СООТНЕСЕНИЕ ОДНОРОДНЫХ МНОГОМЕРНЫХ ВЫБОРОК С ФИКСИРОВАННЫМ ВЕРОЯТНОСТНЫМ РАСПРЕДЕЛЕНИЕМ

Е.Е. Жук, Д.Д. Дусь

Белорусский государственный университет, Минск, Беларусь
zhukee@mail.ru, dzianisdus@gmail.com

Пусть в пространстве наблюдений $\mathbb{R}^N$ $(N \geqslant 1)$ зарегистрированы $m \geqslant 1$ случайных выборок $X^{(1)}, \ldots, X^{(m)}$, для которых верны следующие условия:

1) каждая выборка $X^{(i)} = \{x_t^{(i)}\}_{t=1}^{n_i}$ состоит из независимых в совокупности одинаково распределенных векторов-наблюдений $x_t^{(i)} \in \mathbb{R}^N,\ t = \overline{1, n_i}$ (где n_i — объём $X^{(i)}$) и порождена собственным законом распределения вероятностей, имеющим плотность

$$p_i(x) \geqslant 0, \quad x \in \mathbb{R}^N : \int\limits_{\mathbb{R}^N} p_i(x)\,dx = 1, \quad i = \overline{1, m}, \tag{1}$$

вообще говоря, неизвестную;

2) выборки $X^{(1)}, \ldots, X^{(m)}$ независимы между собой в совокупности.

Пусть также задана некоторая фиксированная (гипотетическая [1, 2]) плотность распределения вероятностей

$$p(x) \geqslant 0, \quad x \in \mathbb{R}^N : \int\limits_{\mathbb{R}^N} p(x)\,dx = 1. \tag{2}$$

Задача состоит в том, чтобы из выборок $\{X^{(i)}\}_{i=1}^m$ выбрать ту, которая «ближе» из всех представленных к гипотетической плотности (2) в терминах сходства вероятностных распределений.

Формально задача сводится к построению решающего правила (РП)

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

соотносящего одну из выборок $\{X^{(i)}\}_{i=1}^m$ с фиксированной гипотетической плотностью (2).

Для построения РП (3) воспользуемся принципом максимального правдоподобия [1–4]:

$$d = d(X^{(1)}, \ldots, X^{(m)}) = \arg\max_{i \in M} P(X^{(i)}); \quad P(X^{(i)}) = \prod_{t=1}^{n_i} p(x_t^{(i)}), \quad i \in M, \tag{4}$$

где $P(X^{(i)})$ — гипотетическая функция правдоподобия [1, 2], вычисленная для выборки $X^{(i)}$.

Теорема. *Пусть конечны следующие интегралы:*

$$\int\limits_{\mathbb{R}^N} |\ln(p(x))| p_i(x)\,dx < +\infty, \quad i \in M.$$

Если для значений

$$H_i = H(p_i(\cdot), p(\cdot)) = \int\limits_{\mathbb{R}^N} \ln(p(x)) p_i(x)\,dx, \quad i \in M, \tag{5}$$

выполняется условие: $\exists d^0 \in M :\ H_{d^0} > H_i,\ \forall i \neq d^0,\ i \in M$, *и все представленные выборки* $\{X^{(i)}\}_{i=1}^m$ *имеют одинаковый объём:* $n_i = n,\ i \in M$, *тогда для РП* (4) *справедливо*:

$$d = d(X^{(1)}, \ldots, X^{(m)}) \xrightarrow{\text{п.н.}} d^0, \quad n \to +\infty; \quad d^0 = \arg\max_{i \in M} H_i. \tag{6}$$

Приведенная выше теорема (соотношение (6)) позволяет аналогично [4] определить обобщение традиционного риска [1–3] для РП (4):

$$r = r(d(X^{(1)}, \ldots, X^{(m)})) = P\{d(X^{(1)}, \ldots, X^{(m)}) \neq d^0\}. \tag{7}$$

Следствие. *Пусть РП* (4) *используется для соотнесения двух выборок* $(m = 2)$ *одинакового объёма* $(n_1 = n_2 = n)$ *и пусть выполняются условия*:

$$G_i = \int\limits_{\mathbb{R}^N} (\ln(p(x)))^2 p_i(x)\, dx < +\infty, \quad G_i - H_i^2 \neq 0, \quad i = 1, 2,$$

где H_i, $i = 1, 2$ *— величины из* (5). *Тогда риск* (7) *РП* (4) *может быть вычислен асимптотически* (*при* $H_1 \neq H_2$) :

$$\frac{r}{\tilde{r}} \to 1, \quad n \to +\infty; \quad \tilde{r} = \Phi\Bigg(-\sqrt{n} \frac{|H_1 - H_2|}{\sqrt{G_1 + G_2 - (H_1^2 + H_2^2)}} \Bigg),$$

где $\Phi(z) = (1/\sqrt{2\pi}) \int_{-\infty}^{z} \exp(-w^2/2)\, dw$, $z \in \mathbb{R}$ *— функция распределения стандартного гауссовского закона.*

Полученные результаты проиллюстрированы также для важного практического случая, когда плотности (1), (2) — многомерные нормальные (гауссовские) [1–3].

Литература

1. Харин Ю. С., Зуев Н. М., Жук Е. Е. *Теория вероятностей, математическая и прикладная статистика.* Мн.: БГУ, 2011.

2. Боровков А. А. *Математическая статистика.* М.: Наука, 1984.

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

4. Жук Е. Е. *Метод максимального правдоподобия для отнесения многомерных однородных выборок к классам и его риск* // Весці НАН Беларусі. Сер. фіз.-мат. навук. 2013. № 3. С. 38–42.