Statistical and Geometrical Approaches to Homogeneity Testing

Dmitriy Klyushin and Kateryna Golubeva

Taras Shevchenko National University of Kyiv, Ukraine, 03680, Kyiv, Akademika Glushkova Avenue 4D

Abstract

Generalizations based on finite training sets are based on the fundamental compactness hypothesis, which states that objects belonging to the same class in the feature space should be close to each other and far from objects of other classes. This concept became generally accepted and was carefully studied. Meanwhile, the hypothesis of compactness is purely geometric in nature and essentially uses the concept of a metric in the feature space. In particular, it generates in a completely natural way the well-known nearest neighbor method, which uses a classifier based on geometric distance. However, this hypothesis ignores the probabilistic nature of the features. For unimodal distributions it works well, but in the general case, this hypothesis may not hold and the generalization becomes incorrect. We propose an alternative approach based on the homogeneity hypothesis. We call homogeneous objects the objects whose features have the same distribution. From a statistical point of view, this means that they belong to the same general population. The use of the universal measure of homogeneity (Petunin's p-statistic) makes it possible to effectively use the apparatus for testing statistical hypotheses about the homogeneity of features for both non-intersecting and largely overlapping samples that do not satisfy the compactness hypothesis, and also to build new variants of statistical featureless discriminant analysis. Instead of metrics in traditional featureless recognition, we propose to use the Petunin's heterogeneity measure. This approach is rigorously substantiated mathematically and has demonstrated high efficiency in practical applications, specifically, in breast cancer screening.

Keywords [1]

Discriminant analysis, relational analysis, Kolmogorov–Smirnov test, Wilcoxon test, compactness hypothesis, homogeneity hypothesis

1. Introduction

The complexity of the pattern recognition is closely related to the compactness hypothesis which allows generalizations to be made based on finite training sets. The meaning of this hypothesis is intuitively clear: similar objects should be close to each other in the feature space, and dissimilar ones are far away. This definition has an obvious geometric character since the concepts of nearness and farness depend on the metric used. In a completely natural way, it generates the simplest classifier that displays a typical pattern of thinking by precedents. This simplest classifier is called the nearest neighbor method and recognizes the tested objects by their proximity to the training objects.

The hypothesis described above has a significant drawback, since it does not take into account the probabilistic nature of the training data. Typically, the data forms samples extracted from some sample space. For correct identification of objects represented by these samples, it is necessary to take into account their random nature. The distance between random samples in the sample space is also random. Therefore, a mechanism is needed that would make it possible to correctly assess the proximity between random samples. The geometric hypothesis of compactness does not work in this case. To solve the problem, we introduce into consideration the concept of homogeneity of objects, which means that these objects belong to the same general population. Since objects are identified with samples of features, their homogeneity should be assessed using criteria for testing statistical hypotheses about homogeneity. The purpose of the chapter is to describe the hypothesis alternative to the compactness hypothesis and propose the modification of the featureless (relational) discriminant analysis using a homogeneity measure instead of a distance. The paper is organized in the following

Information Technology and Implementation (IT&I-2021), December 01-03, 2021, Kyiv, Ukraine
EMAIL: dokmed5@gmail.com (A. 1), golubeva.kateryna@gmail.com (A. 2)
ORCID: 0000-0003-4554-1049 (A. 1), 0000-0003-4307-6470 (A.2)

way. Section 1 contains an introduction and a general description of the problem. Section 2 provides a survey of non-parametric tests for homogeneity and describes a homogeneity measure (p-statistics) that is the main subject of our research. Section 3 describes connection between the p-statistics and featureless (relational) discriminant analysis. Section 4 describes results of numerical comparisons of popular homogeneity measures (the Kolmogorov–Smirnov statistics and the Wilcoxon statistics) with the p-statistics. Here we show how the p-statistics allow reduction of dimensions and arranging features with respect to their significance for recognition. Section 5 contains conclusions and describes possible directions of the work.

2. Two-sample homogeneity measure

Consider samples $x=(x_1,x_2,...,x_n)\in G_1$ and $y=(y_1,y_2,...,y_n)\in G_2$ from populations G_1 and G_2 obeying distribution functions F_1 and F_2 that are absolutely continuous. Let the null hypothesis be $F_1=F_2$ and the alternative hypothesis be $F_1\neq F_2$. The samples drawn from the same populations are called homogeneous. There are many tests for testing the hypotheses on samples` homogeneous: purely nonparametric (Smirnov [1, 2], Dickson [3], Wald and Wolfowitz [4], Mathisen [5], Wilcoxon [6], Mann–Whitney [7], Wilks [8] etc.) and conditionally nonparametric (Pitman [9], Lehmann [10], Rosenblatt [11], Dwass [12], Fisz [13], Barnard [14], Birnbaum [15], Jockel [16], Allen [17], Efron and Tibshirani [18], Dufour and Farhat [19] etc.). Let us consider the Klyushin–Petunin test that is purely non-parametric and does not use any requirements to distribution functions excepting being absolutely continuous [20]. We propose to put this test in the ground of the featureless discriminant analysis. The Hill's assumption $A_{(n)}$ [21] states that if random values $x_1,x_2,...,x_n\in G$ are exchangeable and belong to absolutely continuous distribution then

$$P\left(x\in\left(x_{(i)},x_{(j)}\right)\right)=\frac{j-i}{n+1},\ \ j<i, \tag{1}$$

where x is a sample value from a population G following an absolutely continuous distribution function F, and $x_{(i)}$ and $x_{(j)}$ are the *i*-th and *j*-th order statistics. This assumption was proved for independent identically distributed random values [22] and for exchangeable identically distributed random values [23]. It is a basis of a nonparametric test for samples homogeneity [20]. Thus, computing the relative frequency h_{ij} of the event $y_m\in\left(x_{(i)},x_{(j)}\right)$ for the elements of y, we can estimate a deviation h_{ij} from $\frac{j-i}{n+1}$. To do this we use a confidence interval for binomial proportion. Let for definiteness construct the Wilson confidence interval $I_{ij}^{(n)}=\left(p_{ij}^{(1)},p_{ij}^{(2)}\right)$ where

$$\begin{aligned}p_{ij}^{(1)}&=\frac{h_{ij}n+g^2/2-g\sqrt{h_{ij}(1-h_{ij})n+g^2/4}}{n+g^2},\\ p_{ij}^{(2)}&=\frac{h_{ij}n+g^2/2+g\sqrt{h_{ij}(1-h_{ij})n+g^2/4}}{n+g^2}.\end{aligned} \tag{2}$$

The significance level of this interval depends on the parameter *g*. When *g* = 3 the significance level of $I_{ij}^{(n)}$ does not exceed 0.05 [20]. P-statistics, which is a homogeneity measure of samples *x* and *y*, is defined by the equation

$$h=\frac{2\#\left\{p_{ij}=\frac{j-i}{n+1}\in I_{ij}^{(n)}\right\}}{(n-1)n}. \tag{3}$$

As far as the p-statistics is the relative frequency of the event $\left\{p_{ij}=\frac{j-i}{n+1}\in I_{ij}^{(n)}\right\}$, similar to the above mention case we may construct the Wilson interval I for the p-statistics an use it as a basis of the test: if the upper bound of I is greater than 0.95, the null hypothesis is accepted, else the null hypothesis is rejected.

If the null hypothesis is true, the events $\left\{ p_{ij} = \frac{j-i}{n+1} \in I_{ij}^{(n)} \right\}$ form a generalized Bernoulli scheme [24, 25]. If the alternative hypothesis is true, the these events form a modified Bernoulli scheme. If the null hypothesis can be either true or false, this trial scheme is called Matveichuk–Petunin scheme [26]. If the null hypothesis holds, $\lim\limits_{n\to\infty} \frac{j-i}{n+1} \in (0,1)$, and $\lim\limits_{n\to\infty} \frac{i}{n+1} \in (0,1)$, then the asymptotic significance level β of a sequence of confidence intervals $I_{ij}^{(n)}$ is less than 0.05 [20].

Real samples usually contain rounded numbers and often repeated elements (ties) occur. Thus, we must distinguish a hypothetical sample that is drawn from hypothetical population G containing absolutely precise numbers and an empirical sample drawn from an empirical population $\tilde{G}$ containing rounded measurement. Therefore, we shall have a sample $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ approximating a hypothetical sample $x = (x_1, x_2, ..., x_n)$ Let $x_{(1)} < x_{(2)} < ... < x_{(n)}$ and $\tilde{x}_{(1)} < \tilde{x}_{(2)} < ... < \tilde{x}_{(m)}$ be variational series of hypothetical and empirical samples.

If a number x^* is drawn from G independently from x the Hill assumption (1) holds, hence:

$$p\left(x^* \in \left[x_{(k)}, x_{(k+1)}\right)\right) = \frac{1}{n+1}, \tag{4}$$

where $k = 0,1,...,n$, $x_{(0)} = -\infty$, and $x_{(n+1)} = \infty$ .Then

$$p\left(\tilde{x}^* \in \left[\tilde{x}_{(i)}, \tilde{x}_{(j)}\right)\right) \approx \gamma_i + \gamma_{i+1} + ... + \gamma_{j-1} + \frac{j-i}{n+1}, \tag{5}$$

where $t_l = t\left(\tilde{x}_{(l)}\right)$ is the multiplicity of $\tilde{x}_{(l)}$. If $\tilde{x}$ does not contain ties then $\gamma_i = 0$.

Consider the null hypothesis that hypothetical absolutely continuous distribution functions F_1 и F_2 of hypothetical populations G_1 and G_2 are identical. Suppose, we have empirical samples $\tilde{x} = (\tilde{x}_1, ..., \tilde{x}_n) \in \tilde{G}_1$ and $\tilde{y} = (\tilde{y}_1, ..., \tilde{y}_n) \in \tilde{G}_2$, where $\tilde{G}_1$ and $\tilde{G}_2$ are corresponding empirical populations. Construct the Wilson confidence interval $I_{ij} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ for the probability of the event $\left\{\tilde{y}_k \in \left(\tilde{x}_{(i)}, \tilde{x}_{(j)}\right)\right\}$ using its observed relative frequency. Let us denote $N = \# I_{ij} = \frac{n(n-1)}{2}$ and compute the empirical p-statistics $h = \frac{1}{N} \# \left\{ \frac{j-i}{n+1} \in I_{ij} \right\}$. Then, construct a confidence interval $I = \left(p^{(1)}, p^{(2)}\right)$ for probability $p\left(\left\{ \frac{j-i}{n+1} \in I_{ij} \right\}\right)$ using h. If the upper bound of I is greater than 0.95, the null hypothesis is accepted, else the null hypothesis is rejected.

3. P-statistics and relational discriminant analysis

The relational discriminant analysis is developed in the papers of Petunin et al. [27], Duin and Pekalska [28–35] etc. The main idea of relational discriminant analysis consists in the replacing a presentation of objects in feature space via vectors of feature (hypothesis on vector space) by a proximity (similarity) to some training set using a distance in a metric space. This idea is very productive and valuable, but implicitly it used a concept of geometric proximity space with a metrics. This approach is invalid for random samples. Let us imagine data on a pool of cells (e.g. their areas) measured in microscopic research. A researcher obtains a sample of real values but not an ordered vector. So, to use a distance to measure proximity to training samples is impossible. That is why the concept of homogeneity measure is very useful for such cases. Despite the large number of statistical tests for samples homogeneity only Kolmogorov-Smirnov statistics, U-statistics (Wilcoxon test) and p-statistics have the properties allowing numerical estimating the samples homogeneity (similarity in the sense of belonging to the same population).

For example, we can use the p-statistics to solve the problem of dimension reduction and feature selection. Compute the proximity measure between samples from G_1 and G_2 with respect to two features: *i*th and *j*th. Consider the matrixes of features of *k*th object from G_1 and *l*th object from G_2 (n is the number of features and m is the number of measured values of every feature) :

$$u_k = \begin{pmatrix} x_{11}^{(k)} & x_{12}^{(k)} & \dots & x_{1n}^{(k)} \\ x_{21}^{(k)} & x_{22}^{(k)} & \dots & x_{2n}^{(k)} \\ \dots & \dots & \ddots & \dots \\ x_{m1}^{(k)} & x_{m2}^{(k)} & \dots & x_{mm}^{(k)} \end{pmatrix}, \ v_l = \begin{pmatrix} y_{11}^{(l)} & y_{12}^{(l)} & \dots & y_{1n}^{(l)} \\ y_{21}^{(l)} & y_{22}^{(l)} & \dots & y_{2n}^{(l)} \\ \dots & \dots & \ddots & \dots \\ y_{m1}^{(l)} & y_{m2}^{(l)} & \dots & y_{mn}^{(l)} \end{pmatrix}.$$

Consider the *i*th columns corresponding to *i*th feature from u_k and v_l: $X_i^{(k)} = \left(x_{1i}^{(k)}, x_{2i}^{(k)}, ..., x_{mi}^{(k)}\right)^T$ and $Y_i^{(l)} = \left(y_{1i}^{(l)}, y_{2i}^{(l)}, ..., y_{mi}^{(l)}\right)^T$. Then, compute p-statistics for samples (not vectors!) $X_i^{(k)}$ and $Y_i^{(l)}$ and construct the vector of p-statistics for u_k and v_l with respect to every feature:

$$\mu_{kl}^{(1)} = \rho\left(X_1^{(k)}, Y_1^{(l)}\right),\ \mu_{kl}^{(2)} = \rho\left(X_2^{(k)}, Y_2^{(l)}\right),\ ...,\ \mu_{kl}^{(n)} = \rho\left(X_N^{(k)}, Y_N^{(l)}\right).$$

Then, compute the average p-statistics

$$\nu_k^{(1)} = \frac{1}{N}\sum_{t=1}^{N}\mu_{kt}^{(1)},\ \ \nu_k^{(2)} = \frac{1}{N}\sum_{t=1}^{N}\mu_{kt}^{(2)},\ ...,\ \nu_k^{(n)} = \frac{1}{N}\sum_{t=1}^{N}\mu_{kt}^{(n)}$$

for u_k and an object from G_2 with respect to *i*th feature. This scheme may be applied for comparing u_k with other object from G_1:

$$\overline{\nu}_k^{(1)} = \frac{1}{N-1}\sum_{s=1,s\neq 1}^{N}\mu_{ks}^{(1)},\ \ \overline{\nu}_k^{(2)} = \frac{1}{N-1}\sum_{s=1,s\neq 1}^{N}\mu_{ks}^{(2)},\ ...,\ \overline{\nu}_k^{(n)} = \frac{1}{N-1}\sum_{s=1,s\neq 1}^{N}\mu_{ks}^{(n)}.$$

Pairing p-statistics we form a proximity vector space corresponding to *i*th and *j*th features: $\left(\nu_t^{(i)}, \nu_t^{(j)}\right)$ and $\left(\overline{\nu}_s^{(i)}, \overline{\nu}_s^{(j)}\right)$, $i, j = 1, 2, ..., m; t, s = 1, 2, ..., n.$ Now, in the proximity vector space we have two sets of points consisting of average interclass homogeneity measure and average intraclass homogeneity measure. Thus, we may use in the proximity space any classificator developed for metric spaces. The average intraclass homogeneity measure allows estimating intrinsic diversity of objects in the population, and the average intraclass homogeneity measure allows estimating the feature significance.

4. Experiments and results

To estimate the sensitivity and specificity of the tests we have carried out numerical experiments using samples from normal distribution with various parameters describing the degree of overlapping. We used samples containing 40 random numbers with the same mean and different standard deviations (location shift) and with different means and the same standard deviation (scale shift). We computed the p-statistics with its lower and upper confidence bounds, the Kolmogorov–Smirnov statistics and its p-value and the Wilcoxon statistics and its p-value. The sensitivity of the Klyushin–Petunin test [20] was estimated as the relative frequency of the event when the upper confidence bound for the *p*-statistics is less that 0.95 when distributions are different. The sensitivity of the Kolmogorov–Smirnov test and the Wilcoxon signed-rank test was estimated as the relative frequency of the event when corresponding p-value ≤ 0.05 when distributions were different. The sensitivity of the Klyushin–Petunin test is considered as the relative frequency of the event when the upper confidence bound for the *p*-statistics is greater that 0.95 for identical distributions. The specificity of the Kolmogorov–Smirnov and the Wilcoxon signed-rank test is considered as the relative frequency of the event when p-value ≥ 0.05 when distributions are identical. In this way we have tested two statistical hypotheses: on location shift and on a scale. The null hypothesis on location shift states that the locations of both distributions are the same. The null hypothesis on scale shift states that the variances of both distributions are the same. The alternative hypotheses, in opposite, state that the distribution functions are different. These cases are illustrated at Fig. 1 and Fig. 2. The results are provided in Tables 1–8. Note a remarkable property of the p-statistics. It not only correctly recognizes heterogeneous samples but also demonstrate monotonic decreasing as the location shift increases.

As it was expected, in the case of location shift both the Kolmogorov–Smirnov test and the Wilcoxon signed-rank test work perfectly. The effectiveness of the Kolmogorov–Smirnov test is explained by the increasing discrepancy between the empirical distributions as the location shift increases. The Wilcoxon signed rank test was developed namely for this case and effectively recognizes inversions. However, when we test the scale shift hypothesis the situation changes. Now, the distribution functions are largely overlapped and the discrepancy between them is not very significant. Moreover, the Wilcoxon signed-rank test poorly recognizes the inversions between largely overlapped samples. These statements are justified by the following results (Table 4–6).

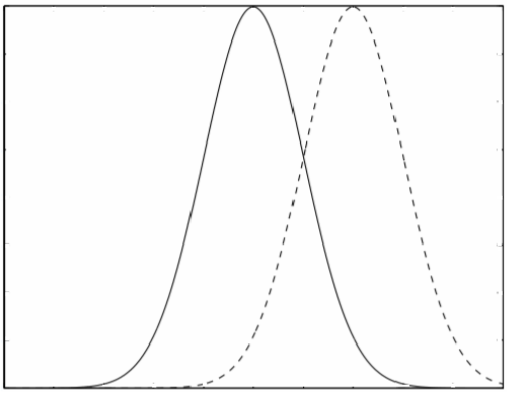

Fig. 1. Probability densities with different means and the same variance (location shift)

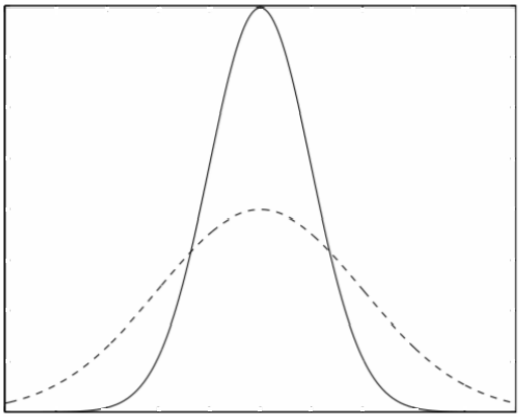

Fig. 2. Probability densities with the same mean and different variance (scale shift)

Table 1

Upper bound of the confidence intervals for the p-statistics (location shift without ties)

Distribution	N(0,1)	N(1,1)	N(2,1)	N(3,1)	N(4,1)
N(0,1)	1.000	0.771	0.590	0.426	0.375
N(1,1)	–	1.000	0.827	0.519	0.375
N(2,1)	–	–	1.000	0.697	0.430
N(3,1)	–	–	–	1.000	0.566
N(4,1)	–	–	–	–	1.000

As we see, the Kolmogorov–Smirnov test has failed in the case of largely overlapped samples in more than almost a half of the cases, and the Wilcoxon signed-rank test has failed in all the cases. The Klyushin–Petunin test fails in almost a third of cases of very overlapped samples following the distributions N(0,3), N(0,4) and N(0,5).

In practice, data are often rounded and ties occur in samples. To simulate this effect we use the same samples as in previous experiments but have rounded them up to two decimal digits. After rounding, samples in average contained four ties. The results are provided in Tables 7–12. As expected from the theoretical point of view, the results of the Kolmogorov–Smirnov test and the

Wilcoxon signed rank test have not changed comparing with the case without ties because the ties do not affect on the Kolmogorov–Smirnov test statistics and the Wilcoxon signed rank test statistics. Thus, we do not provide the results for these cases because they are the same as in Tables 5 and 6.

It is easy to see, that ties have affected on the p-statistics and have slightly changed the monotonic decreasing of the p-statistics with respect to the increasing of the location shift. The Klyushin-Petunin test, as the Kolmogorov–Smirnov test, fails comparing distributions N(0,3), N(0,4) and N(0,5), analogously to the case without ties. But it was effective in the cases when the Kolmogorov–Smirnov test failed and the Wilcoxon signed rank test failed.

Table 2

P-value of the Kolmogorov–Smirnov test (location shift without ties)

Distribution	N(0,1)	N(1,1)	N(2,1)	N(3,1)	N(4,1)
N(0,1)	1.000	0.0002	<0.0001	<0.0001	<0.0001
N(1,1)	–	1.000	0.0002	<0.0001	<0.0001
N(2,1)	–	–	1.000	<0.0001	<0.0001
N(3,1)	–	–	–	1.000	<0.0001
N(4,1)	–	–	–	–	1.000

Table 3

P-value of the Wilcoxon signed-rank test (location shift hypothesis without ties)

Distribution	N(0,1)	N(1,1)	N(2,1)	N(3,1)	N(4,1)
N(0,1)	1.000	0.001	<0.0001	<0.0001	<0.0001
N(1,1)	–	1.000	0.006	<0.0001	<0.0001
N(2,1)	–	–	1.000	<0.0001	<0.0001
N(3,1)	–	–	–	1.000	<0.0001
N(4,1)	–	–	–	–	1.000

Table 4

Upper bound of the confidence intervals for the p-statistics (scale shift hypothesis without ties)

Distribution	N(0,1)	N(0,2)	N(0,3)	N(0,4)	N(0,5)
N(0,1)	1.000	0.741	0.531	0.581	0.570
N(0,2)	–	1.000	0.866	0.767	0.762
N(0,3)	–	–	1.000	0.979	0.964
N(0,4)	–	–	–	1.000	0.999
N(0,5)	–	–	–	–	1.000

Table 5

P-value of the Kolmogorov–Smirnov test (scale shift without ties)

Distribution	N(0,1)	N(0,2)	N(0,3)	N(0,4)	N(0,5)
N(0,1)	1.000	0.014	0.001	<0.0001	<0.0001
N(0,2)	–	1.000	0.029	0.014	0.097
N(0,3)	–	–	1.000	0.766	0.405
N(0,4)	–	–	–	1.000	0.766
N(0,5)	–	–	–	–	1.000

Table 6

P-value of the Wilcoxon signed-rank test (shift hypothesis without ties)

Distribution	N(0,1)	N(0,2)	N(0,3)	N(0,4)	N(0,5)
N(0,1)	1.000	0.202	0.221	0.221	0.900
N(0,2)	–	1.000	0.087	0.158	0.785
N(0,3)	–	–	1.000	1.000	0.314
N(0,4)	–	–	–	1.000	0.795
N(0,5)	–	–	–	–	1.000

Table 7
Upper bound of the confidence intervals for the p-statistics (location shift hypothesis with ties)

Distribution	N(0,1)	N(1,1)	N(2,1)	N(3,1)	N(4,1)
N(0,1)	1.000	0.652	0.511	0.390	0.337
N(1,1)	–	1.000	0.801	0.472	0.355
N(2,1)	–	–	1.000	0.697	0.430
N(3,1)	–	–	–	1.000	0.562
N(4,1)	–	–	–	–	1.000

Table 8
Upper bound of the confidence intervals for the p-statistics (scale shift hypothesis with ties)

Distribution	N(0,1)	N(1,1)	N(2,1)	N(3,1)	N(4,1)
N(0,1)	1.000	0.412	0.323	0.395	0.394
N(1,1)	–	1.000	0.786	0.695	0.644
N(2,1)	–	–	1.000	0.983	0.963
N(3,1)	–	–	–	1.000	0.977
N(4,1)	–	–	–	–	1.000

Thus, we have demonstrated the prevalence of the p-statistics over the Kolmogorov–Smirnov and the Wilcoxon signed rank tests. The Klyushin–Petunin test on homogeneity based on the p-statistics has high sensitivity and specificity both in cases location and scale shifts and in the cases when sample arbitrary overlapped. It is a universal test for test homogeneity of two samples and may be successfully use in application of featureless discriminant analysis as a substitution of a metrics. The p-statistics effectively estimates the homogeneity of the sample and takes values from the interval (0,1). This greatly facilitates its use in the classification of objects defined by samples, in comparison with the Kolmogorov-Smirnov test and Wilcoxon signed rank test.

Since relational discriminant analysis is based on describing the proximity of data, the concept of statistical homogeneity fits perfectly with its concept. Instead of comparing feature vectors representing objects, we can compare samples containing their features. This makes it possible to replace the hypothesis of compactness with the hypothesis of homogeneity and to reduce recognition to assessing the homogeneity of samples. Thus, the p-statistic and the Klyushin-Petunin test can be a valuable tool for pattern recognition in the paradigm of relational discriminant analysis.

5. Conclusions and future work

For correct generalization based on finite training sets, it is necessary to correctly state the fundamental postulates. Relational discriminant analysis is based on the compactness hypothesis, which states that objects belonging to the same class in the feature space must be close, and objects from different classes must be distant. This geometric hypothesis is not correct in the case of assessing the proximity between objects that are characterized not by vectors (ordered sets of features), but by random samples (unordered sets of features), since in such cases it makes no sense to talk about a geometric distance (metric). We propose to base the relational analysis on the hypothesis of homogeneity, which states that objects from the same class (homogeneous) belong to the same general population, that is, the samples of features that characterize them have the same distribution, and objects from different classes belong to different general populations (heterogeneous), that is, the samples of features that characterize them have different distributions. For a numerical assessment of the homogeneity of the samples, we propose to use the Petunin p-statistics, which showed high sensitivity and specificity in experiments both in testing the hypothesis of a mean shift and in testing the hypothesis of the scale shift, in contrast to the statistics of Kolmogorov-Smirnov and Wilcoxon signed rank tests. The proposed approach is strictly mathematically justified and has demonstrated high efficiency in practical applications.

The future direction of the work is to estimate theoretical power of the proposed test and develop its multivariate version. In particular, Petunin's ellipses and ellipsoids, which are constructed on the basis of Hill's assumption, are of great interest. With their help, one can unambiguously order random points in a multidimensional space according to their statistical depth, similar to the Mahalanobis distance, detect outliers, and change points in multidimensional time series. Such tasks often arise in control systems for the timely detection of deviations from the normal operation. In the one-dimensional case, p-statistics, in contrast to the Kolmogorov–Smirnov statistics, are resistant to random noise and universal, in contrast to the Mann–Whitney–Wilcoxon statistics.

Promising methods are the transformation of multidimensional data into one-dimensional samples for subsequent classification. For this, Fisher's linear discriminant analysis transformations can be applied. Similarly, in the multivariate case, one can consider the average p-statistic calculated from one-dimensional samples, followed by an estimate of its average value. Corresponding experiments show the high efficiency of p-statistics in comparison with traditional methods.

The nonparametric approach allows one to effectively solve the problems of classifying one-dimensional and multidimensional data, identify the most and least probable elements of the sample, and rank them using statistical peeling. This is of great importance in medical applications, because due to the unambiguous ranking of multivariate data, it becomes possible to assess the individual risk of a particular patient, and not just the probability of his belonging to a certain group. Such applications have become the focus of a new area of research in artificial intelligence, which is called explainable artificial intelligence. The approach described in the paper is fully consistent with the concept of explained artificial intelligence, since machine learning problems solved using p-statistics allow for an accurate probabilistic interpretation.

The universal and robust nature of p-statistics (its robustness to outliers and independence from the type of hypothesis on mean or variance shift) makes it an indispensable useful tool in statistical studies of data of any size, both small and large. With a small sample size of the strategy, p-statistics work well in combination with the bootstrap and jackknife methods. In such cases, similarly to the multidimensional case, the averaged p-statistics is used, which surpasses traditional analogues in its properties. In the case of large data, the p-statistic can be computationally difficult, but fragmenting large samples and averaging the p-statistic solves this problem as well.

Thus, the presented work proves the high accuracy, sensitivity and specificity of p-statistics, its robustness and superiority over the traditional Kolmogorov–Smirnov and Mann–Whitney–Wilcoxon statistics. Excellent characteristics and a wide field of applications allow us to hope that p-statistics will become a very useful tool for solving many problems of data analysis and machine learning.

6. References

[1] N. V. Smirnov, Estimate of difference between empirical distribution curves in two independent samples. Bulletine of Moscow State University 2 (1939) 3–14

[2] N. V. Smirnov, On the deviations of an empirical distribution curve. Matematicheskii Sbornik 6 (1939) 3–26

[3] W. G. Dixon, A criterion for testing the hypothesis that two samples are from the same population. Annals of Mathematical Statistics 11 (1940) 199–204. doi:10.1214/aoms/1177731914

[4] A. Wald, J. Wolfowitz , On a test whether two samples ate from the same population. Annals of Mathematical Statistics 11 (1940) 147–162. doi:10.1214/aoms/1177731909

[5] H. C. Mathisen, A method of testing the hypothesis that two samples are from the same population. Annals of Mathematical Statistics 14 (1943) 188–194. doi:10.1214/aoms/1177731460

[6] F. Wilcoxon, Individual comparisons by ranking methods,. Biometrika 1 (1945) 80–83. doi:10.2307/3001968

[7] H. B. Mann, D. R. Whitney, On a test of whether one of the random variables is stochastically larger than other. Annals of Mathematical Statistics 18 (1947) 50–60. doi:10.1214/aoms/1177730491

[8] S. S. Wilks, A combinatorial test for the problem of two samples from continuous distributions. In: Proceeding of Fourth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, 1961, pp 707–717

[9] E. J. G. Pitman, Significance tests which may be applied to samples from any populations, Journal of Royal Statistical Society Series A. 4 (1937) 119–130. doi:10.2307/2984124

[10] E. L. Lehmann, Consistency and unbiasedness of certain nonparametric tests, Annals of Mathematical Statistics 22 (1947) 165–179. doi:10.1214/aoms/1177729639

[11] M. Rosenblatt, Limit theorems associated with variants of the von Mises statistic. Annals of Mathematical Statistics 23 (1952) 617–623. doi:10.1214/aoms/1177729341

[12] M. Dwass, Modified randomization tests for nonparametric hypotheses,. Annals of Mathematical Statistics 28 (1957) 181–187, doi:10.1214/aoms/1177729038

[13] M. Fisz, On a result be M. Rosenblatt concerning the Mises–Smirnov test, Annals of Mathematical Statistics 31 (1960) 427–429. doi:10.1214/aoms/1177705905

[14] G. A. Barnard, Comment on "The spectral analysis of point processes" by M.S. Bartlett. Journal of Royal Statistical Society Series B, 25 (1963) 294

[15] Z. W. Birnbaum, Computers and unconventional test-statistics. In: F. Prochan and R. J. Serfling (Eeds). Reliability and Biometry. SIAM, Philadelphia, PA, 1974, pp. 441–458

[16] K.-H. Jockel, Finite sample properties and asymptotic efficiency of Monte Carlo tests. Annals of Statistics 14 (1986) 336–347

[17] D. L. Allen, Hypothesis testing using L1-distance bootstrap. American Statistician 51 (1997) 145–150. doi:10.1080/00031305.1997.10473949

[18] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap. Vol. 57 of Monographs on Statistics and Applied Probability. New York: Chapman-Hall, 1993.

[19] J.-M. Dufour, A. Farhat, Exact nonparametric two-sample homogeneity tests for possibly discrete distributions. Center for Interuniversity research in Quantitative Economics (CIREQ). Preprint 2001-23. California Press, 2001, pp. 707–717

[20] D. A. Klyushin, Yu. I. Petunin, A Nonparametric Test for the Equivalence of Populations Based on a Measure of Proximity of Samples,. Ukrainian Mathematical Journal 55 (2003) 181–198. doi:10.1023/A:1025495727612

[21] B. Hill, Posterior distribution of percentiles: Bayes' theorem for sampling from a population, Journal of the American Statistician Association, 63 1968 677–691. doi:10.1080/01621459.1968.11009286.

[22] I. Madreimov, Yu. I. Petunin, Characterization of a uniform distribution using order statistics. Teoriya Veroyatnostey i Matematicheskaya Statistika, 27 (1982) 96–102.

[23] R. I. Andrushkiw, D. A. Klyushin, Lysyuk V. N., Yu. I. Petunin, Construction of the bulk of general population in the case of exchangeable sample values. In: Proceedings of the International Conference of Mathematics and Engineering Techniques in Medicine and Biological Science (METMBS'03), Las Vegas, Nevada, USA, 2003, pp 486–489

[24] S. Matveichuk, Yu. Petunin, A generalization of the Bernoulli model occurring in order statistics. I., Ukrainian Mathematical Journal 42 1990 459–466. doi:10.1007/BF01071335.

[25] S. Matveichuk, Yu. Petunin, A generalization of the Bernoulli model occurring in order statistics. II., Ukrainian Mathematical Journal 43 (1991) 728–734. doi:10.1007/BF01058940.

[26] N. Johnson, S. Kotz, Some generalizations of Bernoulli and Polya-Eggenberger contagion models, Statistical Papers 32 (1991) 1–17. doi:10.1007/BF02925473.

[27] Yu. I. Petunin., D. A. Klyushin, R. I. Andrushkiw, Nonlinear algorithms of mattern recognition for computer-aided diagnosis of breast cancer. Nonlinear analysis, 30 (1997) 5431–5336

[28] R. P. W. Duin, D. de Ridder, D. M. J. Tax, Experiments with a featureless approach to pattern recognition. Pattern Recognition Letters 18 (1997) 1159–1166. doi: 10.1016/S0167-8655(97)00138-4

[29] R. P. W. Duin, E. Pękalska D. de Ridder, Relational discriminant analysis, Pattern Recognition Letters 20 (1999) 1175–1181. doi:10.1016/S0167-8655(99)00085-9

[30] R. P. W. Duin, M. Loog, E. Pekalska, D. M. J. Tax, Feature-based dissimilarityspace classification. In: D. Unay, Z. Cataltepe, S. Aksoy (Eds.), Recognizing Patterns in Signals, Speech, Images, and Videos, ICPR 2010, vol. 6388, Springer, 2010, pp. 46–55.

[31] E. Pekalska, R. P. W. Duin, On combining dissimilarity representations. In: J. Kittler, F. Roli (Eds.), Multiple Classifier Systems, volume. 2096 of Lecture Notes in Computer Science. Springer-Verlag, 2001, pp. 359–368.

[32] E. Pekalska, R. P. W. Duin. The Dissimilarity Representation for Pattern Recognition. Foundations and Applications. World Scientific, Singapore, 2005.

[33] E. Pękalska, R. P. W. Duin, Dissimilarity-based classification for vectorial representations. In: Proceedings of 18th International conference on pattern recognition, Hong Kong, China, 2006, volume III, pp.: 137–140

[34] E. Pękalska, R. P. W. Duin, Beyond traditional kernels: Classification in two dissimilarity-based representation spaces. IEEE Transactions on Systems, Man, and. Cybernetics, Part C: Applications and Reviews 38 (2008) 729–744. doi: 10.1109/TSMCC.2008.2001687

[35] E. Pękalska, R. P. W. Duin, P. Paclík, Prototype selection for dissimilarity-based classifiers. Pattern Recognition. 39 (2006) 189–208. doi:10.1016/j.patcog.2005.06.012