The advantage of the Mahalanobis distance to assess differences multidimensional characteristics of electronic texts

Anastasia O. Shumskaya

Department of Complex Information Security
Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia
Shumskaya.AO@gmail.com

Abstract

The article presents the effectiveness of the Mahalanobis distances in case of the text origin identification. In order to calculate this metrics were used text features of the original texts and the texts generated on the basis of the original texts. As a generation method there were used the synonymy and Markov chain method.

1. Introduction

Recognizing the authorship is actively studied and expandable area of the scientific knowledge. Many scientists investigate the existing and develop new methods by which the problems associated with the attribution of text arrays can be solved.

The problems of the text origin identification become especially important at the background of anonymity and plagiarism when computer nets are used. So there are needed new methods able to give the answer whether the text was written by a human being (natural text) or it is artificial. The artificial text is the text generated by a special program-generator.

2. Statistical text attribution methods

Statistical analysis methods are based on the fact that the author's manner can be recognized by some text feature or by the set of such features – named author's invariant. The examples of text features may be average word length, N–gram frequency, some words repetition rate.

Statistical methods are commonly used when solving attributing problems. Their advantages are higher calculating and training speed compared with the machine methods, also universality. The drawback is the necessity of the author's invariant extraction that requires additional computation. The volume of this computation depends on the problem under consideration and on the quality of the texts [1].

3. Data set for experiment with artificial texts

It's practically impossible to select the identification text features that can exactly tell one author from another. So it is considered enough that the parameter can recognize different groups of authors. It is assumed that exists a big group of authors whose average value of some parameter differs greatly. In such a case the parameter will not be able to recognize texts of different authors of the same group. They could be recognize using simultaneously many parameters having different character.

Investigation allowed to find parameters typical for such methods of artificial texts generation as synonymy and Markov chain method and features not typical for them. In both cases characteristics of the artificial texts logically followed the suit of their original. In addition, in case of synonymy there were changes according to the level of the artificial text singularity. It proves the influence of the generation algorithm on some parameters variation.

Calculations done allow to pick out some features that would greatly change in every as well as the features that changed less.

The most changed text characteristics of synonymy:

- the number of service words;
- frequency of some words;
- the number of short words.

Less changing text characteristics of synonymy:

- the number of long words;
- the average word length.

The most changed text characteristics of Markov chain method:

- the number of sentences;
- frequency of some words.

Less changing text characteristics of Markov chain method:

- the number of short words;
- the number of long words;
- the average word length;
- the number of service words.

In order to get more correct estimation of text characteristic variation found by calculation may be used some math methods, namely the fitting criterion.

The author used Mahalanobis distance calculation that should prove the possibility to use the statistics methods when solving problems of the texts origin and detecting the artificial texts.

There were used following designations in experiment: v_original – average value vector of nature (origin) texts, v_Symonim – average value vector of texts generated with program SyMonym, v_Article – average value vector of texts generated with program Article Clone Easy, v_Delirium – average value vector of texts generated with program Delirium.

4. Mahalanobis distance calculation experiment

In statistics the Mahalanobis distanse is a measure of distance between two random values. It generates the Euclidean

distance. It is used to identify and gauge similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant [2]. In other words, it has a multivariate effect size.

To use Mahalanobis distance in order to recognize text origin is needed to calculate covariance matrix. As usual one can calculate it based on the known class sets. Then you need to calculate the Mahalanobis distance from a given point to the selected class and appreciate it.

Input data for Mahalanobis distance experiment are

- average value vectors for the following sets of texts: nature texts, texts generated by the program SyMonim, texts generated by the program Article Clone Easy, generated by the program Delirium;
- some input text with a known origin that could be attributed to a certain class (taking part in this experiment). It should be mentioned that as the input texts there were taken the text that were not used for calculating the averaged vector.

The calculation is performed by the classical formula:

$$Dm = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}, \quad (1)$$

where x – input text value vector; μ - average value vector of some text class, S - integrated covariance matrix.

Block-diagram of calculation algorithm is shown in fig. 1.

Results of calculation for the vectors mentioned above are given in Table 1 (synonymy) and Table 2 (Markov chain method).

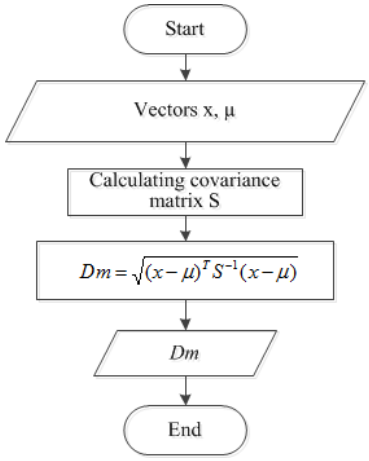

Figure 1: Block-diagram of calculation Mahalanobis distance algorithm in article

Table 1: Results of calculation Mahalanobis distance (synonimy)

Input text (vector x)	Mahalanobis distance between vector x and average vector of some class (vector μ)		
	v_original	v_Symonim	v_Article
Nature (origin) text	0,13	0,269	0,416
Text generated with SyMonim	0,142	0,134	0,107
Text generated with Article Clone Easy	0,259	0,375	0,072

Table 2: Results of calculation Mahalanobis distance (Markov chain method)

Input text (vector x)	Mahalanobis distance between vector x and average vector of some class (vector μ)	
	v_original	v_Delirium
Nature (origin) text	0,210	0,265
Text generated with Delirium	0,062	0,014

The results of calculations are similar to the results of Euclidean distance calculation. It can be mentioned that for synonymy regularity obtain the lowest Mahalanobis distance for texts that belong to one class, also the greatest measure the distance from other classes with more unique text (Article Clone Easy).

In case of Markov chain method results are identical calculations Euclidean metric: text characteristics weakly changed in the generated code, so the difference between the distance to two different classes in the input text is not as revealing as in case of synonimy. Repetition of observational data suggests that for this method it's needed to change the characteristics set to become results that are relevant for identifying a text generated on the basis of this algorithm.

5. Conclusion.

The experimental results show the volume of similarity random input text with specially researched samples artificial texts. It is assumed that this and similar calculations can afford to develop the most effective way to identify the artificial generation of texts.

Mahalanobis distance can be used to identify the artificial generation of text. The minimum distance was calculate from input text vector to average vector of class that this text belongs to.

As the average value vector for the class under test is necessary to use text set with high (over 65%) uniqueness of the texts. The identifying characteristics of this parameter weakly manifested with the lower rate.

In case of the Markov chains method it's needed to change invariant to achieve a more significant difference between the numerical values.

This requires testing of the methods described on the identification of different inputs and other methods of generating artificial texts.

6. References

[1] Romanov A.S., *Development and research of mathematical models, methods and software of information processes in the text author identification*, W-Spektr, Tomsk, 2011.

[2] Hachumov M.V., *"Distances, metrics and cluster analysis"*, Artificial intelligence and decision- theory, Boston, 1989. Vol.1:81-89, 2012.