Principal Component Analysis and Neural Networks for Authorship Attribution

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: A common problem in statistical pattern recognition is that of feature selection or feature extraction. Feature selection 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.

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 performed using neural networks. Then, first of the principal components are used to distinguish authors of the texts by the use of multilayer preceptor type artificial neural networks.

Keywords: principal components, authorship attribution, stylometry, text categorization, function words, classification task, stylistic features, syntactic characteristics, multilayer preceptor, artificial neural network.

1. INTRODUCTION

Authorship attribution is probably the oldest of the all text categorization problems, as old as writing itself. Although it is also possibly the least well organized disciplines, and its history is marred with the mishandling of statistical techniques, it still promises to provide useful applications in spheres as diverse as law, security, and education.

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 (Holmes 1998). In 1887, Mendenhall began investigating this hypothesis, searching for a characteristic difference in the distribution of different-sized words in writings of different languages and presentation styles. 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 (Williams 1975). 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.

However, it was later noted by Williams that this difference was more likely attributable to the different styles of composition: Mendenhall had compared Bacon's prose to the blank verse of Marlowe and Shakespeare (Williams 1975). Williams examined the prose and verse of a fourth contemporary, Sir Philip Sydney, and found they were differed in much the same way as Bacon's and Shakespeare's writings. Williams concluded that Mendenhall had misclassified Shakespeare's writings as prose. In Smith's words, "Mendenhall's method now appears to be so discredited that any serious student of authorship should discard it" (cited in Juola 2006).

Authorship studies also began independently around the same time in Russia, it seems, with Morozov proposing a model for measuring style that garnered the interest of A. Markov (Kukurushkina et al. 2002). In the West, it took 30 years or so for Mendenhall's studies to be resumed by other linguists. George 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 (Smith 2008). G. Udny 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 (Holmes 1998).

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) (Juola 2006).

With Mosteller and Wallace's study on the Federalist Papers the needed breakthrough came at last in 1963. In 1787 and 1788, John Jay, Alexander Hamilton and James 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 (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 (Smith 2008).

Despite the fact that Hamilton and Madison have otherwise very similar styles—nearly identical sentence length distributions, as noted by (Juola 2006)—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 (quoted in Holmes 1998). 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 (see, for instance, Kjell 1994, Holmes & Forsyth 1995, Bosch & Smith 1998, and Fung 2003).

2 PROBLEM DEFINITION

In this paper author attribution is considered as an application of principal component analysis, and as a classification task (Chaski, C. 2001 , 2005). Texts studied are literary works of three Bosnian writers, Ivo Andrić (1892-1975) , M. Meša Selimović (1910-1982), and Derviš Sušić (1925 – 1990). Feature 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 (Hayes 2008).

The novels authored by Ivo Andrić, M. Meša Selimović, and Derviš Sušić 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; 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, and “Pobune” by Derviš Sušić.

2.2 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 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 1. 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 1.

Table 1. Paragraph averages and variances of the textual descriptors used in this research

	Na Drini Ćuprija	
Textual descrs.	Mean	Variance
Sentence length	84.331	2090.92
Word length	2.157	2.877
Word count	79.208	5861.724
Sentence count	4.395	16.886
Comma count	6.432	45.95
dots count	0.052	0.135
i count	5.375	35.072
ili count	0.250	0.514
je count	2.798	11.991
se count	1.852	4.823
pa count	0.140	0.216
da count	1.935	6.853
ne count	0.637	1.695
kao poput count	0.662	1.106
Total		8080.760

	Derviš	
Textual descrs.	Mean	Variance
Sentencelength	58.710	2053.855
Word length	2.155	3.460
Word count	60.362	4756.432
Sentenc count	5.012	29.411
Comma count	7.130	87.211
dots count	0.002	0.002
i count	2.235	9.659
ili count	0.302	0.688
je count	2.552	11.531
se count	1.615	4.478
pa count	0.098	0.133
da count	2.262	9.613
ne count	0.968	2.718
kao poput	0.480	1.007
Total		6970.200

	Pobune	
Textualdescrs.	Mean	Variance
Senten length	33.0478	1337.3416
Word length	2.5459	3.0985
Word count	24.5825	1040.4906
Sentencecount	3.4843	17.0118
Comma count	2.6660	16.4196
dots count	0.2526	0.6327
i count	0.6910	1.8709
ili count	0.09390	0.1397
je count	0.6305	1.8402
se count	0.6221	1.2021
pa count	0.0731	0.0846
da count	0.8601	2.334
ne count	0.4196	0.6708
kaopoput	0.0793	0.1192
Total		2423.2562

As it is seen, there is statistical differences between the usage 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.

3 PRINCIPAL COMPONENT ANALYSIS

The methods of Mosteller and Wallace have proved as enduring as the problem they investigated: they were only modestly altered when Burrows described his method of stylometric analysis in a series of papers published in the late 1980s and early 1990s (Holmes 1998; see, for instance, Burrows 1992). The Burrows method essentially involves computing the frequency of each of a list of function words (larger than that of Mosteller and Wallace), 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 (Holmes 1998). 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 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 (Binongo 2003). 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 (Holmes & Forsyth 1995, Holmes et al. 2001, and Peng & Hentgartner 2002).

3.1 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 $\boldsymbol{X}$,

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

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

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

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 kth variable is

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

the sample variance is defined by

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

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

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

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 . Then ith principal component is given by the linear combination of the original variables in the data matrix $\boldsymbol{X}$ (Kolman and Hill 2004):

(7)

The variance of is , and

The total variance of the data in $\boldsymbol{X}$ is equal to the sum of eigenvalues:

(8)

—— (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.

4. PRINCIPAL COMPONENTS OF SAMPLE TEXTS

Next, matrices of sample covariances for the textual descriptors for the texts authored by Ivo Andrić: Na Drini Ćuprija, and M. Meša Selimović: Derviš i Smrt, and for other four books 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 $\boldsymbol{X}_{tva}$, and $\boldsymbol{X}_{mesa}$. 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 Ivo Andrić: Na Drini Ćuprija, M. Meša Selimović: Derviš i Smrt, and Derviš Sušić: Pobune are shown in Table 2.

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

	Na Drini Ćuprija	
Princ. Comp.	Variance	% Variance covered
1	7447.1542	75.600
2	2376.6700	24.127
3	8.1300	0.083
4	5.3100	0.054
5	3.1990	0.032
6	2.8112	0.029
7	2.1528	0.022
8	1.5691	0.016
9	1.3450	0.014
10	0.8301	0.009
11	0.7779	0.008
12	0.4775	0.005
13	0.1486	0.002
14	0.0742	0.001
	8080.7600	100
	Derviš	

Princ. Comp.	Variance	% Variance covered
1	5374.7584	77.112
2	1561.3045	22.400
3	14.2116	0.204
4	6.1523	0.088
5	3.3358	0.048
6	2.8844	0.041
7	2.0270	0.029
8	1.6440	0.024
9	1.5300	0.022
10	1.0789	0.015
11	0.7001	0.010
12	0.4516	0.006
13	0.1167	0.002
14	0.0024	0.000
	6970.2000	100

	Pobune	
Princ. Comp.	Variance	% Variance covered
1	1796.9083	74.5801
2	598.6175	24.8454
3	4.8226	0.2002
4	3.7082	0.1539
5	1.74690	0.0725
6	0.9694	0.0402
7	0.7969	0.0331
8	0.5848	0.0243
9	0.4671	0.0194
10	0.3633	0.0151
11	0.1363	0.0057
12	0.1178	0.0049
13	0.0895	0.0037
14	0.0391	0.0016
	2409.3677	100

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

It is found that the interval [0, 500] covers the support of first principal components of all 400 paragraph random samples and of all texts, while for the second principal components the region is [-100, 300]. The interval [0, 500] is divided into 25 bins, and frequencies of the data in the principal components are counted. The same is done for the second principal components of the texts. Then these two data combined as coordinates of points in the two dimensional Euclidean plane.

Figure 1. and Figure 2. In the below displays two dimensional plot of the 100 different random samples of Cuprija na Drina data and Derviš i Smrt data. Apparently lower values of the first principal component are more common in Derviš i Smrt data. These figures are writerprints of Ivo Andrić, and M. Meša Selimović respectively.

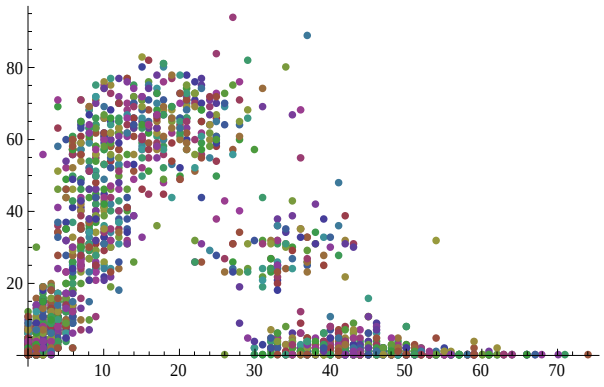

Figure 1. Points representing frequencies in the first and second principal components of Ivo Andrić; Cuprija data in the two dimensional Euclidean plane.

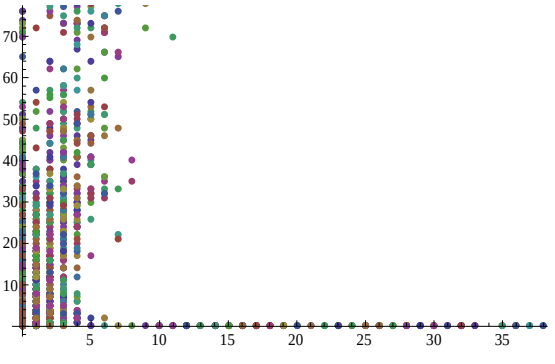

Figure 2 Points representing frequencies in the first and second principal components of Meša Selimović; Derviš i Smrt data data in the two dimensional Euclidean plane.

To check whether these patterns are writing prints of the two authors, two more books of Ivo Andrić; Znakovi Pored Puta, and Proklet Avlija, and one other book of Meša Selimović; Tvrdjeva, as well as Pobune authored by a third novelist Derviš Sušić are investigated.

The comparison of the frequencies in the first principal components of the three books authored by Ivo Andrić: Cuprija na Drina, Znakovi Pored Puta, Proklet Avlija are shown in Figure 3 below. The writing print of Ivo Andrić is the lower peaks – less than 70 – at the lowermost values of the principal components.

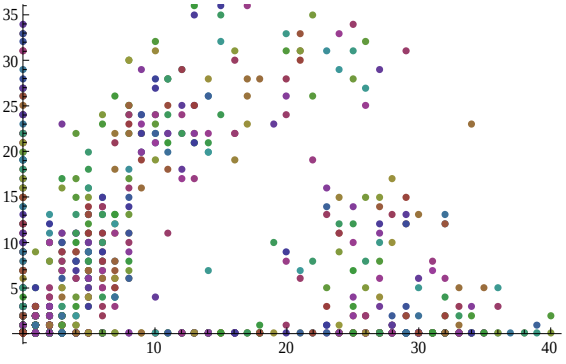

Figure 3. Points representing frequencies in the first and second principal components of Ivo Andrić: Znakovi Pored Puta data in the two dimensional Euclidean plane.

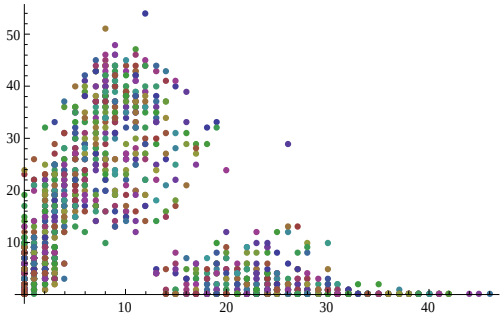

Figure 4. Points representing frequencies in the first and second principal components of Ivo Andrić: Proklet Avlija data in the two dimensional Euclidean plane.

Points representing frequencies in the first and second principal components of the other book authored by Meša Selimović; Tvrdjeva is shown in Figure 5. The writing print of Meša Selimović is revealed as twice higher peaks compared to the corresponding Ivo Andrić peaks.

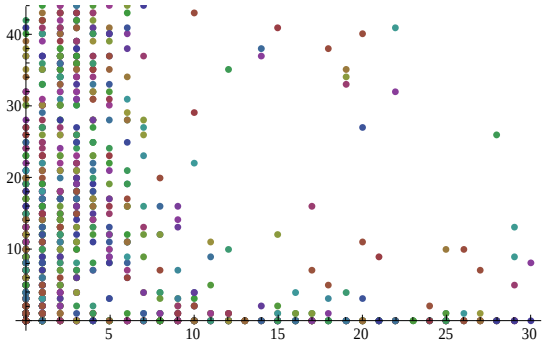

Figure 5. Points representing frequencies in the first and second principal components of the other book authored by Meša Selimović; Tvrdjeva.

To show the author specific character of these patterns, we compare them with a third author's text Pobune (Sušić 1966). Pattern of the points representing frequencies in the first and second principal components of Pobune are dramatically different from the ones of Ivo

Andrić, and Meša Selimović. Figure 6 displays these features.

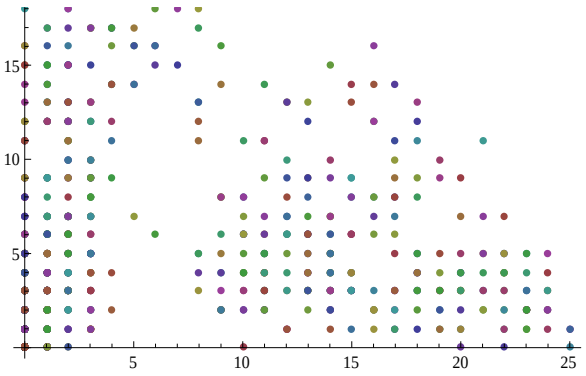

Figure 6. Pattern of the points representing frequencies in the first and second principal components of Pobune.

The frequency profile of first and second 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 multilayer perceptrons to do the job.

5. 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 (McCuloch and Pitts 1943) definition of unsupervised learning rule by Hebb in (Hebb 1949), and the first ever implementation of Rosenblatt's perceptron in (Rosenblatt 1958). 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 (Minsky and Papert 1969), 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.

5.1 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 built 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 fourteen input terminals, five hidden neurons in one hidden layer, and an output layer with one neuron is chosen.

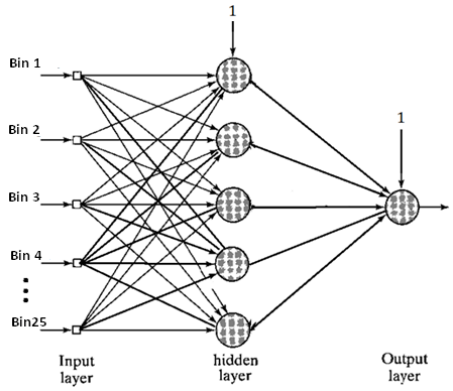

Fig. 3. Signal flow graph of the chosen neural network

5.2 Activation Functions

Activation or transfer function of a neuron is a rule that defines how it reacts to data received through its inputs that all have certain weights.

Among the most frequently used activation functions are linear or semi-linear function, a hard limiting threshold function or a smoothly limiting threshold such as a sigmoid or a hyperbolic tangent. Due to their inherent properties, whether they are linear, continuous or differentiable, different activation functions perform with different efficiency in task-specific solutions.

For classification tasks antisymmetric sigmoid tangent hyperbolic function is the most popularly used activation function:

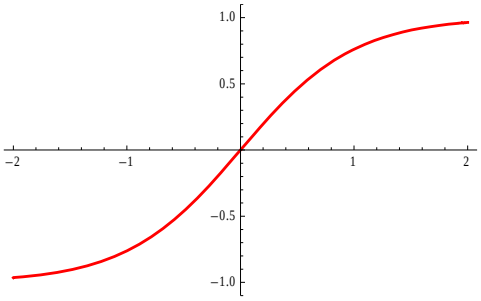

Fig. 1. Antisymmetric sigmoid tangent hyperbolic activation function

5.3 Learning Rules

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.

The term supervised indicates an external teacher who provides information about the desired answer for each input sample. Thus in case of supervised learning the training data is specified in forms of pairs of input values and expected outputs. By comparing the expected outcomes with the ones actually obtained from the network the error function is calculated and its minimization leads to modification of connection weights in such a way as to obtain the output values closest to expected for each training sample and to the whole training set.

In unsupervised learning no answer is specified as expected of the neural network and it is left somewhat to itself to discover such self-organization which yields the same values at an output neuron for new samples as there are for the nearest sample of the training set.

Reinforcement learning relies on constant interaction between the network and its environment. The network has no indication what is expected of it but it can induce it by discovering which actions bring the highest reward even if this reward is not immediate but delayed. Basing on these rewards it performs such re-organization that is most advantageous in the long run [16].

The modification of weights associated with network interconnections can be

performed either after each of the training samples or after finished iteration of the whole training set.

The important factor in this algorithm is the learning rate η whose value when too high can cause oscillations around the local minima of the error function and when too low results in slow convergence. This locality is considered the drawback of the backpropagation method but its universality is the advantage.

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

- the first step was selection of the training and testing examples - *texts to be studied,*
- next stage was taken by the choice of textual descriptors to be analyzed - *the writerprints of the authors of previously selected texts,*
- then followed the third phase of calculating characteristics for all descriptors, *calculation,*
- transform randomly chosen data matrices into matrices with principal components *principal component analysis*,
- count frequencies of principal components in bins of equal length that were later used for training of the neural network, *calculation of frequencies in bins,*
- specification of the network with its architecture and learning method can be seen as the fourth step of the whole procedure, *neural network,*
- the fifth consisted of the actual *training of the network*,
- the sixth stage is *testing*,
- and the final one corresponded to analysis of obtained results and coming up with some conclusions and possible indicators for improvement, *analysis of obtained results*.

This process is applied to different input data, with a artificial neural network of 25 input terminals, five hidden neurons in one hidden layer and an output neuron.

The input vector $\boldsymbol{x}$ is twenty five dimensional with components frequencies in corresponding bins as shown in the signal flow graph in Figure 3. 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 two authors in a mixed text. We have chosen 100 set of 200 paragraphs from each of the texts. Each 200 paragraph set is transformed into its principal components, and only first principal components are taken into account. Hence we have 100 first principal components from each text. Then principal components are transformed into data vectors whose elements are frequencies in 20 uniformly specified bins. The resulting data is a 100×20 matrix for each text.

In the training phase, neural network succeeded to classify Ivo: Cuprija na Drina, and Mesa: Derviš i Smrt paragraphs, with 100% probability of correct classification.

Then the test data consisting of a random mixture of 100 Cuprija and 100 Smrt data is sent to the neural network for classification. Network classified this data with 100% probability of correct classification. Next we sent to the network the data of length 200 from other texts. The correct classification numbers are as follows.

Table 3. Number of correct classifications of 200 test data at the end of the training period of the neural network with Cuprija/Smrt data.

	Cupr	Smrt	Znak	Prok	Tvrd
Ivo	200	1	200	195	175
Meša	0	199	0	5	25
Suc %	100	99.5	100	97.5	12.5

Secondly in the training phase the texts Ivo Andrić: Cuprija na Drina, and Derviš Sušić: Pobune are used. Neural network learned to distinguish the two texts with 100% probability of correct classification. Next test data is sent to the network. This data also classified with 100% probability of correct classification. When test data of length 200 sent to the network from other texts. The correct classification numbers are as follows.

Table 4. Number of correct classifications of 200 test data at the end of the training period of the neural network with Cuprija/Pobune data.

	Cupr	Pob	Znak	Prok
Ivo	200	0	152	192
Sušić	0	200	48	8
Suc%	100	100	76	96

The same is done with the texts Meša: Derviš i Smrt, and Sušić: Pobune. Neural network learned to distinguish the two texts with 96.5% probability of correct classification. Test data classified with 96% probability of correct classification. When test data of length 200 sent to the network from other texts. The correct classification numbers are as follows.

Table 5. Number of correct classifications of 200 test data at the end of the training period of the neural network with Cuprija/Pobune data.

	Smrt	Pobune	Tvrdjava
Meša	199	20	200
Sušić	1	180	0
Success	99.5%	90%	100%

As it is seen from tables above, the neural network is successful in the test data from the texts he trained for. The success in the classification of other books of the same authors are also satisfactory.

6. 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

Andrić, I. 1981. Na Drini Ćuprija, Svjetlost, Sarajevo.

Andrić, I. 1989. Znakovi Pored Puta,, Svjetlost, Sarajevo.

Andrić,I. 1980. Prokleta Avlija, Svjetlost, Sarajevo.

Binongo, J. 2003. Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. Chance 16(2): 9–17.

Bosch, R., and J. Smith. 1998. Separating hyperplanes and the authorship of the disputed federalist papers. American Mathematical Monthly 105(7): 601–8.

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

Chaski, C. 2001. Empirical evaluations of language-based author identification techniques. Journal of Forensic Linguistics. 8(1): 1–65.

Chaski, C. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. International Journal of Digital Evidence 4(1).

Forsyth, R. 1997. Short substrings as document discriminators: An empirical study. Paper presented at ACH/ALLC.

Fung, G. 2003. The disputed Federalist Papers: SVM feature selection using concave minimization. Proceedings of the 2003 Conference on Diversity in Computing. 42–6.

Hayes, J. F. 2008, Authorship Attribution: A Principal Component and Linear Discriminant Analysis of theConsistent Programmer Hypothesis, I. J. Comput. Appl. 15, No. 2, 79-99 (2008).

Hebb, D. O. (1949). The Organization of Behavior. New York: John Wiley & Sons. Introduction
and Chapter 4 reprinted in Anderson & Rosenfeld [1988], pp. 45-56.

Holmes, D. 1998. The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing 13(3): 111–7.

Holmes, D., and R. Forsyth. 1995. The Federalist revisited: New directions in authorship attribution. Literary and Linguistic Computing 10(2): 112–27.

Holmes, D., L. Gordon, and C. Wilson. 2001. A widow and her soldier: Stylometry and the American Civil War. Literary and Linguistic Computing 16(4): 403–20.

Juola, P. 2006. Authorship attribution. Foundations and Trends in Information Retrieval 1(3): 233–334.

Juola, P., J. Sofko, and P. Brennan. 2006. A prototype for authorship attribution studies. Literary and Linguistic Computing 21: 169–78.

Kjell, B. 1994. Authorship determination using letter pair frequency features with neural network classifiers. Literary and Linguistic Computing 9(2): 119–24.

Kolman, B., and D. R. Hill. 2004. Elementary Linear Algebra, Pearson, New Jersey.

Kukushkina, O., A. Polikarpov, and D. Khmelev. 2002. Using literal and grammatical statistics for authorship attribution. Problemy Peredachi Informatsii 37(2).

Mcculloch , W. S., and W.Pill's. (1943). "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.

Minsky, M. L., and S. A. Papert, (1988). Perceptrons, Expanded Edition. Cambridge, MA:
MIT Press. Original edition.

Peng, R., and N. Hengartner. 2002. Quantitative analysis of literary styles. The American Statistician 56(3): 175–85.

Rosenblatt, E, 1958."The Perceptron:A probabilistic model far information storage and organization in the brain," Psychological Review, vol. 65, pp. 386-408.

Selimović, M. M.1966 , Derviš i smrt, Svjetlost, Sarajevo.

Selimović, M. M. 1970. Tvrdjava, Svjetlost, Sarajevo.

Smith, J. 2008. A review of authorship attribution,

Sušića, D. 1966. Pobune, Veselin Masleša, Sarajevo.

Williams, C. 1975. Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. Biometrika 62(1): 207–12.