НЕПАРАМЕТРИЧЕСКАЯ КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ НАБЛЮДЕНИЙ НА ОСНОВЕ ЯДЕРНЫХ ОЦЕНОК ПЛОТНОСТЕЙ С ПРЯМОУГОЛЬНЫМ ЯДРОМ

Паланевич А. С., Жук Е. Е.

Белорусский государственный университет, Минск, Беларусь,
e-mail: alexander.palanevich@gmail.com

Пусть регистрируются случайные наблюдения $x = x(w) \in R^N$ над объектами $w \in \Omega$, принадлежащими к L классам $\{\Omega_1, \Omega_2, \dots, \Omega_L\}$:

$$\Omega_i \cap \Omega_j = \emptyset, i \neq j, i, j = 1,2, \dots, L;$$

$$\bigcup_{i=1}^{L} \Omega_i = \Omega.$$

Обозначим истинный номер класса, к которому принадлежит объект w, *через* $d^0(w)$. Этот номер является дискретной случайной величиной со следующим распределением:

$$P(d^0(w) = i) = \pi_i, i = 1,2, \dots, L;$$

$$\pi_1 + \pi_2 + \dots + \pi_L = 1.$$

Здесь $\{\pi_i, i = \overline{1, L}\}$ – априорные вероятности классов. В рамках каждого из классов наблюдение $x(w)$ описывается условной плотностью распределения:

$$p_i(x) = p(x|d^0(x) = i), i = 1,2, \dots, L, x \in R^N.$$

Поставим перед собой задачу оценки номера класса для нового наблюдения по имеющейся классифицированной выборке (задача дискриминантного анализа). Для этого будем пользоваться байесовским решающим правилом (БРП) [1]:

$$\widehat{d^0}(x) = \underset{i=1,2,\dots,L}{argmax}(\pi_i p_i(x)), x \in R^N,$$

где $\widehat{d^0}(x)$ – оценка номера неизвестного класса для наблюдения x. Мы имеем дело с нерандомизированным решающим правилом [1]. Однако для пользования этим правилом необходимо знать $\{\pi_i, p_i(x), i = 1,2, \dots, L\}$. Так как их точные значения неизвестны, то укажем способ построения оценок для этих величин.

Пусть $X = \{x_1, x_2, \dots, x_n\} \in R^{nN}$ – классифицированная выборка и n_k – число наблюдений из выборки, которые относятся к классу с номером k. Тогда построим оценку для априорных вероятностей [1]:

$$\widehat{\pi_k} = \frac{n_k}{n}, k = 1,2, \dots, L. \quad (1)$$

Условные плотности распределения будем оценивать с помощью ядерных оценок с прямоугольным ядром. Пусть $\Gamma(x)$ – N-мерный параллелепипед с центром в точке x и сторонами $h_i \in R, i = 1,2, \dots, N$, с "объемом" $V = \prod_{i=1}^{N} h_i$. Вводя функцию-индикатор $I_{\Gamma(x)}(y)$, равную единице, если $y \in \Gamma(x)$, и нулю - в противном случае, оценки плотностей запишем следующим образом [2]:

$$\widehat{p_k}(y) = \frac{1}{n_k V} \sum_{j=1}^{n} I_{\Gamma(y)}(x_j) \delta_{k,d^0(x_j)}, x_j \in X, j = 1,2, \dots, n, k = 1,2, \dots, L, y \in R^N. \quad (2)$$

Тогда для оценки номера класса, к которому принадлежит новое наблюдение x^*, получаем следующее подстановочное БРП [1]:

$$\hat{d}(x^*) = \underset{k=1,2,\dots,L}{argmax}\left(\widehat{\pi_k}\widehat{p_k}(x^*)\right), x^* \in R^N. \quad (3)$$

Теперь сформулируем и докажем теорему, придающую построенному БРП содержательный смысл.

Теорема. Пусть по обучающей выборке $X = \{x_1, x_2, \dots, x_n\} \in R^{nN}$ построены упомянутые выше оценки параметров модели $\{\widehat{\pi_k}, \widehat{p_k}(x), k = \overline{1, L}\}$ из (1), (2). Тогда новое наблюдение $x^* \in R^N$, классифицируемое с помощью подстановочного БРП (3), относится к классу Ω_d с номером d, когда количество наблюдений из выборки X, лежащих в $\Gamma(x^*)$ и принадлежащих этому классу, является наибольшим среди остальных классов.

Доказательство. Для простоты рассуждений предположим, что максимум в БРП достигается на одном классе с номером d. Тогда:

$$\hat{d}(x^*) = d \Leftrightarrow \widehat{\pi_d}\widehat{p_d}(x^*) > \widehat{\pi_k}\widehat{p_k}(x^*), k = 1,2, \dots, L, k \neq d.$$

Подставляя полученные ранее оценки параметров модели (1), (2), получаем:

$$\frac{n_d}{n}\frac{1}{n_d V}\sum_{j=1}^{n} I_{\Gamma(x^*)}(x_j)\delta_{d,d^0(x_j)} > \frac{n_k}{n}\frac{1}{n_k V}\sum_{j=1}^{n} I_{\Gamma(x^*)}(x_j)\delta_{k,d^0(x_j)};$$

$$\sum_{j=1}^{n} I_{\Gamma(x^*)}(x_j)\left(\delta_{d,d^0(x_j)} - \delta_{k,d^0(x_j)}\right) > 0.$$

Что и соответствует сформулированной выше интерпретации БРП.

Литература

1. Жук Е.Е., Харин Ю.С. Математическая и прикладная статистика//учебное пособие. –Минск, БГУ, 2005. – С. 192–208.

2. Multivariate Kernel Smoothing and its Applications / José E. Chacón, Tarn Duong –Taylor & Francis Group, LLC, 2018