Testing of hypothesis of random variables independence on the basis of nonparametric algorithm of pattern recognition

Alexandr V. Lapko
Siberian Branch of the Russian Academy of Sciences
Institute of Computer Modelling
Krasnoyarsk, Russia
e-mail: lapko@icm.krasn.ru

Vasily A. Lapko
Institute of Computer Modelling RAS
Reshetnev Siberian State Aerospace University
Krasnoyarsk, Russia
e-mail: lapko@icm.krasn.ru

Ekaterina A. Yuronen
Reshetnev Siberian State Aerospace University
Siberian Federal University
Krasnoyarsk, Russia
e-mail: lapko@icm.krasn.ru

***Abstract*—The new technique of testing of hypothesis of random variables independence is offered. Its basis is made by nonparametric algorithm of pattern recognition. The considered technique doesn't demand sampling of area of values of random variables.**

Keywords—testing of hypothesis, pattern recognition, independent random variables, Parzen–Rosenblatt estimate.

I. INTRODUCTION

The testing of hypothesis about random distributions with use of nonparametric algorithms of a pattern recognition is considered in works [1, 2]. Possibility of transition from a problem of comparison of distribution laws of random values to check of a hypothesis of equality of probability of an error of pattern recognition to threshold value is proved. The offered approach, for example, when checking a hypothesis of uniformity of distribution laws of two sequences of random values consists in realization of the following actions:

- according to the compared random series to create the training selection for the solution of the two-alternate problem of a pattern recognition;

- to carry out synthesis of nonparametric algorithm of the pattern recognition corresponding to criterion of maximum likelihood (for estimation of probability densities of random values in classes are used statistics like Rosenblatt-Parzen);

- n the mode of "the sliding examination" to calculate an assessment of probability of an error of pattern recognition;

- to check a hypothesis of equality of probability of an error of pattern recognition to threshold value.

This approach allows bypassing a problem of decomposition of a range of values of random values which is peculiar to Pearson's criterion.

In this work nonparametric algorithms of a pattern recognition are used at the solution of a problem of testing of a hypothesis of independence of random values.

II. PROPERTIES OF NONPARAMETRIC ESTIMATES OF THE PROBABILITY DENSITY DEPENDENT AND INDEPENDENT CASUAL VELICH

Let's compare asymptotic properties of nonparametric estimates of a probability density in the conditions of dependence and independence of random values. Statistical data are used at synthesis of nonparametric algorithms of pattern recognition at a test of hypothesis about random distributions.

Let there is a selection $V = \left(x^i ,\, i = \overline{1, n}\right)$ of n statistically independent supervision of a two-dimensional random value $x = (x_1 , x_2)$ with a priori unknown probability density $p(x)$. It is known, as x_1 and x_2 are independent.

In these conditions for estimation of a probability density $p(x)$ let's use nonparametric statistics

$$\bar{p}(x) = \bar{p}_1(x_1)\, \bar{p}_2(x_2), \tag{1}$$

where

$$\bar{p}_v(x_v) = \frac{1}{n c_v} \sum_{i=1}^{n} \Phi\left(\frac{x_v - x_v^i}{c_v}\right), \; v = 1, 2\,. \qquad (2)$$

Nuclear functions $\Phi(u_v)$ obey H :

$$\Phi(u_v) = \Phi(-u_v), \quad 0 \le \Phi(u_v) < \infty\,,$$

$$\int \Phi(u_v) du_v = 1, \quad \int u_v^2\, \Phi(u_v) du_v = 1\,,$$

$$\int u_v^m\, \Phi(u_v) du_v < \infty, \quad 0 \le m < \infty, \; v = 1, 2\,.$$

Parameters of nuclear functions $c_v = c_v(n), v = 1, 2$ decrease with body height n. Hereinafter the infinite limits are passed.

Fairly following statement:

Theorem. Let density of probabilities $p_v(x_v)$ of random values x_v, $v = 1, 2$ and their first derivative be limited and continuous; nuclear functions $\Phi(u_v)$ obey H; sequences $c_1 = c_1(n)$, $c_2 = c_2(n)$ of coefficients of a diffuseness of nuclear functions of a nonparametric assessment of a probability density $\bar{p}(x_1, x_2)$ are, that $n \to \infty$ values $c_1 \to 0$, $c_2 \to 0$, and $n c_1 \to \infty$ and $n c_2 \to \infty$.

Then for a nonparametric assessment $\bar{p}(x_1, x_2) = \bar{p}_1(x_1)\bar{p}_2(x_2)$ of a probability density $p(x_1, x_2) = p_1(x_1) p_2(x_2)$ asymptotic expression of a mean squared deviation will register in a look

$$M \iint \left(p_1(x_1) p_2(x_2) - \tilde{p}_1(x_1)\tilde{p}_2(x_2)\right)^2 dx_1\, dx_2 \sim$$

$$\sim \frac{\left(\|\Phi(u)\|^2\right)^2}{n^2 c_1 c_2} + \frac{\|\Phi(u)\|^2 \|p_2(x_2)\|^2}{n c_1} + \frac{\|\Phi(u)\|^2 \|p_1(x_1)\|^2}{n c_2} +$$

$$+ \int\int \left(\frac{p_2(x_2)\, p_1^{(2)}(x_1)\, c_1^2}{2} + \frac{p_1(x_1)\, p_2^{(2)}(x_2)\, c_2^2}{2} \right) dx_1\, dx_2\,. \qquad (3)$$

Here the following designations are used:

$$\|\Phi(u)\|^2 = \int \Phi^2(u)\, du\,; \; \|p_v(x_v)\|^2 = \int p_v^2(x_v)\, du\,;$$

$p_v^{(2)}(x_v)$ - density flexon $p_v(x_v)$, $v = 1, 2$; M - is the sign of expected value.

Convergence in mean squared statistics follows from the analysis of expression (3) when performing conditions of the theorem (1).

At the proof of the theorem the technique offered in work [3] and developed in [4-8] is used.

Let's compare approximating properties of nonparametric estimates of a probability density (1) and

$$\tilde{p}(x) = \frac{1}{n c_1 c_2} \sum_{i=1}^{n} \prod_{v=1}^{2} \Phi\left(\frac{x_v - x_v^i}{c_v}\right). \qquad (4)$$

For this purpose we will define best values c_v from a condition of a minimum of asymptotic expression of a mean squared deviation

$$M \int \left(\bar{p}_v(x_v) - p_v(x_v)\right) dx_v \sim \frac{1}{n c_v} \|\Phi(u)\|^2 + \frac{c^4}{4} \left\|p_v^{(2)}(x_v)\right\|^2.$$

Then it is easy to show that best value of diffuseness coefficient of nuclear function is defined by expression

$$c_v^* = \left[\frac{\|\Phi(u)\|^2}{n \left\|p_v^{(2)}(x_v)\right\|^2} \right]^{\frac{1}{5}}, \; v = 1, 2\,.$$

Substituting c_v^*, $v = 1, 2$ in (3) we will receive

$$W_2 = \left(\frac{\|\Phi(u)\|^2}{n}\right)^{\frac{4}{5}} \left[\left(\frac{\|\Phi(u)\|^2}{n}\right)^{\frac{4}{5}} \left(\left\|p_1^{(2)}(x_1)\right\|^2 \left\|p_2^{(2)}(x_2)\right\|^2 \right)^{\frac{1}{5}} + \right.$$

$$+ \frac{5}{4} \left(\left(\left\|p_1^{(2)}(x_1)\right\|^2 \right)^{\frac{1}{5}} \|p_2(x_2)\|^2 + \left(\left\|p_2^{(2)}(x_2)\right\|^2 \right)^{\frac{1}{5}} \|p_1(x_1)\|^2 \right) +$$

$$\left. + \frac{1}{2} \left(\left\|p_1^{(2)}(x_1)\right\|^2 \left\|p_2^{(2)}(x_2)\right\|^{2^2} \right)^{-\frac{2}{5}} \prod_{v=1}^{2} \int p_v(x_v)\, p_v^{(2)}(x_v)\, dx_v \right].$$

According to results of researches [4] the minimum mean squared deviation of a nonparametric assessment of a probability density $\tilde{p}(x_1, x_2)$ at $k = 2$ and $c_1 = c_2$ is defined by expression

$$W_2' = \frac{5}{2^{7/3}} \left[\left(\frac{\left(\|\Phi(u)\|^2 \right)^2}{n} \right)^4 \times \right.$$

$$\left. \times \left(\int\int \left(p_1^{(2)}(x_1, x_2) + p_2^{(2)}(x_1, x_2) \right)^2 dx_1 \, dx_2 \right)^2 \right]^{\frac{1}{6}},$$

where $p_v^{(2)}(x_1, x_2)$ - probability density flexon $p(x_1, x_2)$ on a variable x_v, $v = 1, 2$.

At terminating $p_v(x_v)$, $p_v^{(2)}(x_v)$, $v = 1, 2$ with body height of volume n of statistical data of value W_2 of a minimum mean squared deviation $\bar{p}(x_1, x_2) = \bar{p}_1(x_1)\bar{p}_2(x_2)$ aspire to zero in proportion to $r = n^{-4/5}$. And the order of similar convergence is higher, than for W_2', which values decrease in proportion to $n^{-4/6}$.

Distinction of approximating properties of nonparametric estimates of a probability density $\bar{p}(x)$, $\tilde{p}(x)$ is a basis of a technique of check of a hypothesis of independence of random values with use of nonparametric algorithms of pattern recognition.

III. TECHNIQUE OF THE HYPOTHESIS TESTING OF INDEPENDENCE OF RANDOM VALUES

There is a selection $V = \left(x_1^i, x_2^i, i = \overline{1, n} \right)$ from n statistically independent supervision of a two-dimensional random value $x = (x_1, x_2)$. Random values are characterized by probability densities $p(x_1, x_2)$, $p_1(x_1)$, $p_2(x_2)$. It is necessary to confirm or disprove a hypothesis H_0 about independence of distribution laws of random values x_1, x_2.

Let's assume that there are two classes Ω_1, Ω_2. The first class Ω_1 is characterized by a probability density $p_1(x_1)p_2(x_2)$. It can be defined as a nonparametric assessment of a probability density $\bar{p}_1(x_1)\bar{p}_2(x_2)$ type (2) which is restored on selection V. The second class Ω_2 is defined by a probability density $p(x_1, x_2)$ and is estimated by statistics (4).

On this basis we will construct nonparametric algorithm of pattern recognition

$$\bar{m}(x): \begin{cases} x \in \Omega_1, \, if \quad \bar{f}_{12}(x) \geq 0 \\ x \in \Omega_2, \, if \quad \bar{f}_{12}(x) < 0, \end{cases} \tag{5}$$

where

$$\bar{f}_{12}(x) = \bar{p}_1(x_1)\bar{p}_2(x_2) - \bar{p}(x_1, x_2).$$

The choice of best values c_v^*, $v = 1, 2$ is carried out from a condition of a minimum of an assessment of probability of an error of pattern recognition

$$\bar{\rho} = \frac{1}{n} \sum_{j=1}^{n} 1(\sigma(j), \bar{\sigma}(j))$$

on the training selection $V' = \left(x^i, \sigma(i), i = \overline{1, n} \right)$. Here all $\sigma(i)$ - are instructions on situation accessory x^i to the first class Ω_1. Decisions $\bar{\sigma}(i)$ are defined by algorithm (5).

Indicator function

$$1(\sigma(j), \bar{\sigma}(j)) = \begin{cases} 0, \, if \quad \sigma(j) = \bar{\sigma}(j) \\ 1, \, if \quad \sigma(j) \neq \bar{\sigma}(j). \end{cases}$$

When forming "decision" $\bar{\sigma}(j)$ the situation x^j is excluded from process of calculation of statisticians $\bar{p}_1(x_1)\bar{p}_2(x_2)$, $\bar{p}(x_1, x_2)$.

To define a minimum error of pattern recognition $\bar{\rho}^*$, which correspond to values c_v^*, $v = 1, 2$.

Using traditional criteria, to check a hypothesis H_1 about equality of probability of an error of pattern recognition to value $1/2$. The initial hypothesis H_0 is fair if the hypothesis H_1 is carried out, differently the hypothesis H_0 is rejected.

The offered technique can is generalized on a problem of hypothesis testing of independence of sets of random values

$$x(1) = \left(x_v, v = \overline{1, k1} \right), \; x(2) = \left(x_v, v = \overline{k1+1, k} \right).$$

ACKNOWLEDGMENT

Difference of approximating properties of nonparametric estimates of probability densities of dependent and independent random values is established. On this basis possibility of application of nonparametric algorithm of pattern recognition in a problem of hypothesis testing of independence of random values is proved. The offered technique allows bypassing difficult formalizable procedure of decomposition of a range of values of random values.

Further development of the offered approach is bound to its generalization on a problem of hypothesis testing of independence of sets of random values.

This work was carried out as part of assignment No. 2.914.2014/K of the Ministry of Education and Science of the Russian Federation.

REFERENCES

[1] A.V. Lapko and V.A. Lapko, "Nonparametric algorithms of pattern recognition in the problem of testing a statistical hypothesis on identity of two distribution laws of random variables," Opt. Instrum. Data Proc., vol. 46, no. 6, pp.545-550, 2010.

[2] A.V. Lapko and V.A. Lapko, "Comparison of empirical and theoretical distribution functions of a random variable on the basis of a nonparametric classifier," Opt. Instrum. Data Proc., vol. 48, no. 1, pp.37-41, 2012.

[3] V.A. Epanechnikov, "Nonparametric estimator of multidimensional probability density," Teor. Veroyatn. Ee Primen., vol. 14, Iss. 1, pp. 156–161, 1969.

[4] A.V. Lapko and V.A. Lapko, "Properties of nonparametric estimates of multidimensional probability density of independent random variables," Informatika i systemy upravleniya, vol. 31, no. 1, pp.166-174, 2012.

[5] A.V. Lapko and V.A. Lapko, "Optimal selection of the number of sampling intervals in domain of variation of a one-dimensional random variable in estimation of the probability density," Measur. Techn., vol. 56, pp. 763 – 767, 2013.

[6] A.V. Lapko and V.A. Lapko, "Properties of the nonparametric decision function with a priori information on independence of attributes of classified objects," Opt. Instrum. Data Proc., vol. 48, no. 4, pp.416-422, 2012.

[7] A.V. Lapko and V.A. Lapko, "Effect of data incompleteness on the approximation properties of nonparametric estimation of the two-dimensional probability density of independent random variables," Opt. Instrum. Data Proc., vol. 50, no. 1, pp.68-74, 2014.

[8] A.V. Lapko and V.A. Lapko, "Regression estimate of the multidimensional probability density and its properties," Opt. Instrum. Data Proc., vol. 50, no. 2, pp.148-153, 2014.