m)}\right)\right) = \mathrm{P}\left\{d\left(X^{(1)}, ..., X^{(m)}\right) \ne d^o\right\}; \tag{13}$$

$$d^o = \arg\max_{i \in M} H_i .$$

Если же все $\{H_i\}_{i \in M}$ равны между собой ($D^o = \{1, \ldots, m\} = M$, $|D^o| = m$), то очевидно, что $r = 0$.

3. Асимптотическое вычисление риска в случае двух выборок одинакового объема. Модель Фишера. Пусть с гипотетическим распределением (2) необходимо соотнести две ($m = 2$) выборки $X^{(1)} = \{x_t^{(1)}\}_{t=1}^n$, $X^{(2)} = \{x_t^{(2)}\}_{t=1}^n$ одинакового объема ($n_1 = n_2 = n$). РП (4) можно записать в виде ($M = \{1,2\}$):

$$d(X^{(1)}, X^{(2)}) = \begin{cases} 1, & \text{если } \ \bar{\xi}_n(X^{(1)}, X^{(2)}) \le 0; \\ 2, & \text{если } \ \bar{\xi}_n(X^{(1)}, X^{(2)}) > 0, \end{cases} \tag{14}$$

где

$$\bar{\xi}_n(X^{(1)}, X^{(2)}) = \frac{1}{n} \sum_{t=1}^{n} \ln\left(\frac{p(x_t^{(2)})}{p(x_t^{(1)})} \right), \tag{15}$$

а $p(\cdot)$ – гипотетическая плотность из (2).

Риск r из (12), (13) для РП (14), (15) примет вид:

$$r = \begin{cases} P\{\bar{\xi}_n(X^{(1)}, X^{(2)}) \le 0\}, & \text{если } H_1 < H_2; \\ 1 - P\{\bar{\xi}_n(X^{(1)}, X^{(2)}) \le 0\}, & \text{если } H_1 > H_2; \\ 0, & \text{если } H_1 = H_2; \end{cases} \tag{16}$$

где H_1, H_2 – величины из (6).

Т е о р е м а 2. *Пусть в случае двух выборок* ($m = 2$) *одинакового объема* ($n_1 = n_2 = n$) *выполнены условия* У1, У2, *а плотности* $p_1(\cdot)$, $p_2(\cdot)$ *из* (1) *и плотность* $p(\cdot)$ *из* (2) *таковы, что*

$$G_i = \int_{R^N} \left(\ln\left(p(x)\right)\right)^2 p_i(x) dx < +\infty, \ G_i - H_i^2 \ne 0, \ i = 1,2, \tag{17}$$

тогда риск (16) *может быть вычислен из асимптотического соотношения* ($H_1 \ne H_2$):

$$\frac{r}{\widetilde{r}} \to 1, \ n \to +\infty; \ \widetilde{r} = \Phi\left(-\sqrt{n} \frac{|H_1 - H_2|}{\sqrt{G_1 + G_2 - (H_1^2 + H_2^2)}} \right), \tag{18}$$

где $\Phi(z) = \dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\dfrac{w^2}{2} \right) dw$, $z \in R$ *– функция распределения вероятностей стандартного нормального закона* $N_1(0,1)$.

Д о к а з а т е л ь с т в о. В условиях *У*1, *У*2 случайные величины

$$\xi_t = \ln\left(\frac{p(x_t^{(2)})}{p(x_t^{(1)})}\right),\ t=\overline{1,n},$$

независимы в совокупности и одинаково распределены с математическим ожиданием:

$$\mathrm{E}\{\xi_t\} = \mathrm{E}\left\{\ln\left(\frac{p(x_t^{(2)})}{p(x_t^{(1)})}\right)\right\} = \mathrm{E}\left\{\ln\left(p(x_t^{(2)})\right)\right\} - \mathrm{E}\left\{\ln\left(p(x_t^{(1)})\right)\right\} = H_2 - H_1$$

и конечной ненулевой дисперсией:

$$\mathrm{D}\{\xi_t\} = \mathrm{D}\left\{\ln\left(p(x_t^{(2)})\right)\right\} + \mathrm{D}\left\{\ln\left(p(x_t^{(1)})\right)\right\} = G_2 - H_2^2 + G_1 - H_1^2 = G_1 + G_2 - (H_1^2 + H_2^2) < +\infty,$$

где учтено условие (17).

Согласно центральной предельной теореме Леви – Линдеберга [1], асимптотически нормально распределена случайная величина $\overline{\xi}_n(X^{(1)}, X^{(2)}) = \frac{1}{n}\sum_{t=1}^{n}\xi_t$, для которой с учетом нормировки имеем:

$$\sqrt{n}\,\frac{\overline{\xi}_n(X^{(1)}, X^{(2)}) - \mathrm{E}\{\xi_t\}}{\sqrt{\mathrm{D}\{\xi_t\}}} = \frac{\overline{\xi}_n(X^{(1)}, X^{(2)}) - (H_2 - H_1)}{\sqrt{G_1 + G_2 - (H_1^2 + H_2^2)}} \to N_1(0,1),\ n \to +\infty. \qquad (19)$$

Из (16) с учетом (19) и известного свойства функции распределения $\Phi(\cdot)$ стандартного нормального закона: $\Phi(-z) = 1 - \Phi(z)$, $z \in R$, получаем соотношение (18).

Из результата (18) теоремы 2 видно, что если $H_1 \neq H_2$,то эффективность соотнесения с гипотетическим распределением выборок $X^{(1)}$, $X^{(2)}$ при помощи РП (14), (15) повышается (риск уменьшается) с увеличением различия значений H_1 и H_2 между собой, а также с ростом объема n выборок $X^{(1)}$, $X^{(2)}$. Если же $H_1 = H_2$, то $r = 0$, и неважно, какая из выборок $X^{(1)}$ или $X^{(2)}$ будет соотнесена с гипотетическим распределением (2).

Пусть теперь плотности $p_1(\cdot)$, $p_2(\cdot)$ и $p(\cdot)$ многомерные нормальные (гауссовские):

$$p_i(x) = n_N(x \mid \mu_i, \Sigma),\ i = 1,2;\quad p(x) = n_N(x \mid \mu, \Sigma),\ x \in R^N, \qquad (20)$$

где

$$n_N(x \mid \overline{\mu}, \Sigma) = (2\pi)^{-\frac{N}{2}} \mid \Sigma \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\overline{\mu})^{\mathrm{T}}\Sigma^{-1}(x-\overline{\mu})\right)$$

– плотность N-мерного нормального закона распределения вероятностей («т»– символ транспонирования); $\mu_1 = \mathrm{E}\{x_t^{(1)}\}$, $\mu_2 = \mathrm{E}\{x_t^{(2)}\}$ – N-вектора математических ожиданий наблюдений из выборок $X^{(1)}$ и $X^{(2)}$ соответственно, $\mu \in R^N$ – математическое ожидание гипотетического распределения, а Σ – общая для всех распределений невырожденная ковариационная $(N \times N)$-матрица ($|\Sigma| \neq 0$).

Отметим, что плотностями (20) часто на практике адекватно описываются реальные статистические данные, и такая модель известна как модель Фишера [3, 4].

С учетом (20) для асимптотического значения риска $\tilde{r}$ из (18) получаем ($n \to +\infty$):

$$\tilde{r} = \Phi\left(-\sqrt{n}\,\frac{\left|\rho^2(\mu,\mu_1) - \rho^2(\mu,\mu_2)\right|}{2\sqrt{N + \rho^2(\mu,\mu_1) + \rho^2(\mu,\mu_2)}}\right), \qquad (21)$$

где

$$\rho(\mu,\mu_i) = \sqrt{(\mu-\mu_i)^{\mathrm{T}}\Sigma^{-1}(\mu-\mu_i)}$$

– расстояние Махаланобиса [1, 3, 4] между μ и μ_i ($i = 1, 2$). Использованы выражения для моментов случайных величин, имеющих нецентральное χ^2-распределение, приведенные в [5], которые позволили получить следующие выражения для величин $\{H_i, G_i - H_i^2\}_{i=1}^2$ из (6) и (17):

$$H_i = \mathrm{E}\left\{\ln\left(n_N(x_t^{(i)} \mid \mu, \Sigma)\right)\right\} = \ln\left((2\pi)^{-\frac{N}{2}} \mid \Sigma \mid^{-\frac{1}{2}}\right) - \frac{1}{2}\mathrm{E}\left\{(x_t^{(i)} - \mu)^{\mathrm{T}} \Sigma^{-1} (x_t^{(i)} - \mu)\right\} =$$

$$= \ln\left((2\pi)^{-\frac{N}{2}} \mid \Sigma \mid^{-\frac{1}{2}}\right) - \frac{1}{2}\left(N + \rho^2(\mu, \mu_i)\right);$$

$$G_i - H_i^2 = \mathrm{D}\left\{\ln\left(n_N(x_t^{(i)} \mid \mu, \Sigma)\right)\right\} = \mathrm{D}\left\{\ln\left((2\pi)^{-\frac{N}{2}} \mid \Sigma \mid^{-\frac{1}{2}}\right) - \frac{1}{2}(x_t^{(i)} - \mu)^{\mathrm{T}} \Sigma^{-1} (x_t^{(i)} - \mu)\right\} =$$

$$= \frac{1}{4}\mathrm{D}\left\{(x_t^{(i)} - \mu)^{\mathrm{T}} \Sigma^{-1} (x_t^{(i)} - \mu)\right\} = \frac{1}{4}\left(2\left(N + 2\rho^2(\mu, \mu_i)\right)\right) = \frac{1}{2}\left(N + 2\rho^2(\mu, \mu_i)\right),$$

где учтено, что случайная величина $(x_t^{(i)} - \mu)^{\mathrm{T}} \Sigma^{-1} (x_t^{(i)} - \mu)$ имеет нецентральное χ^2-распределение с N степенями свободы и параметром нецентральности $\rho^2(\mu, \mu_i)$.

Формула (21) справедлива для $\rho(\mu, \mu_1) \neq \rho(\mu, \mu_2)$ (при этом $H_1 \neq H_2$), и, согласно (21), эффективность соотнесения увеличивается с ростом различия расстояний Махаланобиса $\rho(\mu, \mu_1)$ и $\rho(\mu, \mu_2)$ от гипотетического математического ожидания μ до математических ожиданий μ_1 и μ_2 наблюдений из выборок $X^{(1)}$ и $X^{(2)}$ соответственно. Если эти расстояния равны: $\rho(\mu, \mu_1) = \rho(\mu, \mu_2)$, то риск $r = 0$.

Литература

1. *Харин Ю. С., Зуев Н. М., Жук Е. Е.* Теория вероятностей, математическая и прикладная статистика. Минск, 2011.
2. *Боровков А. А.* Математическая статистика. М., 1984.
3. *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Классификация и снижение размерности. М., 1989.
4. *Жук Е. Е.* // Весці НАН Беларусі. Сер. фіз.-мат. навук. 2013. № 4. С. 101–106.
5. *Хацкевич Г. А.* // Заводская лаборатория. 1994. № 10. С. 49–55.

E. E. ZHUK

STATISTICAL ASSIGNMENT OF RANDOM SAMPLES WITH THE FIXED PROBABILITY DISTRIBUTION BY THE MAXIMUM LIKELIHOOD METHOD

Summary

The problem of statistical assignment of arbitrarily distributed multivariate samples with the fixed probability distribution is considered. The decision rule based on the maximum likelihood method is proposed and its efficiency is analytically investigated. The case of two samples and the Fisher model is investigated.