Authorship Attribution Method Based on KNN Re-Sampling Approach

Oleg Granichin [1], Lev Klebanov [2], Dmitry Shalymov [1,*], Zeev Volkovich [3]
[1] Saint Petersburg State University (Faculty of Mathematics and Mechanics,
and Research Laboratory for Analysis and Modeling of Social Processes), St. Petersburg, Russia
[2] Dept. of Prob. and Stat., Charles University, Prague, Czech Republic
[3] Dept. Soft. Engn., The Ort Braude, Karmiel, Israel
[*] Corresponding author: *dmitry.shalymov@gmail.com*

***Abstract*—The paper deals with the problem of authorship attribution. We assume that texts are generated based on distinct probability sources. The proposed method is based on re-sampling procedure applied to simulate samples from two texts. We use k-nearest neighbors two-sample test to check if samples were drawn from the same population. The method shows high ability to distinguish texts of different origin.**

***Keywords*—Authorship Attribution; re-sampling; two-sample test; KNN.**

I. Introduction

The writing style of an author is determined by a set of certain words and grammatical structures that he chooses to construct sentences and phrases. Methods of Authorship Attribution (AA) are used when the automatic identification of author of a text is needed. Such problems are particularly relevant in the last two decades due to the rapid growth of text documents in digital form, which are necessary to organize and process [1]. AA problems need "brut force" (labor intensive). They are used effectively in areas such as verification of the author's selection of plagiarism (i.e., finding similarities between two texts) [2], author verification (i.e., to decide whether a given text was written by a certain author) [3], author profiling or characterization (i.e., extracting information about the age, education, sex, etc., of the author of a given text) [4], revealing hidden threats (terrorist threats) and others.

The development of new computational methods for AA becomes very actual for the last two decades due to its numerous applications. Chapters 6 and 7 of [5] consider different approaches and try to systemize the methods. One of the first most important work in this direction is [6] where the authorship of "The Federalist Papers" was determined based on Bayesian statistical analysis of common words frequencies. This opened up the field to the exploration of new types of stylometric features and new modeling techniques [1].

From a machine learning viewpoint the task of AA can be considered 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].

One of the most important problems of AA is the allocation of a set of properties (characteristics) which will be used to identify the author. The simplest form of text analysis is simple descriptive statistics. We can easily calculate word lengths, the mean number of syllables per word, number of words per sentence, etc. It is known that no single feature has been found that robustly separates different authors in a large number of cases [8]. There are well known character features for which a text is viewed as a sequence of characters [1]. By this way various character-level measures can be defined, including alphabetic characters count, digit characters count, uppercase and lowercase characters count, letter frequencies, punctuation marks count and so on [9]. This type of information is easily available for any natural language and corpus, and it has been proven to be quite useful to quantify the writing style [10].

In this paper we solve the problem of AA based on the statistical method described in [11]. It is assumed that the "author" is a pseudo-random generator of words that can be specified unique distribution. Then the problem of the definition of style is reduced to the comparison of the relative independence of the texts.

II. Two-Sample Test Methodology

Two-sample hypothesis testing is a statistical analysis approach designed to examine if two samples of independent random elements, drawn from the Euclidean space R^d, have the same probability distribution function. Formally speaking, let $X = X_1, X_2, \ldots, X_m$ and $Y = Y_1, Y_2, \ldots, Y_n$ be a two independent random variables whose distribution functions F and G are unknown. A two-sample problem consists of testing the null hypothesis $H_0 : F(x) = G(x)$ against the alternative $H_1 : F(x) \neq G(x)$.

Apparently, the most useful and general nonparametric method is the KS (Kolmogorov, Smirnov)-test [12] which is a nonparametric test of the equality of continuous, one-dimensional probability distributions. The Kolmogorov–Smirnov statistic $D = \sup_x |\tilde{F}(x) - \tilde{G}(x)|$ measures a distance between the empirical distribution functions $\tilde{F}(x)$ and $\tilde{G}(x)$ of two samples. The test is asymptotic distribution-free. So, the distribution of the test statistic does not depend on the underlying distributions of the data for sufficiently big samples. The test can be also used to compare a sample with a reference probability distribution (one-sample KS-test). In this case the KS-statistic quantifies a distance between

the empirical distribution function of the sample and the cumulative distribution function of the reference distribution.

For the multivariate case many tests have been also derived [13], [14]. The multivariate two sample test based kernel statistics introduced in [15] found many notable applications in the machine learning theory. The two-sample energy test described in [16] 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 proportion are approximately equal for the well mixed samples. From this point of view a cluster validation has been considered in [11]. K-nearest neighbors type coincidences model in the current paper deals with the statistic:

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

which represents the number of all K nearest neighbors type of coincidences. Asymptotic behavior of this statistic has been studied in [17]. It is important to note that the asymptotic normal distribution can hardly be applied in comparison of two real texts due to the inherent heterogeneity. The null hypothesis law can be yet simulated in the spirit of the bootstrap methodology (see, e.g. [18]). Construction of an empirical distribution of the pooled samples indirectly suggests their identical underlying distributions under the null hypothesis. However, once these distributions are, in fact, different, the above procedure (using just "prior mixing") can produce a polluted distribution. Due to this reason we accurate the inference process by means of the procedure described below.

III. The 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 ingredient of the method aiming to reflect the sources structure. We form samples by means of N-grams as connecting sequences of N symbols from a text by the following way

A. Sampling Procedure:

Sample$(N, NWORD, NVEC, F)$,

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

The samples are simulated by as follows:

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 the space R^{NWORD}.

As was mentioned above the normal distribution rarely appears as the null hypothesis distribution in the considered problem. So, 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.

B. Algorithm

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

The comparison is performed as follows:

1) Introduce $F_0 = F_1 \cup F_2$
2) for $iter = 1 : ITER$
3) for $perm = 1 : NPER$
4) $F = random\_permutation(F_0)$;
5) $S_1 = Sample(N, NWORD, NVEC, F)$
6) $S_2 = Sample(N, NWORD, NVEC, F)$;
7) Calculate: $V_{perm} = \{T_K\left(S_1 \cup S_2\right)\}$;
8) end;
9) Construct an empirical P_0 distribution of $\{V_{perm}, perm = 1 : NPER\}$.
10) for $perm = 1 : NPER$
11) $S_1 = Sample(NA, NWORD, NVEC, F_1)$;
12) $S_2 = Sample(NA, NWORD, NVEC, F_2)$;
13) Calculate: $U_{perm} = \{T_K\left(S_1 \cup S_2\right)\}$;
14) end;
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;

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 bigger significantly from 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 $pp < TR$.

IV. Numerical Experiments

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

A. English Text Collections

Three text collections are compared with omitting 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.

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

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

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

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

TABLE I: Comparison of the three text collections

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

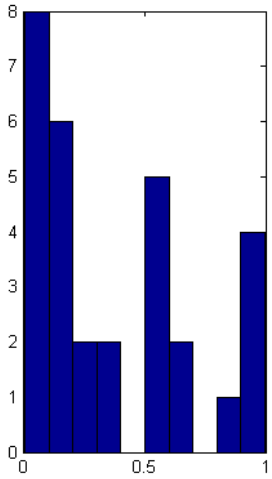

Fig. 1: Histograms of p-values in comparison of the HP collection with itself.

Examples of histogram for asymptotic p-values returned by KNN-test are presented in Fig 1. Each 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 II and III 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 II: 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

B. 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 [19] and for four periods [20].

Three periods of Shakespeare's plays are: I (optimistic)

TABLE III: 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

period (1590-1600), II (tragic) period (1601-1607) and III (romantic) period (1608-1612).

Corresponding text collections for these periods are compared one with another. The corresponding values for p_1 are presented in Table IV.

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 IV: 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 is also known 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.

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: 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 [20].

V. Conclusion

We proposed a new method based on re-sampling approach to distinguish texts with different writing styles. The method is based on comparison of empirical distributions constructed for the two-sample KNN-test statistic for samples drawn from the same source and different ones. We provided numerical experiments that demonstrate a high ability of the proposed method.

Acknowledgment

The work is supported by SPbSU grant 6.37.181.2014.

References

[1] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Assoc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, 2009.

[2] S. M. zu Eissen, B. Stein, and M. Kulig, *Plagiarism Detection Without Reference Collections*. Berlin, Germany: Springer, 2007.

[3] M. Koppel and J. Schler, "Authorship verification as a one-class classification problem," in *Proceedings of the 21st Int. Conf. on Machine Learning. NewYork: ACM Press*, 2004.

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

[5] O. Granichin, V. Volkovich, and D. Toledano-Kitai, *Randomized Algorithms in Automatic Control and Data Mining*. Springer-Verlag: NY, 2015.

[6] F. Mosteller and D. Wallace, "Inference in an authorship problem - a comparative-study of discrimination methods applied to authorship of disputed federalist papers," *JASA*, vol. 58, no. 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] P.Juola, "Authorship attribution," *Foundations and Trends in Information Retrieval*, vol. 1, no. 3, pp. 233–334, 2006.

[9] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *J. Assoc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 378–393, 2006.

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

[11] Z. Volkovich, Z. Barzily, R. Avros, and D. Toledano-Kitay, "On application of the k-nearest neighbors approach for cluster validation," in *XIII Int. Conf. Applied Stochastic Models and Data Analysis (ASMDA 2009)*, 2009.

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

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

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

[15] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample problem," *CoRR*, vol. 0805, 2008.

[16] 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, February 2005.

[17] 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.

[18] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC, ISBN 0-412-04231-2. software, 1993.

[19] F. W. Bradbrook, *Artist and Society in Shakespeare's England*. Brighton, UK: Harvester Press, 1982.

[20] R. P. Halleck, *Halleck's New English Literature*. New York: American Book Company, 1913.