Непараметрическая методика проверки гипотезы о независимости случайных величин и ее применение при анализе данных дистанционного зондирования

А. В. Шаруева

Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева
Российская Федерация, 660037, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: anna-denisyuk@yandex.ru

Аннотация. Проверка гипотезы о независимости случайных величин является одним из основных этапов системного анализа статистических данных. На её результатах осуществляется синтез эффективных алгоритмов принятия решений. Традиционная методика проверки гипотезы о независимости случайных величин основана на использовании критерия Пирсона и содержит трудно формализуемый этап разбиения области значений случайных величин на многомерные интервалы. Предложена методика проверки гипотезы о независимости случайных величин, которая использует непараметрический алгоритм распознавания образов, соответствующий критерию максимального правдоподобия. Её применение позволяет обойти проблему декомпозиции области значений случайных величин на интервалы. Идея подхода состоит в формировании по исходным статистическим данным обучающей выборки для решения двухальтернативной задачи распознавания образов. Каждый класс определяется в предположении независимости либо зависимости случайных величин, что проявляется в различии их законов распределения в классах. В этих условиях появляется возможность замены исходной гипотезы на задачу проверки достоверности отличия вероятностей ошибок распознавания образов в классах. С использованием аппарата теории графов предлагаемая методика развита при формировании наборов независимых случайных величин. Полученные результаты обобщены при проверке гипотезы о независимости случайных величин для больших объёмов статистических данных на основе сжатия исходной информации. Это позволяет на порядки повысить вычислительную эффективность решаемой задачи. В статье обосновывается методика проверки гипотезы о независимости случайных величин, основанная на использовании непараметрического алгоритма распознавания образов в условиях больших объёмов статистических данных. Приводятся результаты сравнения методики с общепризнанным критерием согласия Пирсона при исследовании неоднозначных зависимостей между случайными величинами различной сложности. Эффективность предлагаемой методики подтверждается результатами применения при обработке информации дистанционного зондирования антропогенных территорий в окрестности города Красноярска.

Ключевые слова: проверка гипотезы о независимости случайных величин, ядерная оценка плотности вероятности, регрессионная оценка плотности вероятности, распознавание образов, критерий Пирсона, дистанционное зондирование.

Nonparametric method for testing the hypothesis of independence of random variables and its application in the analysis of remote sensing data

A. V. Sharueva

Reshetnev Siberian State University of Science and Technology
31, Krasnoyarskii rabochii prospekt, Krasnoyarsk, 660037, Russian Federation
E-mail: anna-denisyuk@yandex.ru

Abstract. Testing the hypothesis of independence of random variables is one of the main stages of system analysis of statistical data. Based on its results, the synthesis of effective decision-making algorithms is carried out. The traditional method of testing the hypothesis of independence of random variables is based on the use of the Pearson criterion, which contains a difficult to formalize stage of dividing the range of the values of random variables into multidimensional intervals. A method for testing the hypothesis of independence of random variables is proposed, which uses a nonparametric pattern recognition algorithm corresponding to the maximum likelihood criterion. Its application makes it possible to circumvent the problem of decomposing the range of the values of random variables into intervals. The idea of the approach is to form a training sample based on the initial statistical data to solve a two-alternative pattern recognition problem. Each class is defined under the assumption of independence or dependence of random variables, which is manifested in the difference in their distribution laws in the classes. Under these conditions, it becomes possible to replace the initial hypothesis with the task of checking the reliability of the difference in the probabilities of pattern recognition errors in classes. Using the apparatus of graph theory, the proposed method is developed in the formation of sets of independent random variables. The obtained results are generalized when testing the hypothesis of independence of random variables for large volumes of statistical data based on compression of the original information. This allows increasing the computational efficiency of the problem being solved. The article substantiates the method for testing the hypothesis of independence of random variables, based on the use of a nonparametric pattern recognition algorithm in conditions of large volumes of statistical data. The results of comparing the technique with the generally recognized Pearson consensus criterion in the study of ambiguous dependences between random variables of varying complexity are presented. The effectiveness of the proposed method is confirmed by the results of its application in processing remote sensing information from anthropogenic territories in the vicinity of the city of Krasnoyarsk.

Keywords: testing the hypothesis of independence of random variables, kernel probability density estimation, regression probability density estimation, pattern recognition, Pearson criterion, remote sensing.

Introduction

The universal and generally accepted criterion for testing hypotheses about the distributions of random variables, including their independence, is the Pearson criterion [1]. When using it, it is necessary to solve the problem of partitioning the area of values of random variables into multivariate intervals and to establish the law of distribution of the criterion that determines the dependences between the probabilistic characteristics of random variables. In [2–4] a new approach is proposed that allows simplifying the test of the hypothesis of independence of random variables using a nonparametric algorithm of nuclear-type pattern recognition corresponding to the maximum likelihood criterion. The idea of the approach is to solve a two-alternative problem of pattern recognition. The classes being considered are defined by assumptions about dependence and independence of random variables. On this basis, a training sample is formed from the initial statistical data on observations of random variables and the problem of pattern recognition is solved. The ratio between the estimates of recognition error probabilities of the introduced classes confirms or refutes the hypothesis being considered.

The purpose of this paper is to generalise and develop a nonparametric method of testing the hypothesis of independence of random variables for conditions of the large volume of statistical data and its application in the analysis of information on remote sensing of anthropogenic territories.

Methodology for testing the hypothesis of independence of random variables

Let there be a sample $V = \left(x^i, i = \overline{1, n} \right)$ of the n volume, composed of independent observations of a two-dimensional random variable $x = (x_1, x_2)$. Let us suppose that the V sample is drawn from the general population characterised by the densities of the probabilities $p(x_1) p(x_2)$ or $p(x_1, x_2)$. On the basis of statistical data of V it is necessary to test the hypothesis

$$H_0 : \ p(x_1, x_2) \equiv p(x_1) p(x_2)$$

of independence of random variables x_1, x_2.

To test the H_0 hypothesis let us solve the two-alternative problem of pattern recognition. By classes Ω_1, Ω_2 areas of definition for probability densities $p(x_1) p(x_2)$, $p(x_1, x_2)$ are meant. Under these conditions, the Bayesian decision rule corresponding to the maximum likelihood criterion has the following form

$$m(x): \begin{cases} x \in \Omega_1, \text{if } p(x_1, x_2) < p(x_1) p(x_2), \\ x \in \Omega_2, \text{if } p(x_1, x_2) > p(x_1) p(x_2). \end{cases}$$

In contrast to the traditional formulation of the pattern recognition problem, while synthesizing a decisive rule $m(x)$ there is no a priori training sample containing information about the belonging of the V sample elements to one or another class. This information must be discovered in the process of implementation of the H_0 hypothesis testing methodology, which is based on the following actions.

From the V sample recover probability densities $p(x_1, x_2), p(x_1) p(x_2)$, using their nonparametric Rosenblatt – Parzen type estimates [5; 6],

$$\overline{p}(x_1, x_2) = \frac{1}{n c_1 c_2} \sum_{i=1}^{n} F\left(\frac{x_1 - x_1^i}{c_1} \right) F\left(\frac{x_2 - x_2^i}{c_2} \right),$$

$$\overline{p}(x_1) \overline{p}(x_2) = \frac{1}{n^2 c_1 c_2} \sum_{i=1}^{n} \sum_{j=1}^{n} F\left(\frac{x_1 - x_1^i}{c_1} \right) F\left(\frac{x_2 - x_2^j}{c_2} \right).$$

In the statistics $\overline{p}(x_1, x_2)$, $\overline{p}(x_1) \overline{p}(x_2)$ nuclear functions $F(u_v)$ satisfy the conditions of positivity, symmetry and normalization.

The values of blurring coefficients c_v, $v = 1, 2$ of the nuclear functions decrease as the n volume of the V sample of statistical data increases. Then the nonparametric decision rule for classification of random variables $x = (x_1, x_2)$ is written as follows

$$\overline{m}(x): \begin{cases} x \in \Omega_1, \text{if } \overline{p}(x_1, x_2) < \overline{p}(x_1) \overline{p}(x_2), \\ x \in \Omega_2, \text{if } \overline{p}(x_1, x_2) > \overline{p}(x_1) \overline{p}(x_2). \end{cases}$$

The optimal blurring coefficients of the nuclear functions of the $\overline{m}(x)$ decision rule are chosen on the basis of the analysis of approximation properties of nonparametric estimates of probability densities $\overline{p}(x_1, x_2)$, $\overline{p}(x_1)$, $\overline{p}(x_2)$ from the minimum condition, their corresponding estimates of standard deviations from the $p(x_1, x_2)$, $p(x_1)$, $p(x_2)$. For example, for $\overline{p}(x_1)$ such a criterion is [7–11]

$$\int\limits_{-\infty}^{\infty} \bar{p}^2(x_1)\,dx_1 - \frac{2}{n}\sum_{j=\grave{}}^{n} \bar{p}\left(x_1^j\right).$$

Let us define the estimates of probabilities of pattern recognition errors $\bar{\rho}_1\left(\bar{c}(1)\right)$, $\bar{\rho}_2\left(\bar{c}(2)\right)$ using the $\bar{m}(x)$ dsision rule on the basis of raw statistical data of V at optimal blurring coefficients $\bar{c}(1) = \left(\bar{c}_1(1), \bar{c}_2(1)\right)$, $\bar{c}(2) = \left(\bar{c}_1(2), \bar{c}_2(2)\right)$ of the nuclear functions of statistics $\bar{p}(x_1)\,\bar{p}(x_2)$, $\bar{p}(x_1, x_2)$ respectively.

The values $\bar{\rho}_t\left(\bar{c}(1), \bar{c}(2)\right)$ are calculated in the 'rolling examination' mode on the V sample assuming that its elements belong to the Ω_t class.

$$\bar{\rho}_t\left(\bar{c}(1), \bar{c}(2)\right) = \frac{1}{n}\sum_{j=1}^{n} 1\left(\delta(j), \bar{\delta}(j)\right),\ t = 1, 2,$$

where $\delta(j) = t$ are designations of the type of $x^t = \left(x_1^t, x_2^t\right) \in \Omega_t$;

$$\bar{\delta}(j) = \begin{cases} t, \text{if } x^j \in \Omega_t \\ 0, \text{if } x^j \notin \Omega_t \end{cases}$$

– «solving» the algorithm of $\bar{m}(x)$ about the belonging of the x^j situation to on of the classes Ω_t, $t = 1, 2$.

While calculating $\bar{\rho}_t\left(\bar{c}(1), \bar{c}(2)\right)$ in accordance with the 'rolling examination' methodology the situation $x^j = \left(x_1^j, x_2^j\right)$ from the V sample, which is fed into the algorithm of $\bar{m}(x)$ for control, is excluded from the process of producing statistics $\bar{p}(x_1, x_2)$, $\bar{p}(x_1)\,\bar{p}(x_2)$.

The indicator function is defined by the expression

$$1\left(\delta(j), \bar{\delta}(j)\right) = \begin{cases} 0, \textit{if } \delta(j) = \bar{\delta}(j), \\ 1, \textit{if } \delta(j) \neq \bar{\delta}(j). \end{cases}$$

Let us denote by $\bar{\bar{\rho}}_t$ the value of the estimation of the probability of pattern recognition error assuming that the V sample elements belong to the class Ω_t, $t = 1, 2$. Let us compare the values $\bar{\bar{\rho}}_1$, $\bar{\bar{\rho}}_2$.

Then the H_0 hypothesis is valid if $\bar{\bar{\rho}}_1 < \bar{\bar{\rho}}_2$. Otherwise, at $\bar{\bar{\rho}}_2 < \bar{\bar{\rho}}_1$ the random variables x_1 and x_2 are independent.

When the n volume of the V smple is limited, the problem of confidence estimation of probabilities of pattern recognition errors arises. For its solution, the traditional methodology of confidence estimation of probabilities or Kolmogorov – Smirnov criterion is used.

For example, when using the Kolmogorov – Smirnov criterion, the deviation $\bar{D}_{12} = \left|\bar{\bar{\rho}}_1 - \bar{\bar{\rho}}_2\right|$ is compared to the threshold value [12]

$$D_\beta = \sqrt{-\ln\left(\frac{\beta}{2}\right)\Big/ n}\ .$$

Here β is a probability (risk) of rejecting the hypothesis $\bar{H}_0$: $\rho_1 = \rho_2$. If the ratio $\bar{D}_{12} < D_\beta$ is satisfied, then the $\bar{H}_0$ hypothesis is valid and the risk of rejecting it does not exceed the value of β. At $\bar{D}_{12} > D_\beta$ the $\bar{H}_0$ hypothesis rejected.

Formation of sets of independent random variables

There is a sample of observations $V = \left(x_v^i, v = \overline{1,k}, i = \overline{1,n} \right)$ of the n volume composed of statistically independent observations of the components of the multivariate random variable $x = \left(x_v, v = \overline{1,k} \right)$. The type of the $p(x)$ probability density function is unknown a priori. It is necessary according to the statistics of V, using the hypothesis testing criterion proposed above [13–16]

$$H_{vj}: \ p\left(x_v, x_j \right) \equiv p\left(x_v \right) p\left(x_j \right)$$

For the components x_v, $v = \overline{1,k}$, x_j, $j = \overline{1,k}$, $v > j$, to form the sets of the independent random variables $x(t) = \left(x_v, v \in I_t \right)$, $t = \overline{1,m}$. The m number of sets of components of the random variable x is unknown, and I_t is a set of component numbers that make up the set $x(t)$.

The proposed methodology is based on performing the following steps:

1. In accordance with the above recommendations, to test the H_{vj} hypotheses for each pair of the components $\left(x_v, x_j \right)$ of the multivariate random variable $x = \left(x_v, v = \overline{1,k} \right)$. The number of such pairs corresponds to the value $k(k-1)/2$.

2. Based on the results of step 1, construct an information graph $G(X, A)$, where X is a set of its vertices corresponding to the components of the random variable x, and A is a set of edges. Between the two vertices x_v, x_j there is an edge if the H_{vj} hypothesis is satisfied, i.e. the components x_v, x_j are independent.

3. Analyse the information graph $G(X, A)$ and determine its complete subgraphs $G(X_t, A_t)$, $t = \overline{1,m}$. Each vertice of the subgraph $G(X_t, A_t)$ has an edge if the components of the random variable x are independent. Detect complete subgraphs using algorithms for cutting the original graph, which are based on analysing its adjacency matrix. The components x_v, $v \in I_t$ correwsponding to the vertices of the complete subgraph $G(X_t, A_t)$ form a set of independent random variables.

Modification of the method of testing the hypothesis of independence of random variables in conditions of large volumes of statistical data

With large n volumes of the statistical data $V = \left(x_1^i, x_2^i, i = \overline{1,n} \right)$ regression estimates of probability densities $\overline{p}(x_1, x_2)$, $\overline{p}(x_1)$, $\overline{p}(x_2)$ are used in the proposed methodology. These estimates are based on the compression of the original information, e.g., $V_1 = \left(x_1^i, i = \overline{1,n} \right)$ into the data array $\overline{V}_1 = \left(\overline{p}_1^j, z^j, j = \overline{1,N} \right)$ by decomposing the area of values x_1 into N intervals. Here z^j are centres of sampling intervals of values x_1, and $\overline{p}_1^j = \overline{P}_1^j / \Delta$ is probability density estimation in the jth interval; Δ is a sampling interval length; $\overline{P}_1^j$ is frequency of occurrence of the x_1^i values from the V_1 sample in the interval numbered j. Then the regression estimate of the probability density function $p(x_1)$ according to $\overline{V}_1$ has the form [17; 18]

$$\overline{p}(x_1) = \frac{1}{c_1} \sum_{j=1}^{N} \overline{P}_1^j \Phi\left(\frac{x_1 - z^j}{c_1} \right).$$

The proposed approach allows reducing by orders of magnitude the n volume of initial statistical information when estimating probability densities. The peculiarity of the statistics of the $\overline{p}(x_1)$ type

allows simplifying considerably the choice of coefficients of c blurring of nuclear functions in the $\overline{p}(x_1)$ statistics from the condition of minimum criterion

$$\frac{1}{N}\sum_{i=1}^{N}\left(\overline{p}_1^i - \overline{p}\left(x_1^i\right)\right).$$

By analogy the estimation of the probability densities $p(x_2)$, $p(x_1, x_2)$ is carried out. Regression estimates of probability densities are used in testing the hypothesis of independence of random variables according to the proposed methodology.

Analysing the results of the computational experiment

The effectiveness of the proposed method of testing the hypothesis of independence of two-dimensional random variables and Pearson's criterion in the conditions of ambiguous dependences at different volumes of statistical data has been compared [19–21]. The sensors of random variables x_1, x_2 were formed on the basis of the uniform distribution law x_1, which was used in the calculation of the values of x_2 in the form of nonlinear transformations x_1. At the same time the values of x_2 were superimposed with disturbances with the normal distribution law, which has zero mathematical expectation and standard deviation σ. An example of the values of random variables x_1 and x_2 is shown in Fig. 1.
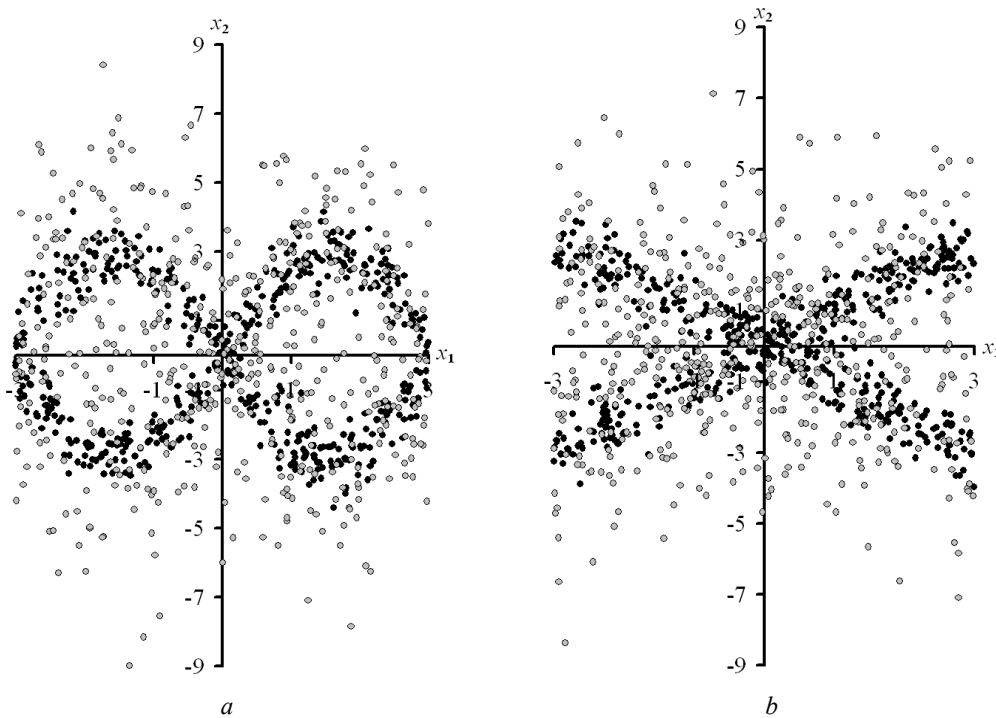

Рис. 1. Значения случайных величин x_1, x_2 из выборки исходных статистических данных V при $n = 500$ и $\sigma = 0{,}5$ (темные точки), а при $\sigma = 2$ (серые точки) при использовании зависимостей различной сложности

Fig. 1. Values x_1, x_2 of random variables from a sample of initial statistical data V at $n = 500$ and $\sigma = 0.5$ (dark dots), and at $\sigma = 2$ (grey dots) when using dependencies of varying complexity

When testing the independence hypothesis of a two-dimensional random component based on the Pearson criterion, the results of the optimal selection of the number of sampling intervals are used [22–24]

$$N^* = \left(\frac{3}{4} \Delta_1 \Delta_2 \left\| p(x_1, x_2) \right\|^2 n \right)^{1/2}.$$

The value $\left\| p(x_1, x_2) \right\|^2 = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} p^2(x_1, x_2)\, dx_1\, dx_2$, and Δ_v is the length of the interval between values of the random value x_v, $v = 1.2$. The works [25–27] are devoted to the traditional formulas of discretization of the range of values of random quantities.

By the results of computational experiment the offered methodology and Pearson's criterion at the analysis of ambiguous dependences between random variables in conditions of relatively small volumes of statistical data and mean square deviations σ of interferences are comparable and unmistakably determine dependence of random variables. This conclusion does not hold for the dependence between random variables (Fig. 1, *a*), when the Pearson criterion does not establish dependence under the conditions $n = 100$ and $\sigma \in [0.5; 2]$. As σ increases, the efficiency of the criteria being compared decreases. This fact is explained by the peculiarities of ambiguous dependences and large values of σ, when the area of definition of random variables hides the desired dependence. With the increase in the *n* volume of initial data the efficiency of the compared criteria for testing the hypothesis of independence of random variables increases. This conclusion is expected, since asymptotic properties of nonparametric estimates of probability densities and frequencies of occurrence of random variables in their two-dimensional intervals rise as *n* increases. The advantage of the proposed methodology for testing the hypothesis of independence of random variables is observed at small values of σ, limited and large *n*. At large *n* and σ, the advantage of Pearson's criterion is often revealed if the procedure of optimal discretisation of the area of values of a two-dimensional random variable is used [22].

Application of the proposed methodology in analysing remote sensing data

The developed methodology was tested when analysing the remote sensing data [2; 28]. The object of the study is anthropogenic territories (quarry, suburban development) in the vicinity of the city of Krasnoyarsk. The initial information was formed on the fragments of Sentinel-2 satellite imagery on 26.08.2021 (Fig. 2). The spectral channels x_j, $j = \overline{1, 9}$ were used. These channels are characterised by wavelengths (nanometres): x_1 – (458–523), x_2 – (543–578), x_3 – (650–680), x_4 – (698–713), x_5 – (733–748), x_6 – (773–793), x_7 – (785–899), x_8 – (1565–1655), x_9 – (2100–2280).

The proposed methodology allows forming pairs of independent and dependent random variables by changing the ratio between their parameters. The application of the methodology allowed us to detect 31 and 29 pairs of spectral features with strong linear dependence for the objects 'quarry' and 'suburban development', respectively. The obtained results are presented in Fig. 3.

Additionally, non-linear dependences between spectral features were found for the object 'quarry'

$$(x_1, x_9), (x_1, x_8), (x_1, x_7), (x_1, x_5), (x_1, x_4)$$

and the object 'suburban development'

$$(x_7, x_9), (x_4, x_9), (x_3, x_9), (x_2, x_9), (x_1, x_9), (x_1, x_8), (x_1, x_7).$$

The obtained results are reliable for all pairs of spectral features, since the condition $\left| \bar{\rho}_1 - \bar{\rho}_2 \right| > D_\beta$ is met at $D_\beta = 0.029$ and the risk $\beta = 0.025$ reject the H_0 hypothesis of equality of values ρ_1, ρ_2.
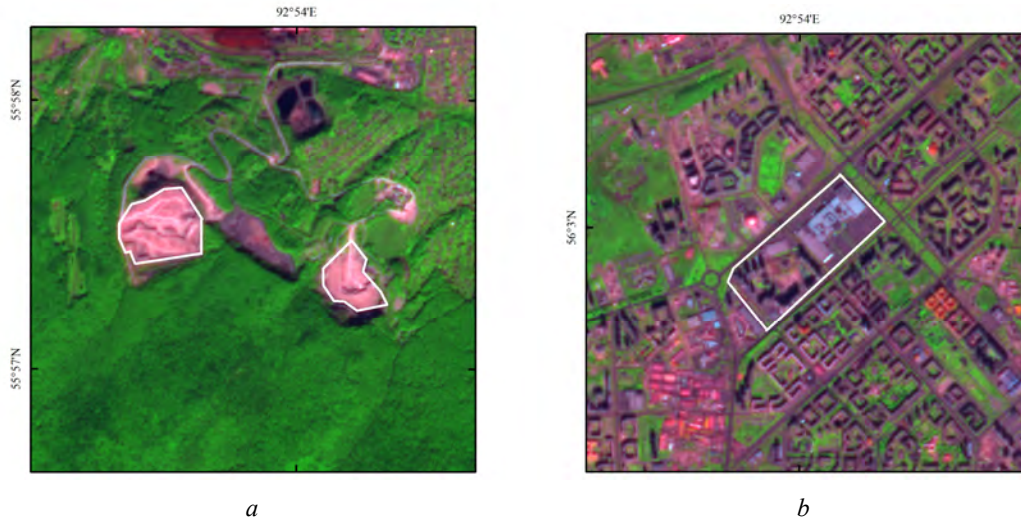
Рис. 2. Фрагменты спутниковой съемки Sentinel-2. Антропогенные территории:
a – карьер; *b* – пригородная застройка

Fig. 2. Fragments of Sentinel-2 satellite imagery. Anthropogenic territories:
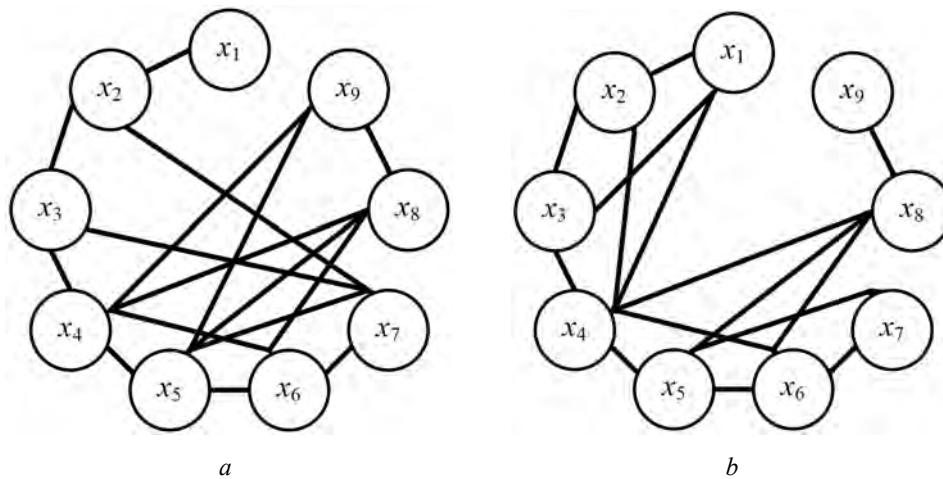*a* – quarry; *b* – suburban development

Рис. 3. Иллюстрация сильной линейной зависимости между парами спектральных признаков (x_i, x_j),
характеризующихся оценками коэффициентов корреляции больше 0,9:
a – карьер; *b* – пригородная застройка

Fig. 3. Illustration of a strong linear relationship between pairs of spectral features (x_i, x_j)
characterized by correlation coefficient estimates greater than 0.9:
a – quarry; *b* – suburban development

The problem of detecting anthropogenic areas from spectral data is considered. The error of their recognition in the space of spectral features $x = \left(x_j, j = \overline{1,9} \right)$ based on the training sample $V = \left(x^i, \sigma(i), i = \overline{1,n} \right)$ is equal to 0.012, where $n = n_1 + n_2$, $n_1 = 3377$ ('quarry', $\sigma(i) = 1$), $n_2 = 5049$ ('Suburban Development' $\sigma(i) = 2$). When excluding from the training sample, for example, the spectral features (x_4, x_5), (x_5, x_6), (x_4, x_5, x_6) the estimates of pattern recognition errors correspond to the values 0.011; 0.01; 0.008. The obtained reduction in pattern recognition errors is not reliable com-

pared to the error estimate in feature space x_j, $j = \overline{1, 9}$. Nevertheless, the obtained result justifies the possibility of reducing spectral features in the synthesis of decision-making algorithms and simplifying their optimisation.

Conclusion

The methodology of testing the hypothesis of independence of pairs of random variables, based on the use of nonparametric algorithm of pattern recognition, allows bypassing the problem of discretisation of the area of the values of random variables into multidimensional intervals. This problem is inherent in the generally recognised Pearson criterion. The conditions of competence of the proposed method and Pearson's criterion in the analysis of unambiguous and ambiguous dependences between random variables are determined. Using the apparatus of graph theory, the proposed method is developed in the formation of sets of independent random variables. The obtained results are generalised in testing the hypothesis of independence of random variables for large volumes of statistical data on the basis of compression of initial information, which allows increasing by orders of magnitude the computational efficiency of the problems being solved. The effectiveness of the proposed methodology is confirmed when analysing remote sensing data of anthropogenic territories and assessing their states. In the presence of a set of spectral features characterised by a strong linear dependence between its pairs, it is possible to reduce the number of spectral features in the recognition of anthropogenic territories with a decrease in the estimate of the probability of error in their recognition.

Библиографические ссылки

1. Академик Пугачёв Владимир Семёнович: к столетию со дня рождения / под ред. И. Н. Синицына. М. : Торус Пресс, 2011. 376 с.

2. Шаруева А. В., Лапко А. В., Лапко В. А. Непараметрические методы проверки гипотез о распределениях случайных величин при анализе данных дистанционного зондирования. Новосибирск : СО РАН, 2024. 189 с.

3. Лапко А. В., Лапко В. А. Проверка гипотезы о независимости двумерных случайных величин с использованием непараметрического алгоритма распознавания образов // Автометрия. 2021. Т. 57, № 2. С. 41–48. DOI: 10.15372/AUT20210205.

4. Лапко А. В., Лапко В. А., Бахтина А. В. Исследование методики проверки гипотезы о независимости двухмерных случайных величин с использованием непараметрического классификатора // Автометрия. 2021. Т. 57, № 6. С. 90–100. DOI: 10.15372/AUT20210610.

5. Parzen E. On estimation of a probability density function and mode // Annals of Mathematical Statistics. 1962. Vol. 33, No. 3. P. 1065–1076. DOI: 10.1214/aoms/1177704472.

6. Епанечников В. А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. Т. 14, №1. С. 156–161.

7. Лапко А. В., Лапко В. А. Анализ методов оптимизации непараметрической оценки плотности вероятности по коэффициенту размытости ядерных функций // Измерительная техника. 2017. № 6. С. 3–8.

8. Лапко А. В., Лапко В. А. Ядерные оценки плотности вероятности и их применение. Красноярск : СибГУ им. М. Ф. Решетнёва, 2021. 308 с.

9. Rudemo M. Empirical choice of histogram and kernel density estimators // Scandinavian Journal of Statistics. 1982. No. 9. P. 65–78.

10. Bowman A. W. A comparative study of some kernel-based non-parametric density estimators // Journal of Statistical Computation and Simulation. 1982. Vol. 21. P. 313–327.

11. Hall P. Large-sample optimality of least squares cross-validation in density estimation // Annals of Statistics. 1983. Vol. 11. P. 1156–1174.

12. Шаракшанэ А. С., Железнов И. Г., Ивницкий В. А. Сложные системы. М. : Высш. шк., 1977. 248 с.

13. Лапко А. В., Лапко В. А., Бахтина А. В. Формирование наборов независимых компонент многомерной случайной величины на основе непараметрического алгоритма распознавания образов // Измерительная техника. 2021. № 9. С. 3–9. DOI: 10.32446/0368-1025it.2021-9-3-9.

14. Непараметрический алгоритм распознавания образов в задаче проверки гипотезы о независимости случайных величин / И. В. Зеньков, А. В. Лапко, В. А. Лапко и др. // Компьютерная оптика. 2021. Т. 45, № 5. С. 767–772. DOI: 10.18287/2412-6179-CO-871.

15. Методика последовательного формирования набора компонент многомерной случайной величины с использованием непараметрического алгоритма распознавания образов / И. В. Зеньков, А. В. Лапко, В. А. Лапко и др. // Компьютерная оптика. 2021. Т. 45. № 6. С. 926–933. DOI: 10.18287/2412-6179-CO-902.

16. Лапко А. В., Лапко В. А., Шаруева А. В. Непараметрический алгоритм распознавания образов в задаче формирования наборов независимых случайных величин // Информатика и системы управления. 2024. Т. 79, №1. С. 81–90. DOI: 10.22250/18142400_2024_79_1_81.

17. Лапко А. В., Лапко В. А. Регрессионная оценка плотности вероятности и её свойства // Системы управления и информационные технологии. 2012. № 3-1 (49). С. 152–156.

18. Лапко А. В., Лапко В. А. Регрессионная оценка многомерной плотности вероятности и её свойства // Автометрия. 2014. Т. 50, № 2. С. 50–56.

19. Лапко А. В., Лапко В. А., Бахтина А. В. Применение непараметрического алгоритма распознавания образов в задаче проверки гипотезы о независимости переменных неоднозначных функций // Измерительная техника. 2022. № 1. С. 17–22. DOI: 10.32446/0368-1025it.2022-01-17-22.

20. Лапко А. В., Лапко В. А., Бахтина А. В. Сравнение методики проверки гипотезы о независимости двухмерных случайных величин, основанной на непараметрическом классификаторе // Искусственный интеллект и принятие решений. 2022. № 1. С. 45–56. DOI: 10.14357/20718594220105.

21. Лапко А. В., Лапко В. А., Бахтина А. В. Сравнение методик проверки гипотезы о независимости случайных величин, основанных на непараметрическом классификаторе и критерии Пирсона // Автометрия. 2023. Т. 59, № 5. С. 36–46.

22. Лапко А. В., Лапко В. А. Выбор оптимального количества интервалов дискретизации области значений двухмерной случайной величины // Измерительная техника. 2016. № 2. С. 14–17.

23. Лапко А. В., Лапко В. А. Метод дискретизации области значений многомерной случайной величины // Измерительная техника. 2019. № 1. С. 16–20. DOI: 10.32446/0368-1025it.2019-1-16-20.

24. Лапко А. В., Лапко В. А. Оценивание параметров формулы оптимальной дискретизации области значений двумерной случайной величины // Измерительная техника. 2018. № 5. С. 9–13.

25. Sturges H. A. The choice of a class interval // Journal of the American Statistical Association. 1926. Vol. 21, No. 153. P. 65–66.

26. Heinhold I., Gaede K. Ingeniur statistic. München: Springler Verlag, 1964. 352 p.

27. Scott D.W. Multivariate Density Estimation: Theory, Practice, and Visualization. New Jersey: John Wiley & Sons, 2015. 384 p.

28. Лапко А. В., Лапко В. А., Бахтина А. В. Применение непараметрической методики проверки гипотезы о независимости случайных величин в условиях большого объёма статистических данных // Измерительная техника. 2023. № 10. С. 17–24. DOI: 10.32446/0368-1025it.2023-10-17-24.

References

1. Sinitsyna I. N. *Akademik Pugachev Vladimir Semenovich: k stoletiyu so dnya rozhdeniya* [Academician Vladimir Semenovich Pugachev: on the centenary of his birth]. Moscow, Torus Press Publ., 2011, 376 p.

2. Sharueva A. V., Lapko A. V., Lapko V. A. *Neparametricheskiye metody proverki gipotez o ra-spredeleniyakh sluchaynykh velichin pri analize dannykh distantsionnogo zondirovaniya* [Nonpara-metric methods for testing hypotheses about distributions of random variables in the analysis of re-mote sensing data]. Novosibirsk, SO RAN Publ., 2024, 189 p.

3. Lapko A. V., Lapko V. A. Testing the Hypothesis of the Independence of Two-Dimensional Random Variables Using a Nonparametric Algorithm for Pattern Recognition. *Optoelectronics, Instrumentation and Data Processing*. 2021, Vol. 57, No. 2, P. 149–155.

4. Lapko A. V., Lapko V. A., Bakhtina A. V. Study of the Method for Verification of the Hypothe-sis on Independence of Two-Dimensional Random Quantities Using a Nonparametric Classifier. *Optoelectronics, Instrumentation and Data Processing*. 2022, Vol. 57, No. 6, P. 639–648.

5. Parzen E. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*. 1962, Vol. 33, No. 3, P. 1065–1076.

6. Epanechnikov V. A. [Non-parametric estimation of a multivariate probability density]. *Theory of Probability & Its Applications*. 1969, Vol. 14, No. 1, P. 156–161 (In Russ.).

7. Lapko A. V., Lapko V. A. Analysis of optimization methods for nonparametric estimation of the probability density with respect to the blur factor of kernel functions. *Measurement Techniques*. 2017, Vol. 60, No. 6, P. 515–522.

8. Lapko A. V., Lapko V. A. *Yadernye otsenki plotnosti veroyatnosti i ikh primenenie* [Kernel probability density estimates and their applications]. Krasnoyarsk, SibGU im. M.F. Reshetnev Publ., 2021, 208 p.

9. Rudemo M. Empirical choice of histogram and kernel density estimators. *Scandinavian Journal of Statistics*. 1982, No. 9, P. 65–78.

10. Bowman A. W. A comparative study of some kernel-based non-parametric density estimators. *Journal of Statistical Computation and Simulation*. 1982, Vol. 21, P. 313–327.

11. Hall P. Large-sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*. 1983, Vol. 11, P. 1156–1174.

12. Sharakshane, A. S., Zheleznov I. G., Ivnitskii V. A. *Slozhnye sistemy* [Complex systems]. Moscow, Vysshaya Shkola Publ., 1977, 248 p.

13. Lapko A. V., Lapko V. A., Bakhtina A. V. Formation of Sets of Independent Components of a Multidimensional Random Variable Based on a Nonparametric Pattern Recognition Algorithm. *Measurement Techniques*. 2021, Vol. 64, No. 9, P. 689–696.

14. Zenkov I. V., Lapko A. V., Lapko V. A., Kiryushina E. V., Vokin V. N. Nonparametric pattern recognition algorithm for testing a hypothesis of the independence of random variables. *Computer Optics*. 2021, Vol. 45, No 5, P. 767–772.

15. Zenkov I. V., Lapko A. V., Lapko V. A., Kiryushina E. V., Vokin V. N., Bakhtina A. V. A method of sequentially generating a set of components of a multidimensional random variable using a nonparametric pattern recognition algorithm. *Computer Optics*. 2021, Vol. 45, No. 6, P. 926–933.

16. Lapko A. V., Lapko V. A., Sharueva A. V. Neparametricheskii algoritm raspoznavaniya obra-zov v zadache formirovaniya naborov nezavisimykh sluchainykh velichin [A nonparametric pattern recognition algorithm in the problem of forming sets of independent random variables]. *Informatika i sistemy upravleniya*. 2024, Vol. 79, No. 1, P. 81–90 (In Russ.).

17. Lapko A. V., Lapko V. A. Regressionnaya otsenka plotnosti veroyatnosti i ee svoistva [Re-gression estimation of probability density and its properties]. *Sistemy upravleniya i informatsionnye tekhnologii*. 2012, No. 3-1 (49), P. 152–156 (In Russ.).

18. Lapko A. V., Lapko V. A. Regression estimate of the multidimensional probability density and its properties. *Optoelectronics, Instrumentation and Data Processing*. 2014, Vol. 50, No. 2, P. 148–153.

19. Lapko A. V., Lapko V. A., Bakhtina A. V. Application of a nonparametric pattern recognition algorithm to the problem of testing the hypothesis of the independence of variables of multi-valued functions. *Measurement Techniques*. 2022, Vol. 65, No. 1, P. 17–23.

20. Lapko A. V., Lapko V. A., Bakhtina A. V. Comparison of the Methodology for Hypothesis Testing of the Independence of Two-Dimensional Random Variables Based on a Nonparametric Classifier. *Scientific and Technical Information Processing*. 2023, Vol. 50, No. 6, P. 572–581.

21. Lapko A. V., Lapko V. A., Bakhtina A. V. Comparison of Methods for Testing the Hypothesis of Independence of Random Variables Based on a Nonparametric Classifier and Pearson's Chi-Squared Test. *Optoelectronics, Instrumentation and Data Processing*, 2023, Vol. 59, No. 5, P. 551–560.

22. Lapko A. V., Lapko V. A. Selection of the Optimal Number of Intervals Sampling the Region of Values of a Two-Dimensional Random Variable. Measurement Techniques. 2016, Vol. 59, No. 2, P. 122–126.

23. Lapko A. V., Lapko V. A. Discretization method for the range of values of a multi-dimensional random variable. *Measurement Techniques*. 2019, Vol. 62, No. 1, P. 16–22.

24. Lapko A. V., Lapko V. A. Estimation of parameters of the formula for optimal discretization of the range of values of a two-dimensional random variable. *Measurement Techniques*. 2018, Vol. 61, No. 5, P. 427–433.

25. Sturges H. A. The choice of a class interval. *Journal of the American Statistical Association*, 1926, Vol. 21, P. 65–66.

26. Heinhold I., Gaede K. W. *Ingeniur statistic*. München, Springler Verlag, 1964, 352 p.

27. Scott D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New Jersey, John Wiley & Sons, 2015, 384 p.

28. Lapko A. V., Lapko V. A., Bakhtina A. V. Application of a nonparametric procedure for testing the hypothesis about the independence of random variables given a large amount of statistical data. *Measurement Techniques*, 2024, Vol. 66, P. 744–754.

Шаруева Анна Владлимировна – заведующий лабораторией дистанционного зондирования, ассистент; Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева. E-mail: anna-denisyuk@yandex.ru. https://orcid.org/0009-0003-4255-4554.

Sharueva Anna Vladimirovna – head of the remote sensing laboratory, assistant; Reshetnev Siberian State University of Science and Technology. E-mail: anna-denisyuk@yandex.ru. https://orcid.org/0009-0003-4255-4554.