Журнал обчислювальної та прикладної математики **2020, № * (***)** Journal of Numerical & Applied Mathematics

УДК 519.23
MSC 47N30

CLASSIFICATION OF MULTIVARIATE SAMPLES USING PETUNIN ELLIPSES

D. A. Klyushin, Ya. V. Shtyk

Faculty of Computer Science and Cybernetics, Taras Shevchenko Kiev National University, Kiev, Ukraine, E-mail: dokmed5@gmail.com.

КЛАСИФІКАЦІЯ БАГАТОВИМІРНИХ ВИБІРОК ЗА ДОПОМОГОЮ ЕЛІПСІВ ПЕТУНІНА

Д. А. Клюшин, Я. В. Штик

Факультет комп'ютерних наук та кібернетики, Київський національний університет імені Тараса Шевченка, Київ, Україна, E-mail: dokmed5@gmail.com.

Abstract. The method of classification multivariate samples using Petunin ellipses is investigated in the paper. Several different types of samples were generated for testing. Based on the calculated accuracy of the criteria advantages and disadvantages of each of the linear and quadratic criteria and the specifics of the method as a whole were discovered. It has been found that both linear and quadratic criteria give high accuracy for samples with small variance. As the variance increases, the accuracy of the linear criterion remains high, the accuracy of the quadratic criterion decreases. Both criteria are resistant to sample noise.
Keywords: Petunin Ellipse, Petunin Statistics, Multivariate Sample, Linear Discriminant, Quadratic Discriminant, Classification.

Анотація. В статті досліджується метод класифікації багатовимірних вибірок за допомогою еліпсів Петуніна. Для тестування було створено вибірки, які мають розподіли різних типів. На основі розрахованої точності запропонованих критеріїв класифікації були виявлені переваги та недоліки кожного з критеріїв та специфіка методу в цілому. Було встановлено, що як лінійні, так і квадратичні критерії дають високу точність для вибірок з невеликою дисперсією. Зі збільшенням дисперсії точність лінійного критерію залишається високою, точність квадратичного критерію зменшується. Обидва критерії стійкі до шуму. Цей важливий факт робить їх корисними у практичному застосуванні. При тестуванні неперервних даних метод будує набір статистик Петуніна, які є постійними для вибірок одного розміру. У зв'язку з цим виникають труднощі в побудові еліпсів Петуніна. Для вирішення описаної проблеми застосовано ймовірнісний підхід.
Ключові слова: еліпс Петуніна, статистика Петуніна, багатовимірна вибірка, лінійний дискримінант, квадратичний дискримінант, класифікація.

1. Introduction

Analyzing data, researchers often face the problem of small amount of available data. With the small amount of data available, achieving high accuracy is an interesting and challenging task. In particular, it is interesting to consider statistical nonparametric methods of classification that are accurate even for small samples (starting from the sample size 40).

To solve this task in the context of breast cancer diagnosis Petunin et al. [1], [2] developed methods of linear and quadratic multivariate analysis using projections of data onto the plane. Thus, they reduced the problem of the analysis of n-dimensional data to the problem of the analysis of 2-dimensional data. This method is based on the concept of the Petunin statistics citeKlyushin, which is a measure of samples' homogeneity and plays a role of coordinate at the plane. In the turn, the Petunin statistics uses the MP-model and Hill' assumption $A_{(n)}$ [4], [5], [6], [7].

The accuracy of the method, obtained in [2] is high, but the amount of test data was limited and small, and the properties of data distributions were not known. So, the purpose of the paper is to determine how accurate this method is at data artificially generated with special properties and what are its advantages and disadvantages. The novelty of the paper is that the linear and quadratic methods proposed in [1] [2] are investigated on the small sample set having different statistical properties for the first time.

2. Petunin ellipse

Let S be a set of equally distributed random variables with an unknown absolutely continuous distribution function F. Let $s_1, s_2, \ldots, s_n$ be sampled values from S. Sorting sampled values in ascending order we get a variation series $s_{(1)}, s_{(2)}, \ldots, s_{(n)}$. Its components called ordinal statistics.

Theorem 1 (Hill's assumption $A_{(n)}$)**.** *If F is an absolutely continuous distribution and $s_{(n+1)}$ is independently sampled from S, then*

$$P\{s_{(n+1)} \in (s_{(i)}, s_{(j)})\} = \frac{j-i}{n+1}$$

for $i = \overline{1 \ldots n-1}$, $j = \overline{1 \ldots n-1}$, $i < j$.

As it will be shown, to analyze the multivariate data using investigated algorithm, it is necessary to construct the Petunin ellipses in the R^2 space. The Petunin ellipse of the set S is the ellipse with the confidence area that includes n random points of set S with the probability $\frac{n-1}{n+1}$. At Figure 1 there is an example of Petunin ellipse.

Let us give a description of construction algorithm. The initial data for the algorithm for the construction of a Petunin ellipse is a set of two-dimensional points $T_n = \{(a_1, b_1), \ldots, (a_n, b_n)\}$. First, we construct the convex hull of the points T_n. Then we find two points (a_i, b_i) and (a_j, b_j) that lying on the diameter of the convex hull. Draw the line M through the points (a_i, b_i) and (a_j, b_j). Find the farthest vertices (a_k, b_k) and (a_l, b_l) of the convex hull from the line M. Draw through the points lines M_1 and M_2 parallel to the line M. It is

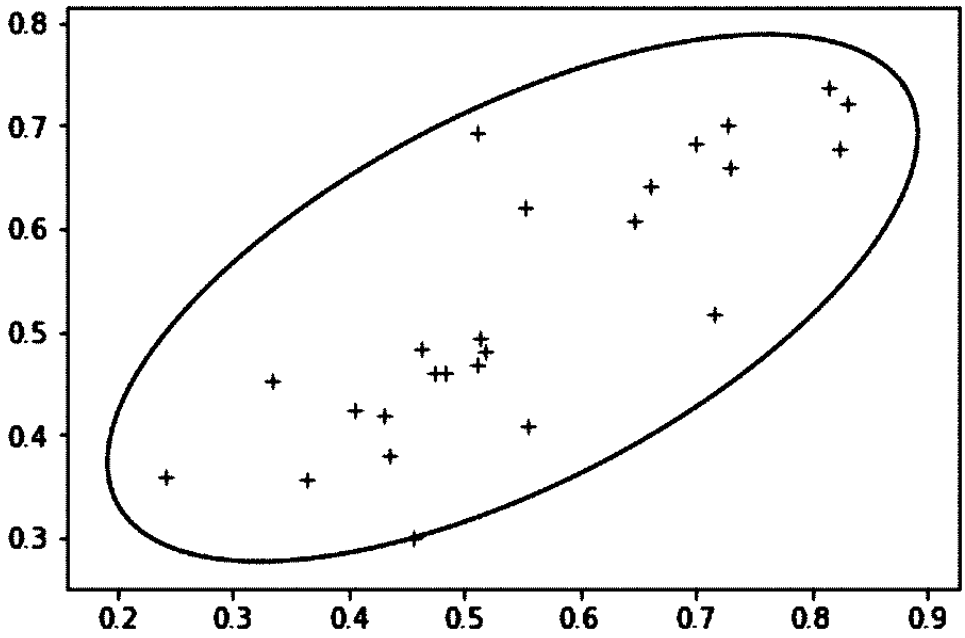

FIGURE 1. Petunin ellipse

important to note that the points (a_k, b_k) and (a_l, b_l) must be on opposite sides of the line M. If all points of the convex hull are on one side of line M, then the line M coincides with line M_2. Draw through the points (a_i, b_i) and (a_j, b_j) two lines M_3 and M_4 perpendicular to the line M. The intersection of the lines M_1, M_2, M_3 and M_4 form a rectangle whose sides have lengths k and m.

Without loss of generality, let $k \leq m$. Let's rotate and move the coordinate system so that the lower left corner of the rectangle is located at the beginning of the new coordinate system. The points $(a_1, b_1), \ldots, (a_n, b_n)$ will turn into points $(a_1^{'}, b_1^{'}), \ldots, (a_n^{'}, b_n^{'})$. Than we perform the system compression by multiplying the abscissas of the obtained points by $\mu = \frac{k}{m}$. Obtained points $(\mu a_1, b_1), \ldots, (\mu a_n, b_n)$ will lie in square K. Let (a_0, b_0) be a center of the square. Find the distances $d_1, d_2, \ldots, d_n$ from the obtained points to the center.

Now we construct a circle centered at the point (a_0, b_0) and radius $R = max(d_1, d_2, \ldots, d_n)$. All points $(a_1^{'}, b_1^{'}), \ldots, (a_n^{'}, b_n^{'})$ lie inside of obtained circle. Stretch the circle along the abscissa with coefficient $\frac{1}{\mu}$. Then we apply the inverse transforms of transfer and rotation and obtain an ellipse in the initial coordinate system. Obtained ellipse will be the Petunin ellipse.

3. Petunin statistics

The Petunin statistics is a measure that shows the likelihood that two samples were generated from the same distribution. In other words, this is a measure of the proximity between samples. It shows how similar the two samples are. Let us have two samples $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$ taken from two general sets S_1 and S_2 with unknown distribution functions F_{S_1} and F_{S_2} respectively. We want to find the probability with which we can argue that $F_{S_1} = F_{S_2}$.

Arrange the elements of each of the samples in ascending order. We obtain variation series $u = (u_{(1)}, u_{(2)}, \ldots, u_{(n_1)})$ and $v = (v_{(1)}, v_{(2)}, \ldots, v_{(n_2)})$. Let event I_{ij} consist in the fact that v_k belong to the interval $(u_{(i)}, u_{(j)})$. According to the Hill's hypothesis: $p(I_{ij}) = \frac{j-i}{n+1} = p_{ij}$. We can find the frequency f_{ij} of event I_{ij} on the given sample $v = (v_1, v_2, \ldots, v_n)$ and find confidence intervals for a given probability p_{ij}.

The formulas for the boundaries of confidence intervals are:

$$p_{ij}^{1,2} = \frac{f_{ij}n_2 + b^2/2 \mp \sqrt{f_{ij}(1-f_{ij})n_2 + b^2/4}}{n_2 + b^2}$$

In this case, we assume that n_2 is small, so according to the rule $'3\sigma'$ we assign $b = 3$. In general case, b must satisfy the condition $\Phi(b) = 1 - \gamma/2$, where Φ - Gaussian distribution function and γ is a given significance level. The number of possible confidence intervals $D_{ij} = (p_{ij}^1, p_{ij}^2)$ is equal to $N = \frac{n_1(n_1-1)}{2}$. Let D be the number of intervals for which the inclusion $p_{ij} \in (p_{ij}^1, p_{ij}^2)$ is true. Find the frequency $f = \frac{D}{N}$. Obtained frequency f will be measure of proximity of the two samples that called p-statistic or Petunin statistic.

5. Proximity measure for a sample and multiple samples

Let $u = (u_1, u_2, \ldots, u_m)$ be a sample from the general set S_1 with distribution F_{S_1}. Let $V = (v_1, v_2, \ldots, v_n)$ where $v_1 = (v_{11}, v_{12}, \ldots, v_{1n}), v_2 = (v_{21}, v_{22}, \ldots, v_{2n}), \ldots, v_m = (v_{m1}, v_{m2}, \ldots, v_{mn})$, be a set which elements is a sample from general set S_2 with distribution F_{S_2}. In the studied method, it was proposed to determine the measure of proximity between u and V by averaging the Petunin statistics.

First, find the value of Petunin statistics $p(v_i, u)$ for $i = \overline{1, \ldots, n}$. We average the obtained values and get a measure of proximity p between the sample and the set of samples:

$$p = \frac{1}{n}\sum_{i=1}^{n} p(v_i, u)$$

Let us try different approach that can be used based on the combining samples $v_1 = (v_{11}, v_{12}, \ldots, v_{1n}), v_2 = (v_{21}, v_{22}, \ldots, v_{2n}), \ldots, v_m = (v_{m1}, v_{m2}, \ldots, v_{mn})$ into one sample. So after combining we obtain sample $N = (v_{21}, v_{22}, \ldots, v_{2n}, v_{21}, v_{22}, \ldots, v_{2n}, \ldots, v_{m1}, v_{m2}, \ldots, v_{mn})$. We use N to find a measure of proximity p between the sample and the set of samples:

$$p = p(N, u).$$

4. Proximity measures between a sample and set of samples in a multidimensional case

Petunin et al. [1], [2] proposed and described a variant of feature-less discriminant analysis, based on tha Petunin ellipses and Petunin statistics. Let us consider groups of objects $V_1 = (a_1, a_2, \dots, a_n)$ and $V_2 = (b_1, b_2, \dots, b_m)$ from classes A and B respectively. Each object is represented via a matrix consisting of rows containing features. For example, in [1] and [2] objects are patients, and a matrix corresponding to a patient consists of rows containing some features of cell nuclei (nuclear area, integral density etc.). Each matrix has a certain number of rows (from 10 to 30). The columns of such matrices form samples from some distributions (from 10 to 30 real numbers). The main idea is to compare the samples pairwise reducing the multivariate case to a sequence of two-dimensional cases where axes are heterogeneity measures between samples and points are the pairs of the heterogeneous measure computed for each object.

Since the amount of data is small at the kth stage of the test we carry out the one-leave-out cross-validation excluding the kth sample P from the group V_1 and getting the group $V_1^{(k)} = (a_1, a_2, \dots, a_{k-1}, a_{k+1}, \dots, a_n)$. Let

$$Y_{M_1}^{(k)} = (y_{1k}^{(1)}, y_{2k}^{(1)}, \dots, y_K^{(1)})$$
$$Y_{M_2}^{(k)} = (y_{1k}^{(2)}, y_{2k}^{(2)}, \dots, y_K^{(2)})$$
$$\dots$$
$$Y_{M_{j_k}}^{(k)} = (y_{1k}^{(j_k)}, y_{2k}^{(j_k)}, \dots, y_K^{(j_k)})$$

be the object a_k, where $Y_i, i = \overline{1, \dots, j_k}$ is sub-objects of object a_k and j_k is a number of sub-objects in $a_k, k = \overline{1, \dots, n}$. In this case the number of features is K.

After exclusion we form a training sample for each feature $x_i, i = \overline{1, \dots, K}$. For feature x_1 the training sample has the form

$$Y_1^{(1)} = (y_{11}^{(1)}, y_{11}^{(2)}, \dots, y_{11}^{(j_1)})$$
$$Y_2^{(1)} = (y_{12}^{(1)}, y_{12}^{(2)}, \dots, y_{12}^{(j_2)})$$
$$\dots$$
$$Y_n^{(1)} = (y_{1n}^{(1)}, y_{1n}^{(2)}, \dots, y_{1n}^{(j_n)});$$

Similar samples obtain for all features. We apply the method described above for finding the heterogeneity measure for each feature. For example, for first feature y_1 the sample is the features y_1 of kth object $(y_{1k}^{(1)}, y_{1k}^{(2)}, \dots, y_{1k}^{(j_k)})$ and the set of samples is a training sample for the feature y_1. As a result, we get a heterogeneity measure $p_k^{(1)}$. In the same way we find heterogeneity measure of the object a_k and training sample for each feature $y_i, i = \overline{1, \dots, K}$. Replacing the object of group V_1 with the object of group V_2 we get heterogeneity measures $(\overline{p_k^{(1)}}, \overline{p_k^{(2)}}, \dots, \overline{p_k^{(K)}})$ for kth object of group V_2. If we replace the group V_1 by V_2 and carry out similar calculations, we obtain heterogeneity measures $(z_i^{(1)}, z_i^{(2)}, \dots, z_i^{(K)}), i = \overline{1 \dots n}$ and $(\overline{z_j^{(1)}}, \overline{z_j^{(2)}}, \dots, \overline{z_j^{(K)}}), j = \overline{1 \dots m}$.

For the ellipses construction we combine the obtained averaged Petunin statistics into pairs $(p_k^{(i)}, p_k^{(j)}), (\overline{p_t^{(i)}}, \overline{p_t^{(j)}}), k = \overline{1,\dots,n}$, $t = \overline{1,\dots,m}$, $i,j = \overline{1,\dots,K}$. Similarly we get the pairs $(z_k^{(i)}, z_k^{(j)}), (\overline{z_t^{(i)}}, \overline{z_t^{(j)}}), k = \overline{1,\dots,n}, t = \overline{1,\dots,m}, i,j = \overline{1,\dots,K}$. So we obtain points in space R^2. Then by point $(p_k^{(i)}, p_k^{(j)}), k = \overline{1\dots n}, i,j = \overline{1\dots K}$ we construct Petunin ellipses E_{ij}. Then we construct Petunin ellipses $\overline{E_{ij}}$ which contains points $(\overline{p_k^{(i)}}, \overline{p_k^{(j)}}), k = \overline{1,\dots,m}, i,j = \overline{1,\dots,K}$. Respectively using points $(z_k^{(i)}, z_k^{(j)})$ and $(\overline{z_k^{(i)}}, \overline{z_k^{(j)}})k = \overline{1,\dots,n}, i,j = \overline{1,\dots,K}$ we obtain Petunin ellipses D_{ts} and $\overline{D_{ts}}$.

The linear discriminant functions is using to separate two sets in space. We construct the linear discriminant functions that separate the set $S_{ij}^p = \{(p_t^{(i)}, p_t^{(j)}),\ t = \overline{1,\dots,n}\}$ from the set $\overline{S_{ij}^p} = \{(\overline{p_k^{(i)}}, \overline{p_k^{(j)}}),\ k = \overline{1,\dots,m}\}$ and set $S_{ij}^z = \{(z_t^{(i)}, z_t^{(j)}),\ t = \overline{1,\dots,n}\}$ from the set $\overline{S_{ij}^z} = \{(\overline{z_k^{(i)}}, \overline{z_k^{(j)}}),\ k = \overline{1,\dots,m}\}$. We construct the linear discriminants using the Roccio method so that the line $l_{ij}^p = \{(x,y) : f_{ij}(x,y) = 0\}$ is perpendicular to line on which the centers of the sets S_{ts}^p and $\overline{S_{ts}^p}$ lie and is located at the same distance from the centers. Moreover, the center of set S_{ts}^p is in the half-plane η_{ij} and the center of set $\overline{S_{ts}^p}$ is in the half-plane μ_{ij}. The function f_{ij}^* is constructed in the same way. So if we have K features we obtained $K(K-1)$ pairs of half-spaces (η_{ij}, μ_{ij}), $(\eta_{ij}^*, \mu_{ij}^*)$.

Let R be an object that needs to be classified. Using the algorithm from Section 2 we find averaged Petunin statistics p and d. We will analyse obtained heterogeneity measures using Petunin ellipses. Introduce the following events: $J_1 = \{(p_R^{(i)}, p_R^{(j)}) \in E_{ij}\}$, $J_2 = \{(p_R^{(i)}, p_R^{(j)}) \in \overline{E_{ts}}\}$, $J_3 = \{(p_R^{(i)}, p_R^{(j)}) \in E_{ts} - \overline{E_{ts}}\}$, $J_4 = \{(p_R^{(i)}, p_R^{(j)}) \in \overline{E_{ts}} - E_t s\}$, $J_1^* = \{(z_R^{(i)}, z_R^{(j)}) \in D_{ij}\}$, $J_2^* = \{(z_R^{(i)}, z_R^{(j)}) \in \overline{D_{ij}}\}$, $J_3^* = \{(z_R^{(i)}, z_R^{(j)}) \in D_{ts} - \overline{D_{ts}}\}$, $J_4^* = \{(z_R^{(i)}, z_R^{(j)}) \in \overline{D_{ts}} - D_{ts}\}$,
$K_1 = \{(p_R^{(i)}, p_R^{(j)}) \in \alpha_{ij}\}$, $K_2 = \{(p_R^{(i)}, p_R^{(j)}) \in \beta_{ij}\}$, $K_1^* = \{(z_R^{(i)}, z_R^{(j)}) \in \alpha_{ij}^*\}$, $K_2 = \{(z_R^{(i)}, z_R^{(j)}) \in \beta_{ij}\}, i < j$ $H_1 = J_3 \bigcup J_4^*$, $H_2 = J_4 \bigcup J_3^*$, $H_3 = J_1 \bigcup J_2^*$, $H_4 = J_2 \bigcup J_1^*$, $H_5 = K_1 \bigcup K_2^*$, $H_6 = K_2 \bigcup K_1^*$.

Let us introduce a relative frequencies $h_i = h(H_i), i = \overline{1,\dots,6}$ of each of the events $H_i, i = \overline{1,\dots,6}$ in N tests, which we can run changing values of i and j. The relative frequency h_1 corresponds to class A excluding overlapping with class B, h_3 corresponds to class A including overlapping with class B, h_2 and h_4 as relative frequency of class B excluding overlapping with class A and relative frequency of class B including overlapping with class B respectively. The frequencies h_5 and h_6 corresponds to the half-planes containing objects of class A and class B respectively.

7. Classification of multivariate samples

In this section, we test the investigated algorithm. To check the correctness of the algorithm, we will construct an experiment with linearly separable samples. We will generate samples using the Gaussian normal distribution.

We have two classes A and B. Each class object is a set of sub-objects. Each object has from 10 to 30 sub-objects. Sub-object is a set of features. Each feature has Gaussian distribution with variance equal to 1 and its own mean value. Let the number of features be K. Objects features of class A have average values from 0 to K. That is, the first feature is generated from the Gaussian distribution $N(0,1)$, the second one is generated from distribution $N(1,1)$, the fifteenth is generated from the distribution $N(14,1)$. Features of class B have Gaussian distributions with averages from 15 to 30 and a variance equal to 1. That is, the first feature of class B is generated from the distribution $N(15,1)$, the second one is generated from $N(16,1)$, the fifteenth is generated from $N(29,1)$.

We generate a sample that consist of 25 objects of each class. Then we perform one-leave-out cross-validation. In fact, at each iteration of the test we divide initial sample into a test sample, which consists of one object, and a training sample that includes all others objects. That is, the result of each iteration of the test is the conclusions about the belonging of the tested object to each of the two classes A and B. Conclusions are presented in the form of 6 statistics $h_i, i = \overline{1,\dots,6}$, which described below. As a result of the algorithm we obtain a set of relative frequencies of each object of initial sample. Analysing this set we can draw conclusion about correctness of the algorithm.

In [1], [2] the quadratic and linear tests for determining the class of a tested object were proposed. These tests were formulated as follows.

Quadratic test: $h_3 > h_4 \Rightarrow A$, $h_3 \leq h_4 \Rightarrow B$.

Linear test: $h_5 > h_6 \Rightarrow A$, $h_5 \leq h_6 \Rightarrow B$

Stable test: $h_3 > h_4 \wedge h_3 > h_2 \Rightarrow A$, $h_3 \leq h_4 \wedge h_1 \leq h_2 \Rightarrow B$

The effect of the stable dominance of the group A is the case when $h_3 > h_4$ and $h_5 > h_6$. Similarly the effect of the stable dominance of the group B is the case when $h_3 \leq h_4$ and $h_5 \leq h_6$. Analyzing h-statistics we obtain the accuracy of the described criteria. The accuracy of each criteria in percents is equal to 100. It can be seen that on a linearly separable sample the method has absolute accuracy. Obtained result allows us to argue that the algorithm is correct and it is possible to use it in practice. Now it is important to check the accuracy of the method on more complex samples. As we can see the accuracy of the criteria has decreased. At the same it is high level of accuracy. Obtained results suggest that for samples with a small variance we can successfully use studied method.

Analyzing h-statistics obtained for each of the cases, it was noticed that the h_1 is give an information of belonging to class A at most cases. The similar relations was noticed between statistics h_2 and class B. In this connection, let us introduce a modified test that is based only on h_1 and h_2 statistics. We formulate the criteria as follows: $h_1 > h_2 \Rightarrow A$, $h_1 \leq h_2 \Rightarrow B$. The accuracy of the introduced criterion for each of the cases is shown in Table 1.

We complicate the classification task by increasing the variance of the Gaussian distribution. Let the variance be 2 for all the distributions that are used. Means of distributions are the same as in the previous section and are equal $(0, 2, \dots, 28)$ and $(1, 3, \dots, 29)$ for class A and B respectively. It should be

noticed that the intervals between the averages of the same features of different classes are equal to 1, and the variance is two times greater than this value.

It is important to understand that generated data is difficult to test. In fact it is not distinguished by human. It can be seen that the linear criterion has better accuracy and can be successfully used in practice. Quadratic criterion can be successfully used for data with small variance.

In most practical tasks, we work with data that is inaccurate. This inaccuracy is called noise. Analyzing it we understand that the actual data may vary. Therefore it is important to check how much noise affects accuracy. For this purpose we will generate a sample with the means $(0, 2, \ldots, 28)$ and $(1, 3, \ldots, 29)$ for classes A and B respectively and variance equal to 1. We add to obtained sample the noise generated from the Gaussian distribution $N(0, 0.2)$.

Analysing accuracy of the tests shown in the Table 1 we are able to argue that the stable test is resistant to noise. Moreover it is seen than both linear and quadratic criteria give good accuracy in this case. For original linearly separated samples the linear, quadratic, stable and modified tests demonstrate the ideal accuracy. For samples with small variance the accuracy of the quadratic test decreases while the accuracy of the linear, stable, and modified tests are quite high. For samples with large variance the accuracy of the quadratic and stable tests decreases. For samples with the Gaussian noise the stable test is the most accurate.

Test	Original			Small variance			Large variance			Gaussian noise		
	A	B	AB	A	B	AB	A	B	AB	A	B	AB
Linear	100	100	100	100	88	94	88	92	90	88	96	92
Quadratic	100	100	100	100	96	98	50	80	50	64	88	76
Stable	100	100	100	100	88	94	68	72	70	72	88	80
Modified	100	100	100	100	100	100	80	84	82	88	100	94

Табл. 1. Accuracy of tests with different variances and noise

8. Conclusions

It has been found that both linear and quadratic criteria give high accuracy for samples with small variance. As the variance increases, the accuracy of the linear criterion remains high, the accuracy of the quadratic criterion decreases. Both criteria are resistant to sample noise. This important fact makes them useful in practical applications. The future scope of the method is connected with the applications in different fields and investigations of more strong variants of relative frequencies comparisons based on confidence intervals.

References

1. Petunin Yu. I., Klyushin D. A., Ganina K. P., Boroday N.V., Andrushkiw R.I. Computer-aided diagnosis of breast cancer. Part 1. Mathematical aspects *Istatistik* 1999. Vol. 2, P. 71–86.

2. Petunin Yu. I., Klyushin D. A., Ganina K. P., Boroday N.V., Andrushkiw R. I. Computer-aided diagnosis of breast cancer. Part 2. Tests and experiments *Istatistik* 1999. Vol. 2., P. 87–105.
3. D. Klyushin and Yu. Petunin "A Nonparametric Test for the Equivalence of Populations Based on a Measure of Proximity of Samples", Ukrainian Math. J., vol. 55, no. 2, pp. 181-198, 2003.
4. S. A. Matveichuk and Yu. I. Petunin, "Generalization of Bernoulli schemes that arise in order statistics", I, Ukrainian Math. J., vol. 42, no. 4, pp. 459–466, 1990.
5. S. A. Matveichuk and Yu. I. Petunin, Generalization of Bernoulli schemes that arise in order statistics", II, Ukrainian Math. J., vol. 43, no. 6, pp. 728–734, 1991.
6. N. Johnson amd S. Kotz, "Some generalizations of Bernoulli and Polya-Eggenberger contagion models", Statist. Paper, vol. 32. p. 1–17, 1991.
7. B.M. Hill, "Posterior distribution of percentiles: Bayes' theorem for sampling from a population", J Am Stat Assoc, vol. 63, pp.677–691, 1968.
8. Lyashko S.I., Klyushin D.A., Alexeyenko V.V. Multivariate ordering using elliptical peeling. *Cybernetics and System Analysis*. 2013. vol. 49,P. 511-–516.

Received: 18.04.2020 / Accepted: 21.05.2020

КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ ВЫБОРОК С ПОМОЩЬЮ ЕЛИПСОВ ПЕТУНИНА

Д. А. Клюшин, Я. В. Штык

Факультет компьютерных наук и кибернетики, Киевский национальный университет имени Тараса Шевченко, Киев, Украина, E-mail: dokmed5@gmail.com

Аннотация. В статье исследуется метод классификации многомерных выборок с помощью эллипсов Петунина. Для тестирования были сгенерированы несколько выборок с разными распределениями. На основании рассчитанной точности критериев были выявлены преимущества и недостатки каждого из линейных и квадратичных критериев и специфика метода в целом. Было обнаружено, что как линейные, так и квадратичные критерии дают высокую точность для выборок с малой дисперсией. С увеличением дисперсии точность линейного критерия остается высокой, точность квадратичного критерия уменьшается. Оба критерия устойчивы к шуму.

Ключевые слова: эллипс Петунина, статистика Петунина, многомерная выборка, линейный дискриминант, квадратичный дискриминант, классификация.