PLISKA
STUDIA MATHEMATICA
BULGARICA

ПЛИСКА
БЪЛГАРСКИ
МАТЕМАТИЧЕСКИ
СТУДИИ

Pliska Stud. Math. Bulgar. **22** (2013), 25–32

PLISKA
STUDIA MATHEMATICA
BULGARICA

CLASSIFICATION OF TEXTS' AUTHORSHIP USING A REGRESSION MODEL ON COMPRESSED DATA*

Diana Dackova, Plamen Mateev

Abstract. An algorithm for text authorship identification is proposed. The procedure is based on the Kolmogorov complexity and uses regression models on the length of the compressed texts. The classification employs the regression parameters estimates. Different combinations of compressor parameters and the preliminary processing on the data are examined using prose texts of a few English classics.

1. Introduction. First known attempts in the search of a specific frequency distribution in a text belongs to Al Kindi in a 9th century [10]. How to use frequencies to decipher encrypted messages is described in his "Manuscript on Deciphering Cryptographic Messages".

Now frequency distributions of key words from a given set are widely used tools of stylometry. The stylometry may be defined as a science (methods, instruments etc.) of classification of given text or text segment. In other words, the aim is to determine authorship, historical authenticity or other similar questions.

An example of accurate statistical approach belongs to A. A. Markov. In his paper [6] he had studied the distribution of vowels and consonants among initial 20000 letters of "Evgenij Onegin" and had used the notion "events which are linked to the chain", so called now "Markov chains".

*The research was partially supported by appropriated state funds for research allocated to Sofia University (contract No 125/2012), Bulgaria.

2010 *Mathematics Subject Classification:* 68T50,62H30,62J05.

Key words: Text authorship identification, Classification, Compression, Linear Regression.

The explosion of computers' dissemination and ability, as well as streams of texts in digital form causes an amazing variety of methods for text processing - text mining. Unfortunately, most of them ignore statistics, "the first and most successful information science".

Our work is provoked and due to the series of papers of M. B. Malyutov [4],[5]. They are based on the intuitive approach of D. V. Khmelyov [2] and his *"Relative Complexity Classifier"*. Malyutov finds out the theoretical frame for Hmelyov's algorithm and its modification called *"Conditional Complexity of Compression"* (CCC) classifier. He approved its usage in the great Shakespeare's problem (the question of his authorship and identity [4]).

The aim of the our article is to explore two dimensional modification of the essentially unidimensional CCC characteristics of text fragment.

The following sections present a short review of CCC, its theoretical justification arguments in favour of our proposal, the results of our experiments and a conclusion.

2. Complexity and MDL principle. In series of articles Kolmogorov introduced the concept of "complexity" of a character string as the length of the shortest software program that reproduces the string. He proves that it is asymptomatically equivalent to Shannon's entropy, which is known to be a lower bound for the compressed size of the string.

The importance of these concepts is the presented opportunity to estimate the entropy: shortest message length $d(A)$ that the event A would happen (appropriately normalized) can be used to estimate probability $P(A)$ using equation:

$$d(A) = -\log P(A).$$

Later the MDL principle (Minimum Description Length) emerges from Kolmogorov's complexity, which is kind of elaborated version of the entropy of an object: the construction of stochastic model that could allow to produce the shortest description of the object and the model itself.

The probability distribution of characters in an evaluated text is characterised by the *entropy of Shannon* [9] of this text. The *Kolmogorov's complexity* [3] is asymptotically equivalent to the entropy but they both are not computable in the general case. It is appropriate to use a model which, according to the *Minimum Description Length principle* [8], permits the shortest description of the data and the model itself. Wyner and Ziv [11] prove that *LZ77 and LZ78 algorithms* are *Universal Compressors* and if used on *stationary ergodic distributed* data then the compression rate is asymptomatically equivalent to the entropy of the source.

2.1. Relative Complexity Classifier. Khmelev proposed in [2] a simple text classifier based on *conditional complexity* estimated via application of compression algorithm by which the relative complexity of text A with respect to

text B may be determined. Let us denote the length of given text A (the number of symbols) by $l(A)$ and the concatenation of two texts A and B as $[AB]$. The length of the concatenated texts is equal to the sum of its ingredients: $l([AB]) = l(A) + l(B)$ and first $l(A)$ symbols of $[AB]$ coinsides with A and the last $l(B)$ symbols coincides with B. The output of compression of text A is denoted by A_c. Khmelyov's definition of *relative complexity* $C(A|B)$ of text A with respect to text B is the difference

$$C(A|B) = l([BA]_c) - l(B_c).$$

The experiments were performed on the corpora of 385 literary texts of 82 writers with total size of about 128 MB. A small text fragment U_i of size 50-100 kB for each author $i = 1, \ldots, 82$ is used as control text. All other texts of author i are concatenated in a single segment T_i.

Classification rule for authorship $\mathcal{A} : \{U_1, \ldots, U_n\} \to \{1, 2, \ldots, n\}$ of the control fragment U_i is

$$\mathcal{A}(U_i) = \arg\min_j \{C(U_i|T_j\}.$$

The experiment is relatively successful. The classification rule achieves 71 out of 82 correctly classified authors with one of 16 tested data compression algorithms. This result is better than the experiments of Khmelev [1] with authorship attribution based on Markov chain models of text.

2.2. Conditional Complexity of Compression Based Test. Malyutov worked on improvement and refinement of the Khmelev's classifier in the form of Student type test. Following [5], we assume that a literary text T, represented as a binary string x^n is characterized by stationary ergodic distribution $P = P_T$ dependent on author's uniqueness. Let the author of T is known undoubtedly and lets another (query) text Q, represented as a binary string y^m, ought to be checked if it does belong to the same author. Denoting P_Q the distribution of the string y^m, the task is to test the hypothesis $\mathrm{H}_0 : P_Q = P_T$, against $\mathrm{H}_1 : P_Q \neq P_T$.

The test statistics is constructed as follows: several (*slices*) non overlapping fragments $y_i,\ i = 1, \ldots, s$, are excerpted from the string y^m. All slices have the same length and are separated with "small" brakes to provide independence of slices. The

$$CCC_i = C(y_i|x^n) = l([x^n y_i]_c) - l([y_i]_c),\ i = 1, \ldots, s,$$

are considered as s independent asymptotically normal distributed individual observations. The last proposition is corollary, of the theorem proved in [5].

The statistics

$$t(Q, T) = \frac{\overline{CCC}.\sqrt{s}}{\sqrt{S^2(CCC)/(s-1)}},$$

where

$$\overline{CCC} = \sum_{i=1}^{s} CCC_i, \quad S^2(CCC) = \sum_{i=1}^{s} (CCC_i - \overline{CCC})^2,$$

has central t-distribution under assumption $\mathrm{H}_0 : P_Q = P_T$. The significance of the difference between authorship is determined as quantile of t-distribution with $s-1$ degrees of freedom.

2.3. Regression based algorithm. Now we propose given text T to be parametrised via simple linear regression.

First step is to do a preliminary processing on the text. Second, prepare set of text fragments of different size. Fragments are chosen randomly. Third, all of them are compressed separately. Forth, we estimate the regression parameters.

We use the fact that the size of compressed fragment may be considered as random variable with normal distribution [5]. The parameters of simple linear regression are considered as characteristics of the given text. The independent variable X is the text fragments length (in kB) and sizes of compressed fragments (in bytes) is the dependent Y variable:

$$Y = \beta_0 + X\beta_1 + \epsilon,$$

where ϵ are independent, equaly distributed with normal distribution $N(0, \sigma^2)$. The estimated regression parameters $\hat{\beta}(T) = \hat{\beta}(\hat{\beta}_0, \hat{\beta}_1)$ and the covariance matrix of those estimates:

$$S(T) = \begin{pmatrix} s_0^2 & s_{01} \\ s_{01} & s_1^2 \end{pmatrix}$$

for the given text T are obtained.

The parameters, which we may variate are: (i) the preliminary processing on the tex; (ii) compression method; (iii) number of observations (the number of text fragments) and (iv) their size.

The algorithm is applied on every text of a given author.

The estimated regression parameters map the text as point in two dimensional space: (intercept $\times$ slope coefficient). In addition, covariance matrix of the estimates is determinate too.

2.4. Classification measure. The Mahalanobis distance between text points will be used as measure of classification quality. Squared Mahalanobis distance between two vectors x and y of the same dimensions and given positive definite square matrix S is defined as:

$$D_S^2(x, y) = D^2(x, y; S) = (x - y)^T S^{-1} (x - y).$$

Let a and b are two texts and $\hat{\beta}(a)$ and $\hat{\beta}(b)$ are estimates of corresponding linear regression coefficients and $S(a)$ and $S(b)$ are their covariance matrices.

A procedure for authorship identification based on regression model of compressed fragments may be evaluated via minimax of similarity based on Mahalanobis distance. Let's A and B are two sets of texts of two authors. Let a is the text of the set A and b is text from set B. All texts from the two sets A and B are subjected to the same procedure of preliminary text processing, fragment extraction and compression. For every text the regression parameters are estimated and mapped on the space (intercept $\times$ slope coefficient).

The procedure is "good" authorship classifier when provides "small" distance between texts of the same author and sufficiently "large" between texts of different authors. Formally the procedure has to ensure:

$$\min_{a\in A, b\in B} \{D(a,b;S(a)), D(b,a;S(b))\} > \max_{c,c'\in C, C\in\{A,B\}} \left\{D(c,c';S(c))\right\}.$$

Here, for simplicity, we denote $D(a,b;S(a))$ as short form of $D(\hat{\beta}(a),\hat{\beta}(b);S(a))$.

3. Experiment.

3.1. Data. Texts of three English authors were used for experiments:

1. Charles Dickens – "A Tale of Two Cities" – 1859, 738 kB; "David Copperfield" – 1850, 1884 kB; "The Pickwick Papers" – 1836, 1694 kB;
2. George Eliot – "Adam Bede" – 1859, 1128 kB; "Daniel Deronda" – 1876, 1694 kB; "The Mill on the Floss" – 1860, 1113 kB;
3. William Thackeray – "The Virginians" – 1857, 1835 kB; "Vanity Fair" – 1848, 1670 kB.

The texts are written in the same language (not translated), close in time, in the same genre and fairly big enough.

3.2. Preliminary processing of the original texts. The texts are saved as *plain text*, 8-bit ASCII coding. Thus, all additional formatting as paragraphs, pages and special characters resulted from the particular edition is removed. Any *spelling errors* (if any) are corrected. The *personal names* are removed so that the individual text and its story would not interfere the style of the author. We decided to leave the *punctuation*, because the authors were better distinguished this way. The *capital letters* are converted into lower case. In the end, the modified text contains lower case letters, punctuation, intervals and new lines.

3.3. Text fragment extraction. The sample of fragments from given text was determined according rules:

- the size of the sample n and the minimal size fragment in the sample k are fixed;
- the fragments are chosen randomly – every fragment begins at a random place in the text;
- the sampling starts with fragments of length k;

- after m fragments with equal length, the length augments with 2 kB;
- the procedure ends when the fixed number n is reached or fragment length exceeds text length.

This procedure is performed on each of the eight texts.

3.4. Compressor parameters. The used compression algorithm is LZMA (Lempel-Ziv-Markov Algorithm) provided by the free software 7-Zip [7]. This is a lossless dictionary based compressing algorithm, similar to LZ77. It has been seen that there is little difference what *level of compression* is applied (fast/normal/ultra) so we were set with normal. The parameter that causes the biggest differences is the *dictionary size*, so we tried with 5kB, 20kB, 50kB, 200kB and 16MB. The best compressing results are achieved with dictionary exceeding the text size but best in terms of distinguishing authors happened to be 5kB. The explanation we provide is that the small dictionary exhibits the words' mutual disposition along with their variety.

3.5. Experimental results. The eight texts of the three authors (3 + 3 + 2) were treated according the described preliminary processing. Then from each text a sample of n fragments was prepared, starting with k as length of the smallest fragment. The number of fragments with equal length m was fixed to 10. The result we have a set of samples $\{s = (k, n, c)\}$, where c is one of the eight texts, $k \in \{10, 12, \ldots, 118\}$ and the volume of sample $n \in \{10, 11, \ldots, 700\}$.

All fragments from the sample s were compressed and the parameters of the regression line $\hat{\beta}(s)$ were estimated as well its covariance matrix $S(s)$. For given values of (k, n) all 42 values of Mahalanobis distances are obtained and the "worst" closeness or minimal Mahalanobis distance between authors was computed:

$$\min D(k,n) = \min \{D(a,b;S(a)) | a \in A, b \in B, \forall (A \neq B)\},$$

where A and B are two of the three authors and (k, n) are given.

The next step was to find the areas of (k, n) for which $minD(k, n)$ is maximized. A scatter plot is shown on the left part of Figure 1 with all pairs (k, n), where the numbers n are on the abscissa and values of $\min D(k, n)$ are on the ordinate axis. On the right of Figure 1 the same is shown for these (k, n) for which $\min D(k, n)$ is less than 0.05. Local maximum (for n) of $\min D$ may be suggested in the interval about $n = 250$ on the both figures.

The range of n in $225 \leq n \leq 285$ is interesting because $\min D \geq 0.231$ there. Within this range of n, maximum values of $\min D$ are observed for $42 \leq k \leq 56$ (see the left part of Figure 2). Authors' equidistant ellipses are apparently separated in those ranges of (k, n).

On the right part of the Figure 2 eight 95% confidence ellipses for the eight texts in the parametric space (intercept × slope coefficient) of regression parame-

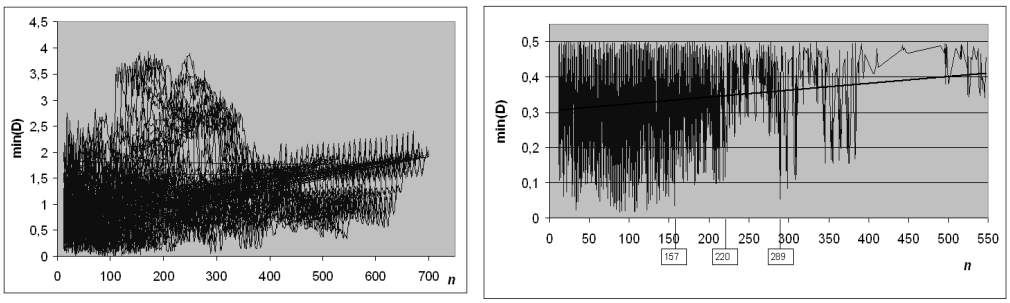

Fig. 1. Minimal Mahalanobis distance $\min D$ for n from 11 to 700, k from 10 to 118 (the wolf head on the left) and its lower part magnified (on the right).

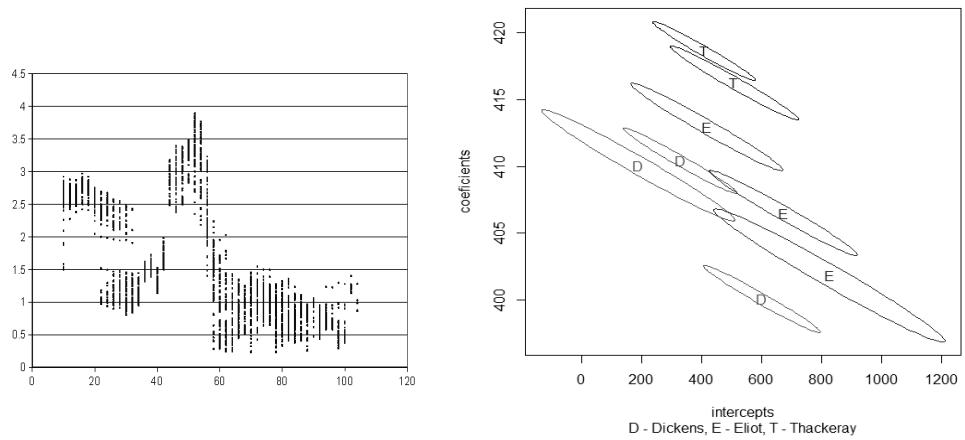

Fig. 2. The minimal Mahalanobis distance $\min D$ as a function of k for $n \in [225;288]$ on the left. Confidence ellipses for regression parameters of the eight texts for $k = 52$ kB and n=250 on the right.

ters estimates for samples with $k = 52$ and $n = 250$ are shown as an illustration.

4. Conclusion. Experiments have shown that two-dimensional parametric characterization of text fragments is a promising approach for authorization and classification. We hope to improve the procedure adjusting the preliminary text processing and exploring change of other algorithm parameters. At next stage the method will be applied to Bulgarian texts and will include experiments in which a fragment of unknown author is concatenated to text fragments of known authorship.

REFERENCES

[1] D. V. Khmelev. Using Markov Chains for Authorship attribution, *Vestn. MGU*, sr. 9, Filolog., **2**, (2000), 115–126.

[2] O. V. Kukushkina, A. A. Polikarov, D. V. Khmelev. Using Literal and Grammatical Statistics for Authorship Attribution. *Problems of Information Transmission* **37** (2000), No 2, 96–108.

[3] A. N. Kolmogorov. Three Approaches to the Quantitative Definition of Information. *Problems of Information Transmission* **1** (1965), 1–7.

[4] M. Malyutov. Review of Methods and Examples of Authorship Attribution. *Review of Applied and Industrial Mathematics* **12** (2005), No 1, 40–79.

[5] M. Malyutov. Compression based homogeneity testing. *Proceedings (Doklady) of Russian Academy of Sciences* **443** (2012), No 4, 427–430.

[6] A. A. Markov. An example of statistical study on text of "Eugeny Onegin" illustrating the linking of events to a chain. Izvestija Imp. Akad. nauk **VI** (1913), No 3, 153 (in Russian).

[7] I. Pavlov. 7-Zip, `http://www.7-zip.org`, 2012.

[8] J. Rissanen. Modelling by the shortest data description. *Automatica*, **14** (1978), 465–471.

[9] C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, **27** (1948), 379–423, 623–656.

[10] Singh, Simon. The code book: the science of secrecy from ancient Egypt to quantum cryptography, NY, Anchor Books, 2000, ISBN 0-385-49532-3.

[11] A. Wyner, J. Ziv. On entropy and data compression. IEEE Transactions on Information Theory, 1991.

Diana Dackova
Faculty of Mathematics
and Informatics
Sofia University
"St.Kliment Ohridski"
5, J. Bourchier Str.
1164 Sofia, Bulgaria
e-mail: `diana.dackova@gmail.com`

Plamen Mateev
Faculty of Mathematics and Informatics
Sofia University "St.Kliment Ohridski"
5, J. Bourchier Str.
1164 Sofia, Bulgaria
and
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia, Bulgaria
e-mail: `p.mateev@gmail.com`