2015 IEEE International Symposium on Intelligent Control (ISIC)
Part of 2015 IEEE Multi-Conference on Systems and Control
September 21-23, 2015. Sydney, Australia

Writing Style Determination Using the KNN Text Model

Oleg Granichin, Natalia Kizhaeva, Dmitry Shalymov, and Zeev Volkovich

Abstract— **The aim of the paper is writing style investigation. The method used is based on re-sampling approach. We present the text as a series of characters generated by distinct probability sources. A re-sampling procedure is applied in order to simulate samples from the texts. To check if samples are generated from the same population we use a KNN-based two-sample test. The proposed method shows high ability to distinguish variety of different texts.**

Index Terms— **Writing style, authorship attribution, two-sample test, re-sampling.**

I. Introduction

In this paper the problem of writing style determination is studied. Writing style is a set of distinctive words, various grammatical structures, or any other measurable pattern that makes the piece of writing unique [1]. When it needs to determine the author of an anonymous text based on the writing style the methods of authorship attribution (AA) are often used. It is supposed for AA that there is a training set of documents with known authorship.

The problem of AA attracts a lot of attention in the last two decades due to numerous applications. Therefore, the development of new computational methods for AA has become very topical. The methods of control theory can be effectively applied to the creation of new methods for data mining and computational intelligence [2].

The AA techniques find use in such attribution problems as author verification (i.e., to decide whether a given text was written by a certain author) [3], plagiarism detection (i.e., to assess similarity of two texts) [4], author profiling or characterization (i.e., to provide information on the age, education, sex, etc., of the author of a given text) [5] and others.

We address the problem of writing style determination using a comparison of the 'randomness' of two given texts. One of the tools, which is reasonably applicable to this purpose, is the two-sample test methodology intended to check if two given samples are drawn from the same population. The Kolmogorov-Smirnov (KS) test is a classical approach for this case.

O. Granichin, N. Kizhaeva, and D. Shalymov are with the Saint Petersburg State University (Faculty of Mathematics and Mechanics, and Research Laboratory for Analysis and Modeling of Social Processes), 198504, Universitetskii pr. 28, St. Petersburg, Russia, O. Granichin is also with the Institute of Problems in Mechanical Engineering, Russian Academy of Sciences, Z. Volkovich is with the Dept. Soft. Engn., The Ort Braude, Karmiel, Israel, oleg_granichin@mail.ru, natalia.kizhaeva@gmail.com, shalydim@mail.com, zeev53@mail.ru

This work was financially supported by the SPbSU (project 15.61.2219.2013 and 6.37.181.2014) and, in part, by the RFBR (grants 13-07-00250 and 15-08-02640).

The normal distribution cannot be confidently established as the limit one here because a text written by one of more co-authors hardly appears to be generated by a single random source. To stabilize the process we apply the following multistage approach. At the first step, we evaluate the null hypothesis distribution, assuming concurrency of the considered writing styles, by comparing samples drawn from the text. Further, samples drawn separately from different texts are matched in order to get the appropriate p-values calculated with respect to the constructed null hypothesis distribution. In the case of the identical writing styles these p-values are uniformly distributed on the interval $[0, 1]$. We compare the obtained p-values distribution with the uniform one by means of a univariate two-sample KS-test.

The article is organized as follows. Section II reviews related work. In section III an overview of Two-Sample Test methodology is provided. Section IV presents the sampling and comparing algorithms. The results of numerical experiments are shown in Section V, followed by Conclusions and Future work discussion.

II. Related work

One of the most influential works in this field is the work of Mosteller and Wallace (1964) [6] where the authorship of 'The Federalist Papers' (a collection of 85 articles and essays written by Alexander Hamilton, James Madison, and John Jay) was assigned based on Bayesian statistical analysis of common word frequencies. It opened up the field to the exploration of new types of stylometric features and new modeling techniques [1].

Since then, various textual measurements were proposed. The simplest ones come from common descriptive statistics. Average word lengths, relative frequency, number of words per sentence, distribution of parts of speech can be easily calculated. They are obtainable for any natural language and corpus (if a proper tokenizer is available), and are useful for evaluating the writing style. However, none of these features robustly separates different authors in a large number of cases [9].

The next level of measures are character features, according to which, a text is viewed as a sequence of characters [1]. The measures include alphabetic and digit characters count, uppercase and lowercase characters count, letter frequencies, punctuation marks count, and so on [10]. More sophisticated approach is to explore frequencies of character n-grams. Among the advantages of this approach are the ability to capture context, use of punctuation, tolerance to grammatical errors.

978-1-4799-7798-7/15/$31.00 ©2015 IEEE

Machine learning techniques made a considerable impact to AA studies. From a machine learning viewpoint, the task of AA can be regarded as a multiclass, single-label text categorization problem when it needs to determine the author of an anonymous text based on a training set of documents with known authorship [7]. Although various learning algorithms can be applied to this task, the performance significantly depends on the choice of features.

III. Two-Sample Test Methodology

Two-sample hypothesis testing is a statistical analysis approach developed to examine if two samples of independent random variables in the Euclidean space R^d have the same probability distribution function. In mathematical notation, let $X = X_1, X_2, .., X_m$ and $Y = Y_1, Y_2, .., Y_n$ be two independent random variables with distribution functions F and G that are unknown. A two-sample problem consists in testing the null hypothesis

$H_0 : F(x) = G(x)$

against the alternative

$H_1 : F(x) \neq G(x)$.

Kolmogorov–Smirnov test [12], [13] is common and general nonparametric method for testing the equality of continuous one-dimensional probability distributions. The Kolmogorov–Smirnov statistic

$$D = \sup_x |\tilde{F}(x) - \tilde{G}(x)|$$

measures the distance between empirical distribution functions $\tilde{F}(x)$ and $\tilde{G}(x)$ of two samples. As the test is asymptotically distribution-free, the test statistic distribution is independent of the underlying distributions of the data for sufficiently large samples. The test is applicable to comparison of a sample and a reference probability distribution (so called one-sample KS-test). In this instance, the KS-statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution.

Numerous tests have been designed for multivariate case. A survey of nonparametric tests and a comparative study are presented in [14] and [15]. Multivariate generalization of Smirnov test is given in [16]. The two-sample energy test [17] is also successful in multidimensional applications. An exact distribution-free test is introduced in [20]. New statistics are also proposed in [18], [19]. A kernel method to comparing distributions, which is introduced in [21] and further developed in [22], [23], is a notable approach with various applications. Related to this methodology, a kernel method for the two-sample problem was independently proposed earlier in [24]. The test uses a characterization theorem stated in [25]. Applications of this approach are shown in [26] (see, also [27]). The above-mentioned energy test can also be interpreted in the framework of this methodology.

A two-sample test statistic is intended to describe mingling quality of items belonging to two disjoint i.i.d. samples S_1 and S_2 . We can measure the mixture merit by means of K-nearest neighbors fractions of the samples quantified at each point. Obviously, these proportions are approximately equal if the samples are well mixed. Cluster validation has been considered from this point of view in the paper [28]. K-nearest neighbors type coincidences model in the current paper deals with the statistic:

$$T_K\,(S_1 \cup S_2) = \sum_{x \in S_1 \cup S_2} \sum_{r=1}^{K} I(\ \begin{matrix} x \text{ and } r\text{-th neighbor} \\ \text{belong to the same sample} \end{matrix}\ ) \quad (1)$$

which represents the number of all K nearest neighbors type of coincidences. Asymptotic behavior of this statistic has been studied in [29]. In fact, the asymptotic normal distribution can barely be applied in the comparison of two real texts owing to the inherent heterogeneity. The null hypothesis law still can be simulated in the spirit of the bootstrap methodology (see, e.g. [34]). Construction of an empirical distribution of the pooled samples indirectly imply their identical underlying distributions under the null hypothesis. At the same time, when these distributions are actually different, the above procedure (using just 'prior mixing') can produce a distorted distribution. Due to this reason we precise the inference process by means of the procedure described below.

IV. Method

To implement our approach we transform the considered texts into two binary files, F_1 , F_2 and introduce $F_0 = F_1 \cup F_2$. Our purpose is to distinguish between the distributions of these files using a re-sampling procedure that is an essential part of the method reflecting the sources structure. We form samples by means of N-grams as connecting sequences of N symbols from a text as follows

Algorithm 1 Sampling procedure

Input parameters:

- F – text file;
- N – attribute (N-gram) size;
- $NWORD$ – number of attributes in a vector (vector dimension);
- $NVEC$ – number of vectors in a sample(sample size).

repeat $NVEC$ times

1) Generate a random number as the starting position for a vector in a file;
2) From this starting position, the sequential set of $NWORD$ attributes is treated as a vector in R^{NWORD} space.

As was mentioned above the normal distribution rarely appears as the null hypothesis distribution in the considered problem. Thus, this probability law is evaluated using the bootstrapping methodology by repeatedly drawing pairs of samples without replacement from F_0. And the values of the T_K test statistic (1) are calculated. At the next step the p-value is evaluated for each statistic value with respect to the null hypothesis distribution obtained in the previous

step. If the null hypothesis is correct then the files cannot be distinguished, this distribution is the uniform one on interval $[0, 1]$. We test such a hypothesis again by means of a one-variate two sample test and consider each one of these assessments as a Bernoulli trial. According to our perception two texts are different by their inner style if the fraction of the rejections in a Binomial sequence of these trials is significantly bigger than 0.5.

Algorithm 2 Main algorithm

Input parameters:

- F_1, F_2 – files being compared;
- $ITER$ – number of the process iterations;
- N – attribute (N-gram) size;
- $NWORD$ – number of attributes in a vector (vector dimension);
- $NVEC$ – number of vectors in a sample (sample size);
- $NPER$ – number of the random permutations in the re-sampling procedure;
- K – KNN quantity;
- TR_KS – prob. threshold below which the null hypothesis in the one-sample KS-test is rejected;
- TR – prob. threshold below which the null hypothesis of the equal styles of F_1 and F_2 is rejected.

1: Introduce $F_0 = F_1 \cup F_2$
2: **for** $iter = 1 : ITER$ **do**
3: **for** $perm = 1 : NPER$ **do**
4: $F = random\_permutation(F_0)$;
5: $S_1 = Sample(N, NWORD, NVEC, F)$;
6: $S_2 = Sample(N, NWORD, NVEC, F)$;
7: Calculate $V_{perm} = \{T_K(S_1 \cup S_2)\}$;
8: **end for**
9: Construct an empirical P_0 distribution of $\{V_{perm}, perm = 1 : NPER\}$.
10: **for** $perm = 1 : NPER$ **do**
11: $S_1 = Sample(NA, NWORD, NVEC, F_1)$;
12: $S_2 = Sample(NA, NWORD, NVEC, F_2)$;
13: Calculate: $U_{perm} = \{T_K(S_1 \cup S_2)\}$;
14: **end for**
15: Calculate $NPER$ p-values: $PV = \{pval_{perm}, perm = 1 : NPER\}$ of $\{U_{perm}, perm = 1 : NPER\}$ with respect to P_0;
16: Use the one-sample KS-test to compare PV with the uniform distribution on $[0, 1]$ and obtain $h_{iter} = 1$ if the null hypothesis is rejected and $h_{iter} = 0$ otherwise;
17: **end for**
18: Test the hypothesis that the fraction of the rejections in the sequence $H = \{h_{iter}, iter = 1 : ITER\}$ is smaller than TR. If this null hypothesis is rejected then the styles of F_1 and F_2 are accepted as different.

Comments Regarding the Algorithm

1) Empirical p-values in the line 15 of the algorithm are calculated according to the formula:
$$PV(U_i) = \frac{\sum_{perm=1}^{NPER} I(V_{perm} > U_i)}{NPER},\ i = 1 : NPER.$$
2) The null hypothesis is rejected in the line 16 if the p-value provided by the one-sample KS-test is smaller than TR_KS.
3) We use the one-sample z-test to determine whether the hypothesized proportion of the rejections in the sequence H is significantly bigger than 0.5. For this aim the following p-value is calculated:
$$pp = 1 - \Phi\left(\frac{\widehat{P} - 0.5}{\sqrt{\left(\widehat{P}\left(1 - \widehat{P}\right)\right)}}\right), \qquad (2)$$
where Φ is the cumulative function of the standard normal distribution, and
$$\widehat{P} = \frac{sum\{H\}}{NPER}.$$
The null hypothesis is rejected if $pp < TR$.

V. Numerical Experiments

We provide several experiments in order to demonstrate the capability of the proposed method.

A. Three Collections of English Texts

The text preprocessing includes omission of all spaces in the files. All comparisons were provided with parameters set as $Iter = 30$, $N = 32bit$, $NWORD = 32$, $NVEC = 64$, $NPER = 50$, $K = 10$ and $TR = TR\_KS = 0.05$.

The first file is denoted by HP (having size of 3,422,603 B). It is composed from the first five books of the Harry Potter of J. K. Rowling series:

- Harry Potter and the Sorcerer's Stone;
- Harry Potter and the Chamber of Secrets;
- Harry Potter and the Prisoner of Azkaban;
- Harry Potter and the Goblet of Fire;
- Harry Potter and the Order of the Phoenix.

The second file is denoted by F (having size of 1,234,583 B) consists of four books of the A. Azimov Foundation series:

- The Story behind the Foundation;
- Forward the Foundation;
- Foundation;
- Foundation and Empire.

The last one (denoted as AC) with the size 2,139,414 B contains 7 books of A. Clarke:

- 2010: Odyssey two;
- 2001: A space Odyssey;
- A Fall of Moondust;
- Against The Fall of Night;
- Expedition to Earth;

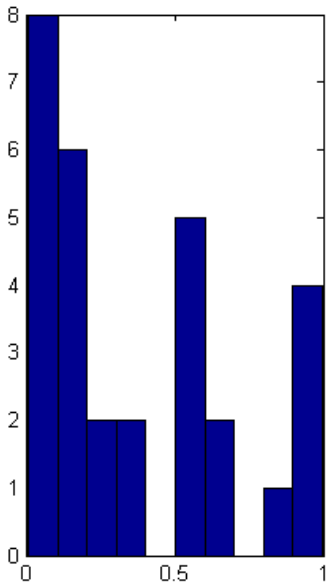

Fig. 1: Histograms of p-values in comparison of the HP collection with itself.

- Space Trilogy 03;
- The Wind from the Sun.

Initially, these collections are compared with each other. The values of pp are presented in Table V.1. Here and in all future tables the sources used for the null hypothesis generation (designated previously as F_1) are placed in the first column. The styles of two files are assumed to be different when the null hypothesis is rejected, i.e. $pp < TR$.

TABLE V.1: Comparison of the three text collections

	HP	F	AC
HP	0.99	0	0
F	0	0.97	0
AC	0	0	1

As we see, our method succeeds to recognize dissimilar files together with the identical styles for all collections.

An example of histogram for asymptotic p-values returned by KS-test is presented in Fig 1. The histogram is built based on bins of a uniform width, chosen to cover the range of p-values. Note that the length of p-values corresponds to the number of iterations $ITER$. The height of each rectangle indicates the number of elements in the bin.

After that we turn to evaluate the so named 'false positive' outcome of the proposed method, when texts written by the same author are recognized as ones possessing different styles. We compare the texts of the HP collection among themselves and the texts of the F collection among themselves. The results presented in Tables V.2 and V.3 reveal that the null hypothesis was 'incorrectly' not rejected in 13 cases (marked in bold) within 41. This fact confirms the reliability of our method, taking into account a sufficiently big variety of the files sizes and matters.

TABLE V.2: Texts comparison from the HP collection

	1	2	3	4	5
1	0.99	**0.99**	0	0	0
2	**0.99**	0.99	0	0	0
3	0	0	0.96	**0.99**	**0.99**
4	0	0	**0.99**	0.99	**0.99**
5	0	0	**0.99**	**0.99**	0.99

TABLE V.3: Texts comparison from the F collection

	1	2	3	4
1	0.99	0	**0.99**	**0.95**
2	0	0.99	0	0
3	**0.96**	0	0.99	**0.99**
4	$1.4 * 10^{-3}$	0	**0.99**	0.99

B. The Chronicles of Narnia

A peculiar result was obtained under comparison of The Chronicles of Narnia by C. S. Lewis. The books enumeration in the Table V.4 corresponds to Harper Collins chronological order:

- The Magician's Nephew
- The Lion, the Witch and the Wardrobe
- The Horse and His Boy
- Prince Caspian
- The Voyage of the Dawn Treader
- The Silver Chair
- The Last Battle

The Chronicles of Narnia narrates the adventures of children in fantasy world Narnia. All books remain fairy tales directed mostly to young readers, whereas Harry Potter series change through time, the main characters grow up, causing the style to differ among the books.

The comparisons were provided with the following values of parameters: $Iter = 50$, $N = 32bit$, $NWORD = 32$, $NVEC = 64$, $NPER = 50$, $K = 10$ and $TR = TR\_KS = 0.05$.

TABLE V.4: Texts comparison from the Chronicles of Narnia

	1	2	3	4	5	6	7
1	1	1	0.99	0.99	0.99	0.99	0.99
2	1	0.99	1	0.99	0.99	0.97	1
3	1	0.99	0.99	0.99	0.99	1	0.99
4	1	0.99	1	1	0.99	0.99	1
5	0.80	0.98	0.99	0.99	1	1	1
6	0.99	0.92	0.99	1	0.99	1	1
7	0.97	0.98	0.99	0.99	0.99	0.99	1

C. Twilight and Fifty Shades of Grey series

The Twilight series by S. Meyer consists of 4 books:

- Twilight
- New Moon
- Eclipse
- Breaking Dawn

The Fifty Shades trilogy by E. L. James was originally developed from Twilight fan fiction. The algorithm captures

this relation as can be seen in Table V.5. The books listed in the first column are:

- Fifty Shades of Grey
- Fifty Shades Darker
- Fifty Shades Freed

The values for parameters in this experiment: $Iter = 30$, $N = 32bit$, $NWORD = 8$, $NVEC = 32$, $NPER = 50$, $K = 10$ and $TR = TR\_KS = 0.05$.

TABLE V.5: Comparison of the Fifty Shades of Grey trilogy and Twilight series

	1	2	3	4
1	**0.07**	0	0	0
2	0	**0.99**	**0**	0.03
3	0	**0.13**	**0.99**	0.05

Denote The Chronicles of Narnia as CN, Twilight series as Tw and Fifty Shades trilogy as 50. HP refers to Harry Potter collection. The false-positive outcome appears only once demonstrating the high correctness of the results.

TABLE V.6: Comparison of the four text collections

	CN	Tw	50Sh	HP
CN	1	0	0.03	0
Tw	0	1	0	0
50Sh	**0.98**	0	1	0
HP	0	0	0	1

D. William Shakespeare's Plays

We also analyzed a set of works of William Shakespeare. There are two main options for the separation of Shakespeare's plays: for three and for four periods [41].

Three periods of Shakespeare's plays are: I (optimistic) period (1590-1600), II (tragic) period (1601-1607) and III (romantic) period (1608-1612).

Respective text collections for these periods are compared one with another. The corresponding values for p_1 are presented in Table V.7.

The First (I) collection has size $2,085,017$ bytes, the Second (II) has size $1,527,055$ bytes ans the Third (III) has size $542,682$ bytes.

All comparisons were provided with the following values of parameters: $Iter = 50$, $N = 32bit$, $NWORD = 32$, $NVEC = 64$, $NPER = 50$, $K = 10$ and $TR = TR\_KS = 0.05$.

TABLE V.7: Comparison of the three periods of Shakespeare's plays

	I	II	III
I	1	**0.99**	$2 * 10^{-3}$
II	**1**	0.98	0
III	0	0	0.99

The null hypothesis was 'incorrectly' not rejected in 2 cases. Which means that accuracy is about 78%.

There exists a separation of Shakespeare's plays into four periods: I (optimistic, sanguine) period (1590-1594), II (more realism, tough-minded) period (1595-1601), III (disappointed) period (1601-1608) and IV (romantic) period (1608-1612). The result of text collections comparison for these four periods is represented in Table V.8.

The First (I) collection has size $866,834$ bytes, the Second (II) has size $1,419,412$ bytes the Third (III) has size $1,325,826$ bytes. and the Forth (IV) has size $542,682$ bytes.

TABLE V.8: Comparison of the four periods of Shakespeare's plays

	I	II	III	IV
I	0.99	0	**1**	0
II	0	0.99	0	**1**
III	**1**	0	1	0
IV	0	0	0	1

The null hypothesis was 'incorrectly' not rejected in 3 cases. The accuracy is about 82% which is better than for the three periods. It corresponds to the fact that four periods of Shakespeare's plays are more widely recognized [41].

E. The Epistels Collection

We have compared six text collections of Epistels from The New Testament that discusses the teachings and person of Jesus. The New Testament is an anthology, a collection of Christian works written in the common Greek language of the first century, at different times by various writers, who were early Jewish disciples of Jesus. Nevertheless it is believed that the authorship of the Epistels belongs to the same author Paul the Apostol.

We used Epistels Corinthians 1 (having size of 42,029 B), Corinthians 2 (having size of 28,228 B), Galatians (having size of 13,953 B), Philippians (having size of 10,285 B), Romans (having size of 43,660 B), and Thessalonians 1 (having size of 9,439 B).

All comparisons were provided with the following values of parameters: $Iter = 30$, $N = 32bit$, $NWORD = 32$, $NVEC = 64$, $NPER = 50$, $K = 10$ and $TR = TR\_KS = 0.05$.

The result of comparison can be seen in Table V.9.

TABLE V.9: Comparison of six Epistels of The New Testament

	Cor 1	Cor 2	Gal	Phil	Rom	Thess 1
Corinth 1	1	**1**	0	0	0	**1**
Corinth 2	**1**	1	0	0	0	**1**
Galat	0	0	1	**1**	**1**	0
Philipp	0	0	**1**	1	**1**	0
Romans	0	0	**1**	**1**	1	0
Thessal 1	**1**	**1**	0	0	0	1

In this experiment the null hypothesis was not rejected 12 times for 36 comparisons (marked as bold items). According to the rejected cases the Epistels mentioned above can be separated in two groups of styles. Group I includes Corinthians 1, Corinthians 2 and Thessalonians 1 Epistels. Group II includes Galatians, Philippians and Romans Epistels.

VI. Conclusion

We offered a new re-sampling method designed to distinguish between texts possessing different writing styles. The method is based on comparison of empirical distributions constructed for the two-sample KS-test statistic for samples drawn from the same source and different ones. The provided numerical experiments show a high capability of the proposed method.

For further investigations the analysis of texts in non-european languages (e.g. Arabic, Hebrew) seems to be perspective. Also the comparison of proposed method with well-known approaches such as Burrow's Delta [35]–[37], Compression models [38], ANOVA [39], Latent Dirichlet allocation [40] etc. is planned to be performed.

References

[1] E. Stamatatos, "A survey of modern authorship attribution methods", *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[2] O. Granichin, Z. V. Volkovich, and D. Toledano-Kitai, *Randomized Algorithms in Automatic Control and Data Mining*, Springer, 2015.

[3] M. Koppel, and J. Schler, "Authorship verification as a one-class classification problem", *Proc. of the 21st International Conference on Machine Learning*, NewYork: ACM Press, p. 62, 2004.

[4] S. Meyer zu Eissen, B. Stein, and M. Kulig, "Plagiarism detection without reference collections", *Advances in Data Analysis*, Berlin, Germany: Springer, pp. 359–366, 2007.

[5] M. Koppel., S. Argamon, A.R. Shimoni, "Automatically categorizing written texts by author gender", *Literary and Linguistic Computing*, vol. 17 no. 4, pp. 401–412, 2002.

[6] F. Mosteller, D.L. Wallace, "Inference in an authorship problem - a comparative-study of discrimination methods applied to authorship of disputed Federalist Papers", *Journal of the American Statistical Association*, vol. 58(302), p. 275, 1963.

[7] F. Sebastiani, "Machine learning in automated text categorization", *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[8] J. Rudman, "The state of authorship attribution studies: Some problems and solutions", *Computers and the Humanities*, vol. 31, pp. 351–365, 1998.

[9] P. Juola, "Authorship attribution", *Foundations and Trends in Information Retrieval*, vol. 1, no. 3, pp. 233–334, 2006.

[10] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques", *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006.

[11] J. Grieve, "Quantitative authorship attribution: An evaluation of techniques", *Literary and Linguistic Computing*, vol. 22 no. 3, pp. 251–270, 2007.

[12] A. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione", *G. Ist. Ital. Attuari*, vol. 4, 1933.

[13] N. Smirnov, "Table for estimating the goodness of fit of empirical distributions", *Annals of Mathematical Statistics*, vol. 19, 1948.

[14] B.S. Duran, "A survey of nonparametric tests for scale", *Communications in statistics - Theory and Methods*, vol. 5, pp. 1287–1312, 1976.

[15] W.J. Conover, M.E. Johnson, and M.M. Johnson, "Comparative study of tests of homogeneity of variances, with applications to the outer continental shelf bidding data", *Technometrics*, vol.23, pp. 351–361, 1981.

[16] J.H. Friedman and L.C. Rafsky, "Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests", *Annals of Statistics*, vol.7, pp. 697–717, 1979.

[17] G. Zech and B. Aslan, "New test for the multivariate two-sample problem based on the concept of minimum energy", *The Journal of Statistical Computation and Simulation*, vol.75, no. 2, pp. 109–119, 2005.

[18] L. Baringhaus and C. Franz, "On a new multivariate two-sample test", *Journal of Multivariate Analysis*, vol. 88, no. 1, pp. 190–206, 2004.

[19] P. Hall and N. Tajvidi, "Permutation tests for equality of distributions in high-dimensional settings", *Biometrika*, vol. 89, no. 2, pp. 359–374, 2002.

[20] P. Rosenbaum, "An exact distribution-free test comparing two multivariate distributions based on adjacency", *Journal of the Royal Statistical Society B*, vol.67, no. 4, pp. 515–530, 2005.

[21] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf and A. Smola, "A kernel method for the two-sample-problem", *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, vol.19, pp. 513–520, 2007.

[22] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf and A. Smola, "A kernel approach to comparing distributions", *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI-07)*, pp. 1637–1641, 2007.

[23] A. Gretton and K.M. Borgwardt and M.J. Rasch and B. Schölkopf and A. Smola, "A Kernel Method for the Two-Sample Problem", *Jornal of Machine Learning Research*, vol.4, pp. 1–10, 2008.

[24] L. Klebanov, "One class of distribution free multivariate tests", *SPb. Math. Society*, Preprint, vol.2003, no. 3, 2003.

[25] A.A. Zinger, A.V. Kakosyan and L.B. Klebanov, "Characterization of distributions by means of the mean values of statistics in connection with some probability metrics", *Stability Problems for Stochastic Models*, VNIISI, pp. 47–55, 1989.

[26] L. Klebanov, T. Kozubowskii, S. Rachev, and V. Volkovich, "Characterization of Distributions Symmetric With Respect to a Group of Transformations and Testing of Corresponding Statistical Hypothesis", *Statistics and Probability Letters*, vol.53, pp. 241–247, 2001.

[27] L. Klebanov, "N-distances and their Applications", Charles University in Prague, The Karolinum Press, 2005.

[28] Z. Volkovich, Z. Barzily, R. Avros. and D. Toledano-Kitay, "On application of the K-nearest neighbors approach for cluster validation", *Proceeding of the XIII International Conference Applied Stochastic Models and Data Analysis (ASMDA 2009)*, 2009.

[29] N. Henze, "A multivariate two-sample test based on the number of nearest neighbor type coincidences", *Annals of Statistics*, vol.16, pp. 772–783, 1988.

[30] R. T. Shackletona, D. C. Le Maitrea, D. M. Richardsona, "Prosopis invasions in South Africa: Population structures and impacts on native tree population stability", *Journal of Arid Environments*, vol. 114, 2015.

[31] P. Farmer, H. Bonnefoi, V. Becette et al.,"Identification of molecular apocrine breast tumours by microarray analysis", *Oncogene*, vol. 24, 2005.

[32] M. L. Goldstein, S. A. Morris, G. G. Yen, "Problems with fitting to the power-law distribution", *The European Physical Journal B*, vol.41, no. 2, 2004.

[33] K. Sharifi, A. Leon-Garcia, "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video", *IEEE Trans. on Circuits and Systems for Video Technology*, vol.5, no. 1, 1995.

[34] B. Efron, R. Tibshirani, "An Introduction to the Bootstrap". Boca Raton, FL: Chapman & Hall/CRC, 1993.

[35] J. F. Burrows, "Delta: A measure of stylistic di erence and a guide to likely authorship". *Literary and Linguistic Computing*, vol. 17, pp. 267–287, 2002.

[36] D. L. Hoover. "Testing burrows's delta". *Literary and Linguistic Comput- ing*, vol. 19, no. 4, pp 453–475, 2004.

[37] S. Stein and S. Argamon, "A mathematical explanation of burrows' delta", *In Proceedings of Digital Humanities 2006*, Paris, France, 2006.

[38] W. Oliveira Jr., E. Justino, L.S. Oliveira, "Comparing compression models for authorship attribution", *Forensic Science International*, vol. 228, pp. 100–104, 2013.

[39] D.I. Holmes, R. Forsyth, "The Federalist revisited: New directions in authorship attribution", *Literary and Linguistic Computing*, vol. 10, no.2, pp. 111–127, 1995.

[40] J. Savoy, "Authorship attribution based on a probabilistic topic model", *Information Processing and Management*, vol. 49, pp. 341–354, 2013.

[41] Four Periods of Shakespeare's Life, http://www.shakespeare-online.com/biography/fourperiods.html