VOL4 NO. 1 March 2015 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: www.scjournal.ius.edu.ba

IUS Soft Computing Research Group

A study in Authorship Attribution: The Federalist Papers

Nesibe Merve Demir
International University of Sarajevo,
Faculty of Engineering and Natural Sciences,
Hrasnicka Cesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina
ndemir@ius.edu.ba

Article Info

Article history:
Article received on February 2015
Received in revised form March 2015

Keywords:
Authorship attribution; stylometry; support vector machine; Mahalanobis Distance

Abstract

In order to authorship attribution techniques, the Federalist Papers have been applied as a testing-ground that twelve of which are claimed by Alexander Hamilton and James Madison. The value of novel stylometric techniques through implementation of them to the Federalist problem is what the paper subjects to. Support vector machines and nearest neighbor techniques alongside Artificial Neural Network techniques are used for classification of selected disputed paper. Encouraging results achieved in the research.

1. INTRODUCTION

The science which focuses on inferring characteristics of the author from the characteristics of documents that is written by that author is called as authorship attribution as a problem with its long history including a broad spectrum of application.

1.1 Authorship Attribution

A text of unknown authorship is appointed to one candidate author in a normal authorship attribution problem that a set of candidate authors for whom text samples of undisputed authorship are available as arranged. So, it can be seen as a multi-class single-label text categorization task according to a machine learning point-of-view (Sebastiani, 2002). As a task, generally it is known also by researchers with a background in computer science as the authorship (or author) identification.
With the attempts for determination of features in order to quantify writing style, research regarding authorship attribution was controlled until the late 1990s that is known as 'stylometry' as a line of research (Holmes, 1994; Holmes, 1998). An important variety of measures had been suggested therefore in which length, word length, word frequencies, character frequencies, and vocabulary richness functions are included. Nearly 1,000 different measures had been suggested that far according to the estimation of Rudman (1998).

The changes have arisen regarding authorship attribution studies since the late 1990s. With a great number of electronic texts available by Internet media including emails, blogs, online forums, etc., the necessity in order to manage the information effectively have increased which resulted with an important impact in scientific areas including information retrieval, machine learning, and natural language processing (NLP), eventually. Authorship attribution technology got affected by the development of the mentioned areas.

The case of author attribution has got a considerable attention and interest recently with an increasing number of data in various forms in which emails, blogs, and messages on the internet and SMS are included. New kinds of applications have existed like web searching, plagiarism

detection, spam email detection as well as finding the authors of disputed or anonymous documents in forensics against cybercrime as additionally to its traditional application as spreading knowledge with respect to the authorship of disputed texts in the classical literature.

1.2 The Federalist Papers

For the purpose of persuading New Yorkers in order to support ratification of the proposed new constitution of the United States of America, seventy seven articles were printed between the years of 1787 – 1788.

'Publius' is published in 1788 which includes the Federalist Papers that are a series of 85 political essays.

Alexander Hamilton, James Madison, and John Jay are accepted as the authors by scholars that the real author(s) were kept as a secret initially. Then, Hamilton and Madison followingly presented their own lists regarding to declare the authorship. 12 essays claimed by Madison and Hamilton were the difference between the two lists. Therefore, 12 of them are under debate for authorship while 73 texts can be considered as having known author(s). The essay numbers 49-58, 62 and 63 are of these 12 authorship texts that under debate. It is concluded by Mosteller and Wallace (1964) that all the essays under debate were written by Madison, including the possible exception that essay number 55 might be written by Hamilton with their conducted early study. Hamilton and Madison claimed twelve of the eighty five papers.

The scholars had opinion division when Mosteller and Wallence published the first edition of their book, Inference and Disputed Authorship: the Federalist. On various historical and stylistic grounds Hamiltonian and Madisonian authorship of the twelve disputed papers were argued in a serious way.

Particular numbers of the Federalist is tabulated by considering particular numbers of the Federalist as a reference of the situation which is found by Mosteller and Wallace.

Mosteller and Wallace in their main study applied Bayes' Theorem for drawing inferences with regard to the probabilities of the competing hypotheses that Hamilton or Madison wrote the disputed Federalist Papers by basing on evidence obtained from the rates of usage of thirty marker words.

Mosteller and Wallace (1984) have caution for their readers: Hamilton's and Madison's styles are unusually similar; new problems, with two authors as candidates, should be easier than distinguishing between Hamilton and Madison.

McColly and Weier (1983) provide an example study by applying to the federalist papers as a test case by claiming a likelihood-ratio approach to attribution in which the assumption of the frequency of occurrence of a particular marker word is distributed as a Poisson random variable is included. According to the chi-square results, Mosteller and Wallace's findings seemed to confirm that Madison had written both disputed papers.

Kjell (1994) has conducted a study on the federalist problem as recently by applying to letter-pair frequencies as input to a neural network. By applying to back-propagation, it was trained to discriminate between Hamilton and Madison. The results confirm those of Tweedie et al.(1994) and Mosteller and Wallace broadly.

Feature Selection

Establishing features that work as effective discriminators of texts under study is one of critical issues in research on authorship analysis which are lexical. In this research thirty-seven textual descriptors are used, numbers of characters, words, sentence, and chosen English words in paragraphs.

Data Architecture

The number of inputs equaled the number of textual descriptors used, thus it is 37. When PCA applied then 20 inputs were used. Two types of input used, with PCA and without PCA. Training data consists of three writers' one article. Also disputed table (no.14) is the main testing data. After using data in that way, training data changed. All training data and another article of each writer's combined and used as combined data. Disputed paper used as only testing data.

2. SUPPORT VECTOR MACHINES

A learning algorithm which carries out binary classification (pattern recognition) and real value function approximation (regression estimation) tasks is a support vector machine (SVM). Non-linearly mapping the n-dimensional input space into a high-dimensional feature space is the aim with it. Through the way of constructing a linear classifier the high-dimensional feature space gets classified. A maximum-margin hyper plane which takes place in the transformed input space is created by the basic SVM.

Vapnik developed the basic concepts of SVM (Support Vector Machines). The opinion which underlies the SVMs' concept is the structural risk minimization where the generalization error is minimized such as true error on unseen examples. It is bordered by the sum of the training set error and a term depending on the VC (Vapnik-Chervonenkis) dimension of the classifier and on the number of training samples as well. The number of free parameters which is applied in the SVM depends on the margin by separating the data and does not depend on the number of input features as differently from many other learning algorithms. While the data vectors are separable with a wide margin, for the context of high-dimensional applications like text document and authorship categorization it is a big advantage.

As Support Vector Machines only compute two-way categorization, Q two-way classification models were generated, where Q is the number of author categories, and each SVM categorization was applied Q times. This produced Q two-way confusion matrices.

On the basis of a small set of training examples, called the support vectors, the best decision surface σ i is determined. Figure 1 shows that σ_i is the best decision

surface or hyper plane (the thicker line) because it is the middle element of the parallel decision surfaces (the thin lines). The support vectors are the small boxes (Sebastiani, 2002: 30):

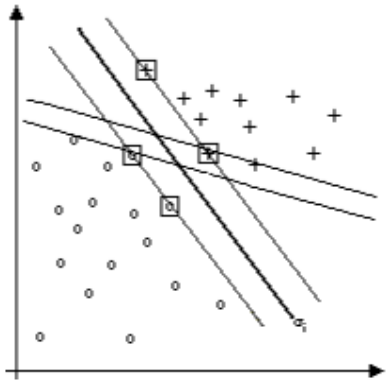

Figure 1. Learning Support Vector classifiers

2.1 Success of Support Vector Machines

Table 2 shows results of support vector machine.

Table 2.Support Vector Machine results

	Hamilton	Jay	Madison
Hamilton	N/A	Hamilton	Madison
Jay	Hamilton	N/A	Madison
Madison	Madison	Madison	N/A

This table should be read as follows: After six SVM's are trained to distinguish texts by author pairs

1) Hamilton/Jay,
2) Hamilton/ Madison,
3) Jay/Hamilton,
4) Jay/ Madison,
5) Madison/ Hamilton, and
6) Madison/Jay,

the disputed text, which indeed belongs to Madison is presented to these six experts. The votes of experts are as follows: 1) Hamilton, 2) Madison, 3) Hamilton, 4) Madison, 5) Madison, and 6) Madison. Two votes for Hamilton, four votes Madison. If we aggregate expertise by majority vote, Madison is guessed as the author of the disputed text, which is correct.

3. NEAREST NEIGHBOR TECHNIQUE WITH MAHALANOBIS DISTANCE

The Mahalanobis distance is a measure of the distance between a point P and a distribution D, introduced by P. C. Mahalanobis in 1936. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D.

We focused on the mahalanobis distance method in the results because it is better than others according to the results derived from data sets. Introduction of mahalanobis distance enables one to take into account the interactions among different components. The principle is the shorter the mahalanobis distance two points, the higher their similarity and more likely they are from same class.

The Mahalanobis distance of an observation

$x=(x_1,x_2,x_3,\ldots x_N)^T$

from a set of observations with mean

$\mu=(\mu_1, \mu_2, \mu_3,\ldots \mu_N)^T$

and covariance matrix S is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}.$$

3.1 Success In Nearest Neighbor Technique with Mahalanobis distance

First, texts by each author is taken as set of observations. For each of these three texts, the mean μ_i, and covariance S_i, i=1,2,3 are computed. Results in Table 3. must be read as follows: when the Mahalanobis distance $D_M(x)$ of the data vectors for paragraphs from the three texts to the three sets are computed, all of the Hamilton paragraphs and Madison paragraphs are nearest to their data sets as expected. Only 5% of Joy paragraphs are found nearest to the Madison dataset. When the Mahalanobis distance of paragraphs of the disputed text to the three data sets are computed, all of them are found nearest to the Madison dataset. That is the author of all paragraphs of the disputed text is Madison, which is exactly correct.

Table 3. Mahalanobis distance results

"o"	"Hamilton"	"joy"	"Madison"	"correct"
"Hamilton"	100	0	0	100
"joy"	0	95	5	95
"Madison"	0	0	100	100
"dMadison"	0	0	100	100

4. ARTIFICIAL NEURAL NETWORKS

Feed-forward and recurrent networks are the two categories in which neural networks that can be split. The flow of data is from input precisely to output cells which can be grouped into layers is what happens in feed-forward networks in which feedback interconnections cannot occur. On the other side, feedback loops are included in recurrent networks of which dynamical properties get so significant.

Author attribution analysis that was performed within research presented in this paper can be seen as the multistage process. Firstly training and testing data were selected. Two papers from each writer and one disputed paper were chosen and calculation for textual descriptors was done. Network with architecture and learning method was designed. After training and testing was done,

disputed paper was used as testing data and result was obtained.

The number of inputs equaled the number of textual descriptors used, thus it is 37. When PCA applied then 20 inputs were used. There is one hidden layer with the same number neurons with input size.

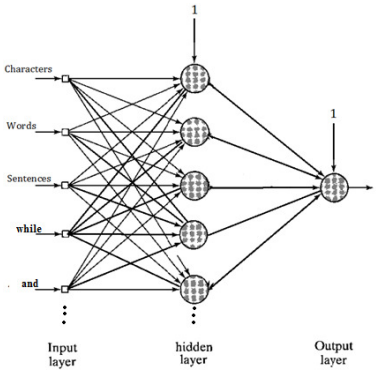

Figure 2. Signal flow graph of the machine

Two types of input used, with PCA and without PCA.

4.1 PCA: Principal Component Analysis

In order to determine the linear combination of variables which accounts for the variations in the data best, Burrows method of principle component analysis (PCA) includes basically computing the frequency of each of a list of function words as well as performing principle component analysis (PCA).The transformed data are basically planned instead of analyzing the result as statistically. For examining for trends that happen as clusters of points, the two-dimensional plots of the first two principal components provide us so. Cluster analysis might come after this step afterwards.

This technique is used for Data Compression algorithms because it reduces the dimensionality of the input. It rotates and scales the data set, which is represented in the vector space. Its goal is to understand the relationships and correlations in a data set.

The application of artificial neural networks with output neurons competing on the data of first principal components is what the author attribution might be pointed out by rather than cluster analysis of the two dimensional plots in the paper. Then, for the textual descriptors for the texts, matrices of sample covariances are calculated.

In order to determine a set of new variables as a linear combination of the original variables in the data matrices, the information is applied in the covariance matrix. Into a decreasing order of importance are derived the new variables that among them the first is called first principal component which accounts for the variation in the original data as much as possible. Second principal component that is the second of them accounts for another, rather smaller portion of the variation and continues in the same way. P components are needed in order to covering all of the variation regarding the original data, if there are p variables. However, most of the variation is covered frequently with a smaller amount of components. PCA is getting its aims the interpretation of the variation and data reduction, in this way.

Training data consists of three writers' one article. Also disputed table (no.14) is the main testing data. After using data in that way, training data changed. All training data and another article of each writer's combined and used as combined data. Disputed paper used as only testing data. The initial values of the synaptic weights and thresholds were picked at random from a uniform distribution inside the range [-1,1].

4.2 ANN Results

After training ANN with the texts authored by these three authors, we send test data from the texts authored by these three authors to ANN. All of the Hamilton paragraphs are identified correctly, while 3% of Joy and Madison paragraphs are misidentified as Hamilton paragraphs. This shows that ANN is trained well.

Table 4. Accuracy of ANN in disputed paper

"o"	"Hamilton"	"Joy"	"Maddison"	"correct"
"Hamilton"	100	0	0	100
"Joy"	3	97	0	97
"Maddison"	3	0	97	97
"dispute"	30	15	55	55

When the paragraphs from the disputed text are exposed to ANN, as seen in Table 4., 30% of paragraphs are misidentified as Hamilton paragraphs. 15% of them are misidentified as Joy paragraphs. But 55% of paragraphs are correctly identified as Madison paragraphs. Since all paragraphs belong to the same author, this author most probably is Madison, which is correct.

When 20 principal components replaced the original data, one gets slightly worst results as expected.

Table 1. Success of accuracy in disputed paper when 20 principal components replace the 47 dimensional real data

"o"	"Hamilton"	"Joy"	"Maddison"	"correct"
"Hamilton"	100	0	0	100
"Joy"	3	97	0	97
"Maddison"	3	0	97	97
"dispute"	40	15	45	45

Maddison is again the winner author, and disputed text is correctly identified, but the second author Hamilton is very close to the winner. Data is compressed, and price is paid.

5. CONCLUSION

Mosteller and Wallace (1964) concluded that Madison wrote the disputed Federalist papers. Our approach lead to the same conclusion for chosen disputed paper and this is what needs to be done in our research. Promising results obtained in this paper shows that we can proceed to find

authors of shorter texts, i. e. e-mail messages. As a future work, to deal with short e-mail messages, feature selection phase must be taken very seriously.

REFERENCES

Holmes, D. I. (1998) "The evolution of stylometry in humanities scholarship." Literary and Linguistic Computing 13(3): pp. 111–117.

McCune, B., Grace, J. B., and Dean L. (2008) Urban Analysis of Ecological Communities, MjM Software Design, p. 45.

Holmes, D. I., Forsyth, R. S. (1995) "The Federalist Revisited: New Directions in Authorship Attribution." Oxford Journals, Volume 10, Issue 2, pp. 111-127.

Greenacre, M. (2008) "Measures of distance between samples." http://www.econ.upf.edu/~michael/stanford/

Burrows, J. (1992) "Not unless you ask nicely: The interpretative nexus between analysis and information." Literary and Linguistic Computing 7(2), pp. 91–109.

Rudman, J. (2012) "The Twelve Disputed 'Federalist' Papers: A Case for Collaboration." Digital Humanities.

Schulz, J. (2007) "Bray-Curtis Dissimilarity." Retrieved from http://www.code10.info/

Teknomo, K. (2015) "Similarity Measurement." http://people.revoledu.com/kardi/tutorial/Similarity/BrayCurtisDistance.html