Studying of KNN Two-Sample Test Approach Applications for Writing Style Comparison of English and Russian Text Collections

Natalia Kizhaeva, Dmitry Shalymov, Oleg Granichin
Saint Petersburg State University
Saint Petersburg, Russia
oleg_granichin@mail.ru, natalia.kizhaeva@gmail.com,
shalydim@mail.com

Zeev Volkovich
ORT Braude College
Karmiel, Israel
vlvolkov@braude.ac.il

***Abstract*—The paper deals with the writing style determination problem. The method designed is based on re-sampling approach and its performance depending on the parameter values is studied. A text is represented as a sequence of characters generated by distinct probability sources. A re-sampling procedure is applied in order to obtain samples from the texts. To check if samples are drawn from the same population a KNN-based two-sample test is used. Numerical examples if different languages are provided, showing the high potential of the method.**

I. Introduction

In this paper the problem of writing style determination is studied. Writing style refers to a manner in which the author writes. It can be expressed as a set of distinctive words, grammatical structures, or any other quatifiable characteristic that makes the piece of writing unique [1]. The problem of determination the author of an unassigned text based on the writing style is a part of more broad range of tasks called authorship attribution (AA).

Authorship analysis becomes a promising area of research due to the increasing amount of real-world electronic texts, like blogs, posts in social networks, emails etc. Various applications include criminal and civil law (digital evidence investigations, copyright disputes [7]), computer security (authorship identification of source code) as well as literary studies (attributing works of unknown or disputed authorship). The development of new computational methods for AA remains topical. The techniques of control theory can be effectively applied to the creation of new methods for data mining and computational intelligence [2].

The AA techniques cope with 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 examine the problem of writing style determination by a comparison of the 'randomness' of two given texts. One of the tools, applicable to this purpose, is the two-sample test 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.

The normal distribution cannot be confidently set as the limit one in this problem because a text written by co-authors does not appear to be generated by a single random source. To stabilize the generative process we apply the following multistage procedure. First, we evaluate the null hypothesis distribution, assuming coincidence of the considered writing styles, by comparing samples drawn from the text. Then, samples drawn separately from different texts are mixed 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.

We continue the reseach started in the paper [6] and study the effectiveness of the algorithm varying the parameter values. The method was applied to Russian texts for the first time.

The article is organized as follows. Section II-A gives an overview of Two-Sample Test methodology. Section II-B presents the sampling and comparing algorithms. The results of numerical experiments are shown in Section III, followed by Conclusions and Future work discussion.

II. Method

A. 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 $\mathbf{R^d}$ 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

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

ISBN 978-952-68397-0-7 (paperback), ISBN 978-952-68397-1-4 (PDF)

Kolmogorov–Smirnov test [8], [9] 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 [10] and [11]. Multivariate generalization of Smirnov test is given in [12]. The two-sample energy test [13] is also successful in multidimensional applications.

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 [14].

Let $|\cdot|$ denote a fixed but arbitrary norm in R^d and set

$$Z_i = \begin{cases} X_i & 1 \le i \le m, \\ Y_{i-m} & m+1 \le i \le l, \end{cases}$$

where $l = m + n$ is the total sample size. The r-th nearest neighbor to Z_i is that point Z_j satisfying $|Z_\nu - Z_i| < |Z_j - Z_i|$ for exactly $r - 1$ values of ν, $1 < \nu < l$, $\nu \neq i, j$.

K-nearest neighbors type coincidences model in the current paper deals with the statistic:

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

which represents the number of all K nearest neighbors type of coincidences.

Asymptotic behavior of this statistic has been studied in [15]. 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. [16]). 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.

B. Algorithm

To implement our approach we transform the considered texts into two binary files, F_1, F_2 and set $F_0 = F_1 \cup F_2$. We aim at distinction between the distributions of these files using a re-sampling procedure that is an essential part of the method reflecting the sources structure. Samples are formed by means of N-grams as connecting sequences of N symbols from a text as shown in Algorithm 1.

Algorithm 1 Sampling procedure

Require:

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

repeat num_vector times

1) Generate a random number – the starting position for a vector in a file;
2) Construct a vector of num_attr sequential attributes.

As was mentioned above the normal distribution rarely appears to be the null hypothesis distribution in the considered problem. So the probability law is evaluated using the bootstrapping methodology by repeatedly drawing pairs of samples without replacement from F_0. Then the values of the T_K test statistic 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.

Comments Regarding the Algorithm

1) Empirical p-values in the line 15 of the Algorithm 2 are calculated as follows:
$$PV(U_i) = \frac{\sum_{perm=1}^{num\_perm} I(V_{perm} > U_i)}{num\_perm},$$
where $i = 1 : num\_perm$.
2) The null hypothesis is rejected in the line 16 if the p-value provided by the one-sample KS-test is smaller than $threshold_{KS}$.
3) In the line 18 we use the one-sample z-test to determine whether the hypothesized proportion of the rejections in the sequence $\{h_{iter}\}$ is significantly bigger than 0.5. For this aim the following p-value is calculated:
$$pp = 1 - \Phi\left(\frac{\widehat{P} - 0.5}{\sqrt{\left(\widehat{P}\left(1 - \widehat{P}\right)\right)}}\right), \qquad (1)$$

Algorithm 2 Main algorithm

Let $F_0 = F_1 \cup F_2$
for $iter = 1 : num\_iter$ **do**
  for $perm = 1 : num\_perm$ **do**
    $F = random\_permutation(F_0)$;
    $S_1 = Sample(N, num\_attr, num\_vector, F)$;
    $S_2 = Sample(N, num\_attr, num\_vector, F)$;
    Calculate $V_{perm} = T_K(S_1 \cup S_2)$;
  end for
  Construct an empirical P_0 distribution of $\{V_{perm}\}$, $perm = 1 : num\_perm$;
  for $perm = 1 : num_perm$ **do**
    $S_1 = Sample(N, num\_attr, num\_vector, F_1)$;
    $S_2 = Sample(N, num\_attr, num\_vector, F_2)$;
    Calculate: $U_{perm} = T_K(S_1 \cup S_2)$;
  end for
  Calculate $PV(U_i)$, $i = 1 : num\_perm$ with respect to P_0;
  Compare PV with the uniform $\mathcal{U}\{0, 1\}$ and obtain $h_{iter} = 1$ if H_0 is rejected and $h_{iter} = 0$ otherwise;
end for
Test hypothesis that the fraction of the rejections in the sequence $\{h_{iter}\}$, $iter = 1 : num\_iter$ is smaller than $threshold$.
If this hypothesis is rejected $\Rightarrow$ the styles of F_1 and F_2 are accepted as different.

where Φ is the cumulative function of the standard normal distribution, and

$$\widehat{P} = \frac{sum(\{h_{iter}\})}{num\_perm}.$$

The null hypothesis is rejected if $pp < threshold$.

III. Numerical Experiments

We provide numerical experiments in order to demonstrate the capability of the proposed method. We performed evaluation with various parameter values and measured the execution time of the comparisons. The preprocessing consists in removing all spaces in a text file.

A. Comparison of English texts

For this experiment two novels written by American authors whose literary friendship was notable were taken. The first file is The Great Gatsby by F. Scott Fitzgerald (denoted as F). The second one is novel A Moveable Feast by E. Hemingway (denoted as H).

The following tables show the values of pp (1). Here and in all future tables the sources used for the null hypothesis generation (F_1 in the Algorithm 2) are placed in the first column. The styles of two files are believed to be different if the null hypothesis is rejected, i.e. $pp < threshold$.

The increase of the number of attributes and the sample size leads to the improvement of the results yet it takes more time to compute.

In case of $num\_attr = 8$, $num\_vector = 16$ the algorithm completely fails to distinguish the two files while it still correctly marks identical styles.

TABLE I. Comparsion with $num\_attr = 8$, $num\_vector = 16$, average time 92.6 sec

	F	H
F	0.99	**0.99**
H	**0.76**	0.99

TABLE II. Comparsion with $num\_attr = 16$, $num\_vector = 32$, average time 96.4 sec

	F	H
F	0.99	**0.07**
H	0.01	0.99

In case of $num\_attr = 16$, $num\_vector = 32$, the null hypothesis was incorrectly not rejected only once. The accuracy of the result enhanced.

TABLE III. Comparsion with $num\_attr = 32$, $num\_vector = 64$, average time 107.8 sec

	F	H
F	0.99	0
H	0	0.99

At the point of $num\_attr = 32$, $num\_vector = 64$ the method succeds to recognize different files and identify the identical books. The further augmentation of the values doesn't affect the result while the execution time rises considerably.

TABLE IV. Comparsion with $num\_attr = 64$, $num\_vector = 128$, average time 130.5

	F	H
F	0.99	0
H	0	0.99

B. Comparison of Russian texts

We conducted experiments on Russian texts as well. We chose authors of different epochs (XIXth, XXth and XXIth centuries) since intuitively their works should differ considerably.

The list of compared literary works:

- *Demons* by F.M. Dostoevsky
- *The Brothers Karamazov* by F.M. Dostoevsky
- *The Luzhin Defense* by V.V. Nabokov
- *Invitation to a Beheading* by V.V. Nabokov
- *Generation* Π by V.O. Pelevin
- *The Life of Insects* by V.O. Pelevin

The comparisons were made with the folowing parameter values: $num\_iter = 50$, $N = 32bit$, $num\_attr = 16$, $num\_vector = 32$, $num\_perm = 50$, $K = 10$ and $threshold = threshold_{KS} = 0.05$.

The null hypothesis was incorectly not rejected in case of comparing Nabokov's *Invitation to a Beheading* with novels by F. M. Dostoevsky (marked as bold in Table V). This fact is caused by the big difference in the file sizes.

The examples of T_K-statistic values are given in Figure 1 (comparison of Nabokov and Pelevin) and Figure 2 (comparison of 2 works by Dostoevsky). Recall, V_{perm} is calculated for samples from mingled file F_0, while U_{perm} is calculated for samples from F_1 and F_2.

TABLE V. COMPARISON OF RUSSIAN TEXTS

	Dost 1	Dost 2	Nab 1	Nab 2	Pel 1	Pel 2
Dost 1	1	0.99	0	**0.99**	0	0
Dost 2	0.99	1	0	**0.76**	0	0
Nab 1	0	0	1	0.12	0	0
Nab 2	**0.99**	**0.76**	0.12	1	0	0
Pel 1	0	0	0	0	1	0.07
Pel 2	0	0	0	0	0.07	1

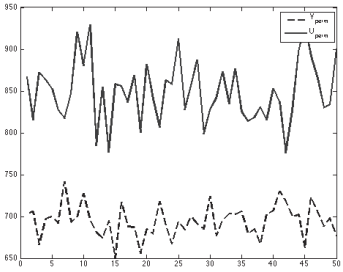

Fig. 1. The values of V_{perm} and U_{perm} in case of different styles

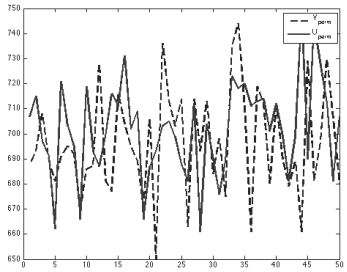

Fig. 2. The values of V_{perm} and U_{perm} in case of identical styles

IV. CONCLUSION

We presented a new re-sampling method designed to disern texts having 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 and its language independence. We studied the influence of the number of attributes and sample size on the accuracy of the result and the execution time.

For further research the analysis of texts in non-european languages (e.g. Arabic, Hebrew) seems to be perspective. Also a comparison of the proposed method with well-known approaches, such as Burrow's Delta [17], [18], [19], Compression models [20], ANOVA [21], Latent Dirichlet allocation [22] and others, is planned to be performed.

ACKNOWLEDGEMENT

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).

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] Granichin O., Kizhaeva N., Shalymov D., Volkovich Z. "Writing style determination using the KNN text model", *In: Proc. of the 2015 IEEE International Symposium on Intelligent Control*, September 21-23, 2015, Sydney, Australia, pp. 900–905, 2015.

[7] 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.

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

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

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

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

[21] 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.

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