Mathematical and Computational Applications, Vol. 19, No. 1, pp. 21-36, 2014

AUTHORSHIP ATTRIBUTION USING PRINCIPAL COMPONENT ANALYSIS AND COMPETITIVE NEURAL NETWORKS

Mehmet Can

International University of Sarajevo, Faculty of Engineering and Natural Sciences
Hrasnićka Cesta 15, 71000 Sarajevo, Bosnia and Herzegovina
mcan@ius.edu.ba

Abstract- Feature extraction is a common problem in statistical pattern recognition. It refers to a process whereby a data space is transformed into a feature space that, in theory, has exactly the same dimension as the original data space. However, the transformation is designed in such a way that the data set may be represented by a reduced number of "effective" features and yet retain most of the intrinsic information content of the data; in other words, the data set undergoes a dimensionality reduction. Principal component analysis is one of these processes. In this paper the data collected by counting selected syntactic characteristics in around a thousand paragraphs of each of the sample books underwent a principal component analysis. Authors of texts identified by the competitive neural networks, which use these effective features.

Key Words- principal components, authorship attribution, stylometry, text categorization, stylistic features, syntactic characteristics, multilayer preceptor, competitive learning, artificial neural network.

1. INTRODUCTION

Problems of authorship have always been attacked with traditional research methods: unearthing and dating original manuscripts, for instance. But since the late 19th century, statisticians have developed “non-traditional” tools that attempt to discern quantifiable patterns within a text or corpus, with the hope that these features will help to reliably identify different authors.

The origin of non-traditional authorship attribution, or stylometry, is often said to be Augustus de Morgan’s suggestion in 1851 that certain authors of the Bible might be distinguishable from one another if one used longer words [1]. In 1887, searching for a characteristic difference in the distribution of different-sized words in writings of different languages and presentation styles, Mendenhall began investigating this hypothesis. In 1901, he turned his methods to Shakespeare, Bacon and Marlowe, and found that while Shakespeare and Marlowe were nearly indistinguishable, they were both significantly and consistently different from Bacon [2]. The difference was mainly observed in the relative frequency of three- and four-letter words: Shakespeare used more four - letter words and Bacon more three-letter words.

Authorship studies also began independently around the same time in Russia with Morozov [3]. In the West, it took 30 years or so for Mendenhall’s studies to be resumed by other linguists. G. Zipf examined word frequencies and determined not a stylometric but a universal law of language, Zipf’s Law: that the statistical rank of a word varies inversely to its frequency [4]. G. U. Yule devised a feature known as

"Yule's characteristic K," which estimated 'vocabulary richness' by comparing word frequencies to that expected by a Poisson distribution, but like Mendenhall's word lengths, this too was later found to be an unreliable marker of style [1]. In fact, most of the measurements proposed in this period proved unhelpful: among others, researchers tried average sentence length, number of syllables per word, and other estimates of vocabulary richness such as Simpson's D index and a simple type/token ratio, a ratio of the number of unique words, or types, to the number of total words, or tokens [5].

The needed breakthrough came at last in 1963 with Mosteller and Wallace's study on the Federalist Papers. In 1787 and 1788, J. Jay, A. Hamilton and J. Madison collectively wrote 85 newspaper essays supporting the ratification of the constitution. Published under the pseudonym "Publius," the authors later revealed which of the Federalist Papers they had written; however, while authorship of 67 were undisputed, 12 were claimed by both Hamilton and Madison. Mosteller and Wallace hoped to characterize each author's style through their choice of function words, such as "to," "by," and so forth. Function words are regarded as good markers of style because they are assumed to be unconsciously generated and independent of semantics, the meaning, or what the author is trying to convey. That is, an author may have a preference for modes of expression, for instance, the active vs. the passive voice that emphasize certain function words, and the same broad set of function words will be used regardless of the topic at hand [4].

Despite the fact that Hamilton and Madison have otherwise very similar styles, nearly identical sentence length distributions, as noted by Juola [5], Mosteller and Wallace found sharp differences in their preference for different function words: for instance, the word "upon" appears 3.24 times per 1000 words in Hamilton, and just 0.23 times in Madison [1]. Adjusting these frequencies with a Bayesian model, they showed that Madison had most likely written all 12 disputed papers. Traditional scholarship had already long come to the same conclusion, but Mosteller and Wallace's conclusion was independent, and thus a great achievement of the then quite exploratory field of stylometry. The Federalist Papers problem is still regarded as a very difficult test case, and as an unofficial benchmark it has been used to test most methods of authorship attribution developed since then [6-9**]**.

In the history of authorship studies, it is proved that Burrows method of principle component analysis (PCA) [10] is very efficient to remove the redundant data dimensions. In the next section this technique is going to be elaborated.

2. PRINCIPAL COMPONENT ANALYSIS

The principle component analysis (PCA) essentially involves computing the frequency of each of a list of function words, and performing principle component analysis (PCA) to find the linear combination of variables that best accounts for the variations in the data. Rather than analyze this result statistically, the transformed data are simply plotted. Two-dimensional plots of the first two principal components supply us with a means to inspect visually for trends, which occur as clusters of points [1]. Later, cluster analysis may follow this step.

This simple but effective method continues to be used today, partly because of the ease with which the results are communicated and interpreted. For example,

Binongo [11] used this method to study the problem of the authorship of L. Frank Baum's last book, which historians had long suspected of being mostly the work of Baum's successor, Ruth P. Thompson. He confirmed this suspicion independently, demonstrating that Thompson was much more prone to use position words such as "up," "down," "over," and "back," than Baum. This was not demonstrated using complex statistical techniques; rather, function word frequencies were tallied, the authors' tallies compared, PCA used to reduce the dimensionality of the data, and the resulting plots inspected: the two authors' works form obvious clusters. Similar procedures can be found in [7], [12-13].

In this paper instead of cluster analysis of the two dimensional plots, the author attribution will be found by the use of artificial neural networks with output neurons competing on the data of first principal components.

2.1 Theory of Principal component Analysis

Multivariate statistics deals with the relation between several random variables. The sets of observations of the random variables are represented by a multivariate data matrix $\boldsymbol{X}$,

Multivariate statistics deals with the relation between several random variables. The sets of observations of the random variables are represented by a multivariate data matrix $\boldsymbol{X}$,

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}. \tag{1}$$

Each column vector $\boldsymbol{u}_k$ represents the data for a different variable. If $\boldsymbol{c}$ is an $p \times 1$ matrix, then

$$\boldsymbol{Xc} = c_1 \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ \vdots \\ x_{n1} \end{bmatrix} + c_2 \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \\ \vdots \\ x_{n2} \end{bmatrix} + \cdots + c_p \begin{bmatrix} x_{1p} \\ x_{2p} \\ x_{3p} \\ \vdots \\ x_{np} \end{bmatrix} \tag{2}$$

is a linear combinations of the set of observations.

Descriptive statistics can also be applied to a multivariate data matrix $\boldsymbol{X}$, the sample mean of the *k*th variable is

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}\,, k = 1,2,\dots,p, \tag{3}$$

the sample variance is defined by

$$s_k^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2\,, k = 1,2,\dots,p. \tag{4}$$

Next we introduce a matrix that contains statistics that relate pairs of variables (x_i, x_k), sample covariance s_{ik}:

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right), i = 1,2,\dots,p\,, k = 1,2,\dots,p. \tag{5}$$

It follows that $s_{ik} = s_{ki}$ and $s_{ii} = s_i^2$, the sample variance.

Matrix of sample covariances

$$\boldsymbol{S_n} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ s_{31} & s_{32} & \cdots & s_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \quad (6)$$

is symmetric.

THEOREM Let $\boldsymbol{S_n}$ be the $p \times p$ covariance matrix related to the multivariate data matrix $\boldsymbol{X}$. Let eigenvalues of $\boldsymbol{S_n}$ be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, and corresponding orthonormal eigenvectors be $\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_p}$. Then ith principal component $\boldsymbol{y}_i$ is given by the linear combination of the original variables in the data matrix $\boldsymbol{X}$ [14]:

$$\boldsymbol{y}_i = \boldsymbol{X}\boldsymbol{u_i}, i = 1,2,..,p. \quad (7)$$

The variance of $\boldsymbol{y}_i$ is λ_i, and $\text{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = 0, i \neq j$. The total variance of the data in $\boldsymbol{X}$ is equal to the sum of eigenvalues:

$$\sum_{j=1}^{p} s_{jj} = \sum_{j=1}^{p} \lambda_j. \quad (8)$$

$$\textit{Proportion of the total variance covered by the kth principal component} = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}. \quad (9)$$

If a large percentage of the total variance can be attributed to the first few components, then these new variables can replace the original variables without significant loss of information. Thus we can achieve significant reduction in data.

In this paper author attribution is considered a classification task, and the most popularly used type of neural networks employed in pattern classification tasks is the feedforward network. Hence, some basic information about artificial neural networks with some stress on feedforward networks will be included as in the following.

3. ARTIFICIAL NEURAL NETWORKS

Nervous systems existing in biological organism for years have been the subject of studies for mathematicians who tried to develop some models describing such systems and all their complexities. Artificial Neural Networks emerged as generalizations of these concepts with mathematical model of artificial neuron due to McCuloch and Pitts described in [15], and the first implementation of Rosenblatt's perceptron in [16]. The efficiency and applicability of artificial neural networks to computational tasks have been questioned many times, especially at the very beginning of their history the book "Perceptrons" by Minsky and Papert [17] caused dissipation of initial interest and enthusiasm in applications of neural networks. It was not until 1970s and 80s, when the backpropagation algorithm for supervised learning was documented that artificial neural networks regained their status and proved beyond doubt to be sufficiently good approach to many problems.

3.1. Multilayer Perceptrons

Multilayer perceptrons have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with a highly popular algorithm known as the error back-propagation algorithm. This algorithm is based on

the error - correction learning rule. As such, it may be viewed as a generalization of an equally popular adaptive filtering algorithm: the ubiquitous least-mean-square (LMS) algorithm.

From architecture point of view neural networks can be divided into two categories: feed-forward and recurrent networks. In feed-forward networks the flow of data is strictly from input to output cells that can be grouped into layers but no feedback interconnections can exist. On the other hand, recurrent networks contain feedback loops and their dynamical properties are very important.

The most popularly used type of neural networks employed in pattern classification tasks is the feedforward network which is constructed from layers and possesses unidirectional weighted connections between neurons. The common examples of this category are Multilayer Perceptron or Radial Basis Function networks, and committee machines.

Multilayer perceptron type is more closely defined by establishing the number of neurons from which it is built, and this process can be divided into three parts, the two of which, finding the number of input and output units, are quite simple, whereas the third, specification of the number of hidden neurons can become crucial to accuracy of obtained classification results.

The number of input and output neurons can be actually seen as external specification of the network and these parameters are rather found in a task specification. For classification purposes as many distinct features are defined for objects which are analyzed that many input nodes are required. The only way to better adapt the network to the problem is in consideration of chosen data types for each of selected features. For example instead of using the absolute value of some feature for each sample it can be more advantageous to calculate its change as this relative value should be smaller than the whole range of possible values and thus variations could be more easily picked up by Artificial Neural Network. The number of network outputs typically reflects the number of classification classes.

The third factor in specification of the Multilayer Perceptron is the number of hidden neurons and layers and it is essential to classification ability and accuracy. With no hidden layer the network is able to properly solve only linearly separable problems with the output neuron dividing the input space by a hyperplane. Since not many problems to be solved are within this category, usually some hidden layer is necessary.

With a single hidden layer the network can classify objects in the input space that are sometimes and not quite formally referred to as simplexes, single convex objects that can be created by partitioning out from the space by some number of hyperplanes, whereas with two hidden layers the network can classify any objects since they can always be represented as a sum or difference of some such simplexes classified by the second hidden layer.

Apart from the number of layers there is another issue of the number of neurons in these layers. When the number of neurons is unnecessarily high the network easily learns but poorly generalizes on new data. This situation reminds auto-associative property: too many neurons keep too much information about training set rather "remembering" than "learning" its characteristics. This is not enough to ensure good generalization that is needed.

On the other hand, when there are too few hidden neurons the network may never learn the relationships amongst the input data. Since there is no precise indicator how many neurons should be used in the construction of a network, it is a common practice to build a network with some initial number of units and when it trains poorly this number is either increased or decreased as required. Obtained solutions are usually task-dependant.

For the purposes of this research¸ a neural network with fifty input terminals and an output layer with six competing neurons is chosen.

3.2. Competitive Learning

In order to produce the desired set of output states whenever a set of inputs is presented to a neural network it has to be configured by setting the strengths of the interconnections and this step corresponds to the network learning procedure. Learning rules are roughly divided into three categories of supervised, unsupervised and reinforcement learning methods.

In competitive learning, as the name implies the output neurons of a neural network compete among themselves to become active (fired). Whereas in a neural network based on Hebbian learning several output neurons may be active simultaneously, in competitive learning only a single output neuron is active at any one time. It is this feature that makes competitive learning highly suited to discover statistically salient features that may be used to classify a set of input patterns [18].

There are three basic elements to a competitive learning rule [19]:

1. A set of neurons that are all the same except for some randomly distributed synaptic weights, and which therefore respond differently to a given set of input patterns.
2. A limit imposed on the "strength" of each neuron.
3. A mechanism that permits the neurons to compete for the right to respond to a given subset of inputs, such that only one output neuron or only one neuron per group is active (i.e., "on") at a time. The neuron that wins the competition is called a winner-takes-all neuron.

Accordingly the individual neurons of the network learn to specialize on ensembles of similar patterns; in so doing they become feature detectors for different classes of input patterns.

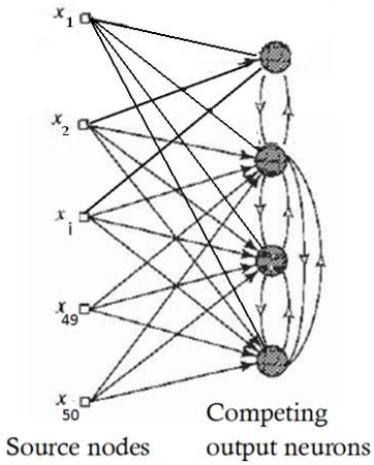

Figure 1. Architectural graph of a simple competitive learning network with feedforward (excitatory) connections from the source nodes to the neurons, and lateral (inhibitory) connections among the neurons; the lateral connections are signified by open arrows.

In the simplest form of competitive learning, the neural network has a single layer of output neurons; each of which is fully connected to the input nodes. The network may include feedback connections among the neurons, as indicated in Figure 1. In the network architecture described herein, the feedback connections perform lateral inhibition, with each neuron tending to inhibit the neuron to which it is laterally connected. In contrast, the feedforward synaptic connections in the network of Figure 6. are all excitatory.

For a neuron *k* to be the winning neuron, it's induced local field $\boldsymbol{v_k}$, for a specified input pattern $\boldsymbol{x}$ must be the largest among all the neurons in the network. The output signal $\boldsymbol{y_k}$ of winning neuron *k* is set equal to one; the output signals of all the neurons that lose the competition are set equal to zero. We thus write

$$\boldsymbol{y_k} = \begin{cases} 1, if\ \boldsymbol{v_k} > \boldsymbol{v_j}\ for\ all\ j, j \neq k \\ 0, \qquad\qquad otherwise \end{cases} \tag{10}$$

where the induced local field $\boldsymbol{v_k}$ represents the combined action of all the forward and feedback inputs to neuron *k*.

Let $\boldsymbol{w_{kj}}$ denote the synaptic weight connecting input node *j* to neuron *k*. Suppose that each neuron is allotted a fixed amount of synaptic weight (i.e., all synaptic weights are positive), which is distributed among its input nodes; that is,

$$\sum_j \boldsymbol{w_{kj}} = 1 \text{ for all } k \tag{11}$$

A neuron then learns by shifting synaptic weights from its inactive to active input nodes. If a neuron does not respond to a particular input pattern, no learning takes place in that neuron. If a particular neuron wins the competition, each input node of that neuron relinquishes some proportion of its synaptic weight, and the weight relinquished is then distributed equally among the active input nodes. According to the standard competitive learning rule, the change Δw_{kj} applied to synaptic weight w_{kj} is defined by

$$\Delta w_{kj} = \begin{cases} \eta(\boldsymbol{x_j} - w_{kj}) & if\ neuron\ k\ wins\ the\ competition \\ 0, & if\ neuron\ k\ loses\ the\ competition \end{cases} \quad (12)$$

where η is the learning-rate parameter. This rule has the overall effect of moving the synaptic weight vector $\boldsymbol{w_k}$ of winning neuron *k* toward the input pattern $\boldsymbol{x}$.

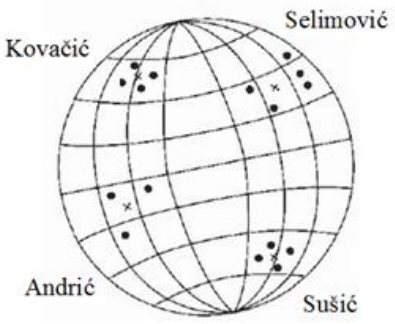

Figure 2. Geometric interpretation of the competitive learning process. The dots represent the input vectors, and the crosses represent the synaptic weight vectors of the four output neurons.

We may use the geometric analogy depicted in Figure 2. to illustrate the essence of competitive learning[19]. In our case, each of 50 dimensional input pattern vector $\boldsymbol{x}$ has unit Euclidean length so that we may view it as a point on an *N*-dimensional unit sphere where *N=50* is the number of input nodes. *N* also represents the dimension of each synaptic weight vector $\boldsymbol{w_k}$. It is further assumed that all neurons in the network are constrained to have the same unit Euclidean length (norm), as shown by

$$\sum_j \boldsymbol{w^2_{kj}} = 1 \text{ for all } k \quad (13)$$

When the synaptic weights are properly scaled they form a set of vectors that fall on the same *N*-dimensional unit sphere. In Figure 2. we show four natural groupings (clusters) of the stimulus patterns represented by dots.

This figure also includes state of the network that results from the use of competitive learning. In particular, each output neuron has discovered a cluster of input patterns by moving its synaptic weight vector to the center of gravity of the discovered cluster. This figure illustrates the ability of a neural network to perform clustering through competitive learning. However, for this function to be performed in a "stable" fashion the input patterns must fall into sufficiently distinct groupings to begin with. Otherwise the network may be unstable because it will no longer respond to a given input pattern with the same output neuron.

4. PROBLEM DEFINITION

In this paper author attribution is considered as an application of principal component analysis, and as a classification task [20-21]. Texts studied are literary works of five ex Yugoslavian writers, Ivo Andrić (1892-1975) – Cuprija na Drini[22], Znanovi [23], Proklet Avlija [24], M. Meša Selimović (1910-1982) – Derviš i Smrt [25], Tvrdjava [26] , Derviš Sušić (1925 – 1990) – Pobune [27], and Ante Kovačić- U registraturi [28].

Features selected to describe texts are lexical and syntactical components that show promising results when used as writer invariants because they are used rather subconsciously and reflect the individual writing style which is difficult to be copied. Principal components of data elicited from texts possess generalization properties that allow for the required high accuracy of classification [29].

The novels selected provide the corpora which are wide enough to make sure that characteristic features found based on the training data can be treated as representative of other parts of the texts and this generalized knowledge can be used to classify the test data according to their respective authors.

Obviously literary texts can greatly vary in length; if the lengths of texts differ essentially, learning levels of the training sets will differ, and this may affect test results. What is more, all stylistic features can be influenced not only by different timelines within which the text is written but also by its genre. The first of these issues is easily dealt with by dividing long texts, such as novels, into some number of smaller parts of approximately the same size.

Described approach gives additional advantage in classification tasks as even in case of some incorrect classification results of these parts the whole text can still be properly attributed to some author by based the final decision on the majority of outcomes instead of all individual decisions for all samples. Whether the genre of a novel is reflected in lexical and syntactic characteristics of it is the question yet to be answered. Hence all together we have selected thousands of paragraphs from "Na Drini Ćuprija, Znakovi Pored Puta, Prokleta Avlija " by Ivo Andrić, "Derviš i Smrt, Tvrdjava" by M. Meša Selimović, “Pobune” by Derviš Sušić, U registraturi by Ante Kovačić.

Feature Selection

Extracting lexical features that work as effective discriminators of texts under study is one of critical issues in authorship analysis. In this research fourteen textual descriptors are used, average sentence length, average word length, number of words, sentences, commas, and conjecture “and”, in Bosnian “i”, and other characteristics in paragraphs listed in the first column of Table 1a, and 1b. Means and variances of the textual descriptors for the texts Ivo Andrić: Na Drini Ćuprija, M. Meša Selimović: Derviš i Smrt, and Derviš Sušić: Pobune are shown in Table 1a, and Ivo Andrić: Proklet Avlja, M. Meša Selimović: Tvrdzava, and Ante Kovacevic: U Registraturi in Table 1b as samples for comparison.

Table 1a. Paragraph averages and variances of the textual descriptors for three of the books used in this research

	Na Drini Ćuprija		Derviš		Pobune	
Textual descrs.	Mean	Variance	Mean	Variance	Mean	Variance
Sentence length	84.331	2090.92	58.710	2053.855	33.0478	1337.3416
Word length	2.157	2.877	2.155	3.460	2.5459	3.0985
Word count	79.208	5861.724	60.362	4756.432	24.5825	1040.4906
Sentence count	4.395	16.886	5.012	29.411	3.4843	17.0118
Comma count	6.432	45.95	7.130	87.211	2.6660	16.4196
dots count	0.052	0.135	0.002	0.002	0.2526	0.6327
i count	5.375	35.072	2.235	9.659	0.6910	1.8709
ili count	0.250	0.514	0.302	0.688	0.09390	0.1397

je count	2.798	11.991	2.552	11.531	0.6305	1.8402
se count	1.852	4.823	1.615	4.478	0.6221	1.2021
pa count	0.140	0.216	0.098	0.133	0.0731	0.0846
da count	1.935	6.853	2.262	9.613	0.8601	2.334
ne count	0.637	1.695	0.968	2.718	0.4196	0.6708
kao poput count	0.662	1.106	0.480	1.007	0.0793	0.1192
Total		8080.760		6970.200		2423.2562

Table 1b. Paragraph averages and variances of the textual descriptors for other three books used in this research (continued)

	Proklet Avlja		Tvirdjava		U Registraturi	
Textual descrs.	Mean	Variance	Mean	Variance	Mean	Variance
Sentence length	297.18	78387.76	127.58	49057.29	264.22	107368.41
Word length	361.16	115682.56	155.24	73005.05	317.71	155506.61
Word count	64.97	3650.86	28.62	2384.83	54.44	4481.05
Sentence count	4.26	15.97	2.46	10.88	7.96	111.66
Comma count	5.15	27.95	3.35	49.48	4.16	32.25
dots count	0.01	0.03	0.01	0.03	0.	0.
i count	4.27	22.32	1.04	5.44	2.57	16.24
ili count	0.25	0.81	0.09	0.13	0.09	0.14
je count	2.5	10.59	1.2	5.42	1.06	2.86
se count	1.44	3.44	0.72	2.22	1.41	3.92
pa count	0.15	0.18	0.09	0.15	0.3	0.48
da count	1.72	4.16	0.95	3.9	0.98	2.83
ne count	0.61	1.06	0.51	1.6	0.53	1.04
kao poput count	0.52	0.68	0.21	0.38	0.	0.
Total		197808		124527		267527

As it is seen, there is statistical differences between the usages of textual descriptors in texts, for instance, Ivo Andrić prefers longer paragraphs. In average Ivo Andrić ‘s paragraphs contain 79 words with variance 5861.7, while Meša Selimović’s average is 62 with variance 4756.4, and Derviš Sušić’s average is 25 with variance 1040.5.

In the next chapter the pattern captured by principal components corresponding to these data will be displayed.

5. PRINCIPAL COMPONENTS OF SAMPLE TEXTS

Next, matrices of sample covariances for the textual descriptors for the texts are computed. The information in the covariance matrix is used to define a set of new variables as a linear combination of the original variables in the data matrices X_{ivo} , X_{mesa} , etc. . The new variables are derived in a decreasing order of importance. The first of them is called first principal component and accounts for as much as possible of the variation in the original data. The second of them is called second principal component and accounts for another, but smaller portion of the variation, and so on.

If there are *p* variables, to cover all of the variation in the original data, one needs *p* components, but often much of the variation is covered by a smaller number of

components. Thus PCA has as its goals the interpretation of the variation and data reduction.

In fact PCA is nothing but the spectral decomposition of the covariance matrix.

Variances and percentage variances covered by fourteen principal components of the textual descriptors for the sample texts consisting randomly chosen 400 paragraphs of six chosen works of six authors are shown in Table 2.

Table 2. Percentages of variances covered by fourteen principal components of the textual descriptors used in this research.

Princ. Comp.	Andrić Cuprija	Selimović Derviš	Sušić Pobune	Kovačić Registraturi
1	75.600	77.112	74.580	63.700
2	24.127	22.400	24.845	35.424
3	0.083	0.204	0.200	0.588
4	0.054	0.088	0.154	0.104
5	0.032	0.048	0.073	0.058
6	0.029	0.041	0.040	0.035
7	0.022	0.029	0.033	0.030
8	0.016	0.024	0.024	0.024
9	0.014	0.022	0.019	0.019
10	0.009	0.015	0.015	0.012
11	0.008	0.010	0.006	0.006
12	0.005	0.006	0.005	0.001
13	0.002	0.002	0.004	0.000
14	0.001	0.000	0.002	0.000
Total	100	100	100	100

Table 2 reveals that the first two principal components cover more than %99 of variances of principal components.

It is seen that first principal component covers around 75% of the variance. Therefore in this article to classify the texts, we will rely on only first principal components. The interval [-500, 350]] is the common support of the first principal components for all of the data. This interval is divided into 50 equal bins. From a text we collect 400 paragraphs randomly hence it is an 400×14 matrix and when the principal components are extracted, the transformed data is also a 400×14 matrix. The first column of this matrix is the first principal component of this sample. This column has 400 entries. Their distributions to the bins are counted. We normalize these bin counts by dividing these counts by the maximum count. We repeat this procedure 500 times, and get the simple average of resulting frequency distributions. We repeat it for each text. The average frequency distributions are shown in the following figures.

Figure 3. in the below displays plot of the normalized mean of 500 normalized frequency distributions for the three books authored by Ivo Andrić: Cuprija na Drina, Znakovi Put, and Proklet Avlija.

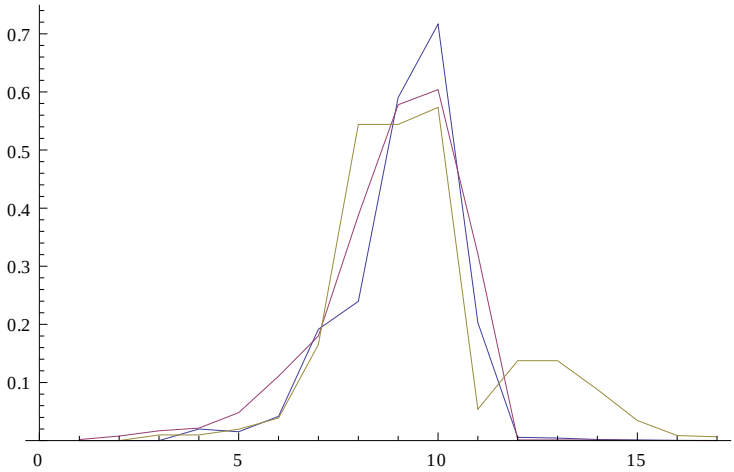

Figure 3. Plot of average frequencies in the 500 first principal components of the samples of Ivo Andrić's three books.

Figure 4. displays plot of the normalized mean of 500 normalized frequency distributions of the three books authored by M. Meša Selimović: Derviš i Smrt, and Tvrdjava. Apparently higher peaks of the first principal components are more common in Meša Selimović's works.

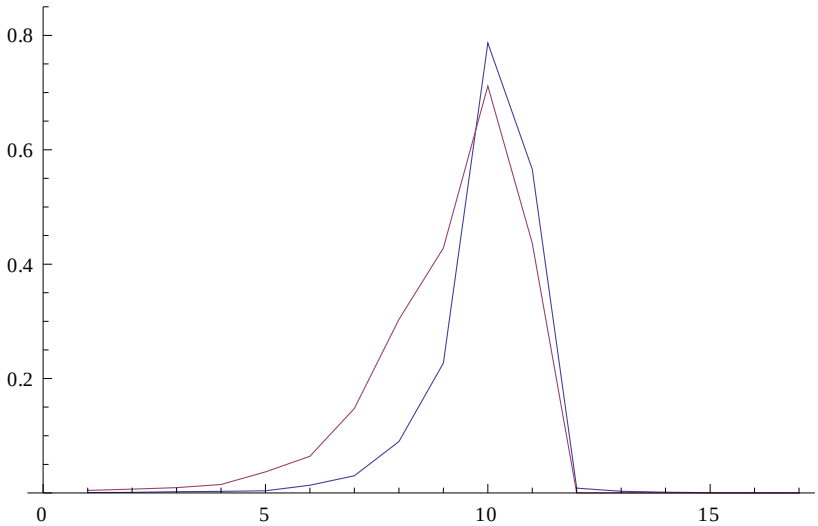

Figure 4. Plot of average frequencies in the 500 first principal components of the samples of Meša Selimović's Derviš i Smrt, and Tvrdjava (with higher peak).

Figure 5. displays plot of the normalized mean of 500 normalized frequency distributions of the two books authored by Derviš Sušić, and Ante Kovačić.

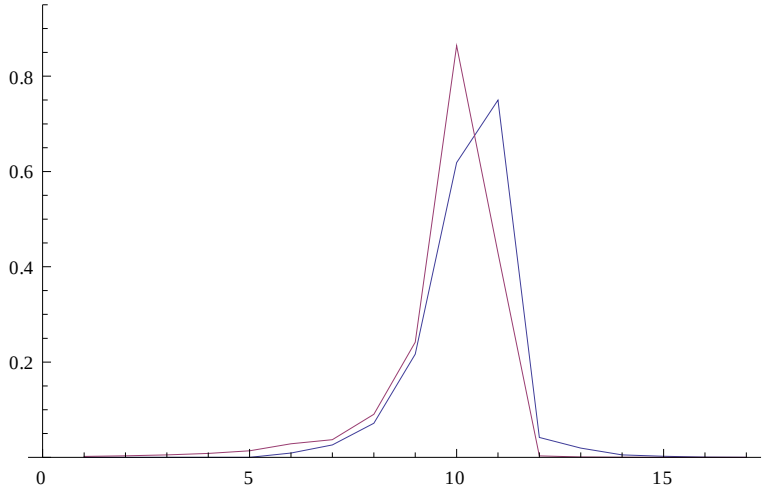

Figure 5. Plot of average frequencies in the 500 first principal components of the samples of books authored by Derviš Sušić (shortest peak), and Ante Kovačić.

The frequency profile of first principal components of the textual data seems to be invariant throughout a text. There are similarities in the frequency profiles of the text authored by the same person. Therefore these frequency profiles can be regarded as writerprints. However a visual identification of the authors of these writerprints seems to be difficult. To help the classification of these writerprints, we propose to take it as a pattern classification task, and use artificial neural networks, more specifically perceptrons with competing neurons to do the job.

6. APPLICATION TO AUTHOR ATTRIBUTION

Author identification analysis that was performed within research presented in this paper can be seen as the multistage process, as follows

- *selection of corpora;* the first step was selection of the training and testing examples,
- *selection of textual descriptors;* next stage was taken by the choice of textual descriptors to be analyzed,
- *calculation;* then followed the third phase of calculating characteristics for all descriptors,
- *principal component analysis*; transform randomly chosen data matrices into matrices with principal components,
- *calculation of frequencies in bin;* count frequencies of principal components in bins of equal length that were later used for training of the neural network,
- *selection of the neural network;* specification of the network with its architecture and learning method can be seen as the fourth step of the whole procedure,
- *training;* the fifth consisted of the actual training,
- *testing;* the sixth stage is testing,

- *analysis of obtained results;* and the final one corresponded to analysis of obtained results and coming up with some conclusions and possible indicators for improvement.

In this paper, the training phase is bypassed by simply choosing average of pattern vectors as synaptic weights for competing output neurons. The testing process is applied to ten sets of test data, with an artificial neural network of 50 input terminals, and four competing output neurons.

The input vector $\boldsymbol{x}$ is 50 dimensional with components as frequencies in corresponding bins as shown in the signal flow graph in Figure 6. Algorithm results in a decision about attribution of paragraphs whose textual description entered in the form of frequencies in bins of principal components as inputs.

Our aim is to train a neural network to distinguish paragraphs authored by four authors in a mixed text. We have chosen 500 sets of 400 paragraphs from each of the texts. Each 400 paragraph set is transformed into its principal components, and only first principal components are taken into account. Hence we have 500 first principal components from each text. Then principal components are transformed into data vectors whose elements are frequencies in 50 uniformly specified bins. The resulting data is a 500×50 matrix for each text.

Training phase is completed simply taking averages of training set as synaptic weights for competing output neurons.

Then the test data consisting of a random mixture of 500 test data from each text, totally 2000 mixed data, is sent to the neural network for classification. The correct classification numbers are as follows.

Table 3. Number of 2000 paragraphs attributed to correct authors.

	Cuprija na Drini	Derviš i Smrt	Pobune	U Registraturi	Percent.
Correct attr.	499	500	500	498	99.85%

When test data for four books are sent individually to the neural network for identification, the author attributions are as in Table 4.

Table 4. The author attributions of the seven books.

Attribution	Cuprija	Znakovi	Proklet	Derviš	Tvrdjava	Pobune	Registraturi
Andrić	499	417	496	500	0	0	0
Selimović	1	0	0	0	267	4	0
Sušić	0	83	4	0	68	496	2
Kovačić	0	0	0	0	165	0	498
Success	99.8%	83.4%	99.2%	100%	53.4%	99.2%	99.6%

As it is seen from tables above, the neural network is successful in the test data from the texts it trained for. The relative weakness for Znakovi, and Tvrdjava is due to changes in the styles of authors in time. The successes in the classification of other books of the same authors are also satisfactory.

7. CONCLUSIONS

The research described in this paper concerning author identification analysis shows that the method of principal component analysis (PCA), when followed by an artificial neural network is an efficient tool. Thus a series of future experiments should include wider range of authors, definition of new sets of textual descriptors, and test for other types and structures of neural networks, and search the possibility of inheritance through translation into other languages.

REFERENCES

1. D. Holmes, The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing* **13(3)**, 111–7, 1998.
2. C. Williams, Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon, *Biometrika* **62(1)**, 207–12, 1975.
3. O. Kukushkina, A. Polikarpov, and D. Khmelev, Using literal and grammatical statistics for authorship attribution. *Problemy Peredachi Informatsii* **37(2)**, 2002.
4. J. Smith, A review of authorship attribution, *Report*, 2008.
5. P. Juola, Authorship attribution, *Foundations and Trends in Information Retrieval* **1(3)**, 233–334, 2006.
6. B. Kjell, Authorship determination using letter pair frequency features with neural network classifiers, *Literary and Linguistic Computing* **9(2)**, 119–24, 1994.
7. D. Holmes, and R. Forsyth, The Federalist revisited: New directions in authorship attribution, *Literary and Linguistic Computing* **10(2)**, 112–27, 1995.
8. R. Bosch, and J. Smith, Separating hyperplanes and the authorship of the disputed federalist papers, *American Mathematical Monthly* **105 (7)**, 601–8, 1998.
9. G. Fung, The disputed Federalist Papers: SVM feature selection using concave minimization, *Proceedings of the 2003 Conference on Diversity in Computing*, 42–6, 2003.
10. J. Burrows, Not unless you ask nicely: The interpretative nexus between analysis and information, *Literary and Linguistic Computing* **7(2)**, 91–109, 1992.
11. J. Binongo, Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution, *Chance* **16(2)**, 9–17, 2003.
12. D. Holmes, L. Gordon, and C. Wilson, A widow and her soldier: Stylometry and the American Civil War, *Literary and Linguistic Computing* **16(4)**, 403–20, 2001.
13. R. Peng, and N. Hengartner, Quantitative analysis of literary styles, *The American Statistician* **56(3)**, 175–85, 2002.
14. B. Kolman, and D. R. Hill, *Elementary Linear Algebra*, Pearson, New Jersey, 2004.
15. W. S. Mcculloch, and W. Pitts, A Logical Calculus of the Ideas Immanent in Nervous Activity, *Bulletin of Mathematical Biophysics*, 5:115-133. Reprinted in Anderson& Rosenfeld [1988], pp. 18-28, 1943.
16. E. Rosenblatt, The Perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Review* **65**, 386-408, 1958.
17. M. L. Minsky, and S. A. Papert, *Perceptrons*, Expanded Edition. Cambridge, MA: MIT Press. Original edition, 1988.

18. A. Jamak, A. Savatic, and M. Can, Principal Component Analysis for Authorships Attribution, *Business Systems Research* **3(2)**, 49-56, 2012.
19. S. Haykin, *Neural Networks A Comprehensive Foundation*, Second Edition, Prentice-Hall, Inc., Simon & Schuster, A Viacom Company Upper Saddle River, New Jersey 07458, 1999.
20. C. Chaski, Empirical evaluations of language-based author identification techniques, *Journal of Forensic Linguistics* **8(1)**, 1–65, 2001.
21. C. Chaski, Who's at the keyboard? Authorship attribution in digital evidence investigations, *International Journal of Digital Evidence* **4(1)**, 2005.
22. I. Andrić, *Na Drini Ćuprija*, Svjetlost, Sarajevo, 1981.
23. I. Andrić, *Znakovi Pored Puta*, Svjetlost, Sarajevo, 1989.
24. I. Andrić, *Prokleta Avlija*, Svjetlost, Sarajevo, 1980.
25. M. M. Selimović, *Derviš i smrt*, Svjetlost, Sarajevo, 1966.
26. M. M. Selimović, *Tvrdjava*, Svjetlost, Sarajevo, 1970.
27. D. Sušić, *Pobune*, Veselin Masleša, Sarajevo, 1966.
28. A. Kovačić, *U registraturi*, Večernjakova bibilioteka, Zagreb 2004.
29. J. F. Hayes, Authorship Attribution: A Principal Component and Linear Discriminant Analysis of the Consistent Programmer Hypothesis, *Journal of Computational and Applied Mathematics* **15(2)**, 79-99, 2008.