SOUTHEAST EUROPE JOURNAL OF SOFT COMPUTING
Available online at www.scjournal.com.ba

Authorship Categorization With Neural Network

Nesibe Merve Demir [a]

[a] Computer Science and Engineering, International University of Sarajevo, Hrasnička cesta 15, 71210 Sarajevo, ndemir@ius.edu.ba

***Abstract*— This paper explores the use of neural networks in author classification. Also exploring the effect of stylometry is another aim of the research. Choosing the algorithm and descriptors are important issues in the research.**

In this paper methods for the multi-topic machine learning of an authorship attribution classifier were investigated using texts from novels as the data set. Artificial neural network is proposed to classify the texts of authors using a set of lexical descriptors and feed-forward neural network using back propagation.

The result shows that Turkish authors Peyami Safa, Orhan Pamuk and Mustafa Necati Sepetcioglu's two novels are successfully classified.

Keywords— Neural networks, author classification, stylometry, principal component analysis, Kohonen's self organizing map.

INTRODUCTION

Authorship categorization is the task of determining the author's a piece of work. In particular, categorizing textual work given other text samples produced by the same author is interested in.

Individuals have distinctive ways of speaking and writing, and there exists a long history of linguistic and stylistic investigation into author identification. Specific author features such as frequency of certain words, choice of rhymes, and habits of hyphenation have been used as tests for author attribution. These authorial features are examples of stylistic evidence which is thought to be useful in establishing the authorship of a text document. It is conjectured that a given author's style is comprised of a number of distinctive features or attributes sufficient to uniquely identify the author. Stylometric features used in early authorship attribution studies were character or word based, such as vocabulary richness metrics, word length etc.

In this context, stylometry which is the study of linguistic style; include various measures of vocabulary richness and lexical repetition based on word frequency distributions, to capture the style of a particular author, play an important role [1][2].

Content-dependent features and syntactic features are two types of features. An example syntactic feature is punctuation which is thought to be the graphical correlate of intonation which is the phonetic correlate of syntactic structure [3]. As punctuation is not guided by any strict placement rules, punctuation will vary from author to author. Chaski [4] has shown that punctuation can be useful in discriminating authors. Therefore, a combination of syntactic features may be sufficient to uniquely identify an author.

According to Rudman, over 1,000 stylometric features have been proposed [5]. Just as there is a range of available stylometric features, there are many different analytical techniques using these features for textual analysis of literature at author identification.

A lot of research has been made regarding author identification and many different methods have been proposed. One of them is an adaptive statistical data compression technique PPM algorithm [2]. Other statistical approaches used for author identification are factor analysis, Bayesian statistics, Poisson distribution, multivariate analysis, descriptor function analysis of function words, and Cumulative Sum [6].

So choosing descriptors has an important role as a distributor. In previous researches, different kind of descriptors were used, like word class frequencies, syntactic analysis, word collocations, grammatical errors, number of words, sentences, clauses, and paragraph lengths [7].

Neural networks were used in some researches [8]. The other machine learning approaches used are case based reasoning, support vector machines, etc.

In this research, artificial neural network is used. Network is designed as neural network model and is trained by Kohonen's self organizing Map (SOM) method.

Firstly artificial neural networks and more specifically, Kohonen's self organizing map are introduced. The rest of the paper is organized as follows. Methods / Experiments include detailed description of the methods and techniques used in the project. Results and Discussions include the experimental results of the project which is clearly stated and discussed. General summary and future directions depict the conclusion.

ARTIFICIAL NEURAL NETWORKS

The problems that cannot be formulated and solved mathematically are solved by computers with intuitive method. Artificial intelligent (AI) is the area that develops and improves that specialty of computers. AI systems learn with proved data and then make decision for other cases. AI system is capable of doing three things: store knowledge, apply the knowledge stored to solve problems and acquire new knowledge through experience [9].

An artificial neural network (ANN) is an interconnected group of artificial neurons that uses a mathematical or computational model to process information. It is a software simulation of a "brain" [10]. Neural network (NN) is a machine that is designed to model the way in which the brain performs a particular task or function [9]. The key element of this paradigm is the novel structure of the information processing system.

In a neural-network model, simple nodes, also known as neurons or units are interconnected to form a network. The nodes operate on a principle similar to biological neurons. The incoming synaptic strength of a biological neuron is modeled with a weight of the node. Each node also has an activation function, also known as a transfer function, which dictates when the node will fire. Not only is the structure of neural networks inspired by the biological nervous system, but functions unique to the brain, such as learning, have also been simulated to a certain extent with neural networks.

A neuron forms the basis for designing NN. A neuron has 5 fundamental elements [11]:

Inputs, weights, adder, activation function, output.

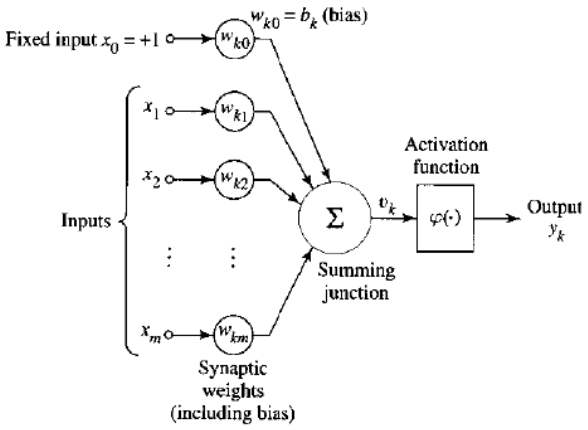

Figure 1.1 Nonlinear model of a neuron

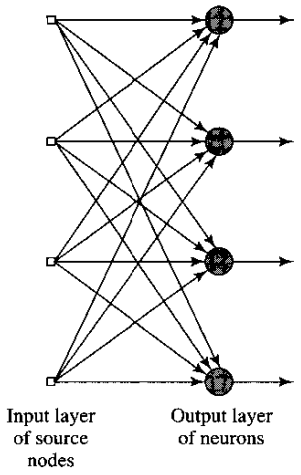

Figure.1.2 Feed-forward network with a single layer of neurons

According to architecture, ANN can be identified in three classes:

1. Single-Layer Feed-forward Networks,
2. Multilayer Feed-forward Networks
3. Recurrent Networks.

One of the common used architecture is Multilayer Feed-forward Network. It consists of input layer, one or more hidden layers and an output layer. These neural networks also are known as multilayer perceptrons (MLPs).

According to learning rules, NN can be divided into three learning rules: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value. A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier. In reinforcement learning, the neural network takes inputs and the machine interacts with its environment by producing actions. In unsupervised learning, the neural network receives inputs but obtains neither supervised target outputs, nor rewards from its environment [12].

MLPs are successful to solve difficult problems by training them supervised with popular error back propagation algorithm devised by Rumelhart et. al., 1986. This algorithm is based on the error-correction learning rule, which consists of two passes through different layers of network: forward pass and backward pass. It uses supervised learning in which the network is trained using the data for which inputs as well as the desired outputs are known [13]. Learning rate, the constant of proportionality, is an important factor of this algorithm. It is important to choose learning rate as large as possible without leading to oscillation. If learning rate is chosen small, iteration will converge too slowly.

MLP is trained by a gradient descent using the backpropagation algorithm to optimize the cost function. For example, the most common cost function is the mean square error criteria which summed the squared error between the desired and actual output vectors.

Choosing data is an important issue in NN. The data is divided into two parts. One is used for training and the second part is used for testing. The training data is a set of data which is used for training a neural network i.e. for adapting the weights of the network until the stopping criterion is met. Testing data is the data that has not been used by the neural network previously. This data is used to test the neural network performance. The performance test measures how well the neural network has learned to generalize. The purpose is training as less number as possible and testing with wider number of data [14].

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar.

KOHONEN'S SELF ORGANIZING MAP

As a clustering technique, Kohonen's self organizing Map (SOM) method finds similar data to another or dissimilar data to another data. It is based on competitive learning. Kohonen's self organizing Map (SOM) method is used to choose different data samples and then choose one sample for data set and calculate distance between all neuron vectors.

The training process is comprised of choosing to neuron whose weight closest to input vector and certifying the neuron as the winning neuron. The weights of all neurons that are around the winning one are adjusted proportionally to the distance.

The algorithm for SOM is as follows:

1. Choose initial values for neurons. This can be done by picking randomly different data samples.

2. Choose input for the data set. This can be done either randomly or by systematically going through the whole data set (cyclic order).

3. Calculate the distance of the selected data sample to all neuron vectors. Typically, the Euclidean distance measure is used. The neuron with the vector closest to the data sample is called the winner neuron.

4. Select winning element according to minimum distance. Update the vector of the winner neuron in a way that moves weight vector toward input vector.

$$\underline{c}_{win}^{(new)} = \underline{c}_{win}^{(old)} + \varphi(\underline{u} - \underline{c}_{win}^{(old)})$$

ϕ (learning rate) is the step size and should be chosen properly. If it is chosen too large, algorithm can not converge. So it is recommended to start with a smaller step size, like 0.5 and decrease the step size in each iteration.

5. If any neuron vector has been moved significantly then go to Step 2; otherwise stop.

EXPERIMENTATION

In this research the neural network was designed to classify the books of Peyami Safa, Orhan Pamuk and Mustafa Necati Sepetçioğlu.

Validation details

Wolfram Mathematica7 was used as an experimental tool. The dataset used for the experiment was split into three sets:

(1) Training set 18%.
(2) Testing set 30%.
(3) Evaluation set 52%.

Training set was used for parameter estimation. During training current condition of the learning process was examined using testing set. To avoid overfitting, those parameters were chosen for which the average training and testing recognition accuracy was the highest. The real recognition accuracy was checked after training using evaluation set that actually was another book of the authors'.

Dataset Description

For the experiments reported in this paper, size of the training set is 200 data and the test set is 337 data from each book. As evaluation set, 585 data was used from each authors' second books.

An important issue of using neural network is choosing proper data. In this research, the main problem is choosing descriptors that would recognize authors' style. Features selected in author identification methods must constitute the author's invariant properties of texts which is an invariant of its author, that it is similar in all texts of this author and different in texts of different authors [4].

Table 2.1 Features that was taken per paragraph

# of character	# of comma
# of "edat"(particles)	# of question mark
# of word	# of exclamation mark
# of sentence	# of dialog sentence
Sentence length	# of "ve (in English "and")"
Word length	# of triple dots (...)
# of "zarf" (adverb)	# of "baglac"(sentence connector)

From data, it is recognized that Peyami Safa uses short sentences if compared with Mustafa Necati Sepetçioğlu. Another important feature is triple dots that Mustafa Necati Sepetcioglu used it too much whereas Peyami Safa and Orhan Pamuk almost never used. It is important to choose distinctive features.

Data have to be normalized in self organizing map. Each set of attributes has to be scaled and their quadratic mean value will be equal to unity [16].

Design

Neural networks have been intensively used in the area of pattern recognition and have increasingly received considerable attention in various areas such as signal processing, pattern recognition and automatic control [14].

Firstly, the data is prepared by using Principal Component Analysis. Instead of using 14 features in the algorithm, 2 features are sent to algorithm after Principal component analysis done. It helped to save the time while there is no change in the success.

After the preparation of the data, it was necessary to choose centers. Basically training was done by choosing average of training sets as weight for deciding output neurons. Classifying was done by SOM algorithm that continues to iterate till error is less than the epsilon value that was chosen as 10^6, by using the centers.

The neural network is trained with the aim of classifying paragraphs of three authors.

After training, the weights are recorded. During testing and evaluation, that information is used. In testing process, mixed data set is sent that contains three authors' same books' paragraphs.

RESULTS AND DISCUSSION

Training Results

200 paragraphs were chosen from Peyami Safa(P.S.)'s novel "Bir Akşamdı", 200 paragraphs were chosen from Mustafa Necati Sepetçioğlu(M.N.S.)'s novel "Bir Büyülü Dünya ki" and 200 paragraphs were chosen from Orhan Pamuk(O.P.)'s novel "Kar" for training process.

In the table 3.1 below, the results of classification at the end of training by the machine are given.

Table 3.1 Correct classifications of the *training* data

	P.S.	***M.N.S.***	***O.P***	***Total***
# of Data	200	200	200	600
#of correct classified data	149	134	136	419
% correct classification	74.5	67	68	69.83

Testing Results

At the testing part, 337 paragraphs from each novel, 1011 paragraphs totally were used.

The results of the classification performed during testing are displayed in Table 3.2.

Table 3.2 Correct classifications of the *testing* data

	P.S.	***M.N.S.***	***O.P.***	***Total***
# of Data	337	337	337	1011
# of correct classified data	261	230	222	713
% correct classification	77.4	68.2	66	70.53

Evaluating Results

After training and testing, another set of data used for checking success of classification with same authors' different novels.

For that part, 585 paragraphs were collected from Peyami Safa's novel "Yalnızız", 585 paragraphs were collected from Mustafa Necati Sepetçioğlu's novel "Anahtar"and Orhan Pamuk's novel "Benim Adim Kirmizi". As a final output of classification process, the success was 69%. Table 3.3 shows the results in detail.

Table 3.3 Correct classifications of the *evaluating* data

	P. S.	***M.N. S.***	***O.P***	***Total***
# of Data	585	585	585	1755
# of correct classified data	363	421	430	1214
% correct classification	62	71.9	73.5	69.1

The machine identifies 62% Peyami Safa's data and 71.9% of Mustafa Necati Sepetçioğlu's data and 73.5% Orhan Pamuk's data from different novels.

CONCLUSION

In this paper, an approach is described for writer identification using three Turkish writers and two novels from each one. Our proposed method is based on the combination of optimal local and global feature subset.

In this paper as analytical technique neural network using Kohonen's self organizing Map (SOM) method is chosen. Fourteen descriptors are used as discriminators of texts. Principal component analysis was used to interpret the variation and to reduce the data.

Neural networks can successfully be used with choosing proper descriptors. With a wider set of textual descriptor, a higher success can be achieved. Language property is important in that sense. Having more information about writer's distinctive specialty and language may help for choosing appropriate features. Some future experiments can be done with defining more features and data from more novels. Also it is necessary to examine the results with more writers. Finally, it should be fine to try other methods of neural network and compare the results.

ACKNOWLEDGEMENT

The author would like to acknowledge the ideas and guidance during the research, of Prof. Dr. Mehmet Can. The author also would like to thank Amir Jamak for providing the code that extracts the necessary statistical information from the novels.

REFERENCES

[1] McEnery, T., & Oakes, M. (2000). "Authorship Identification and Computational Stylometry", In Dale, R., Moisl, H., & Somers, H. (Ed.), Handbook of Natural Language Processing, New York: Marcel Dekker. pp. 545 – 562.

[2] D. Pavelec, L. S. Oliveira, E. Justino, F. D. Nobre Neto, and L. V. Batista (2009), "Author Identification using Compression Models", 10th International Conference on Document Analysis and Recognetion.

[3] Egmont-Petersen M., de Ridder D., and Handels H. (2002), "Image Processing with Neural Networks: A Review", Pattern Recognition Journal, vol. 35, pp. 2279-2301

[4] S. Doan, S. Horiguchi, "An efficient feature selection using multi-criteria in text categorization for naive Bayes classifier", WSEAS Transactions on Information Science & Applications, vol. 2, no. 2, pp. 98–103, 2005

[5] J. Rudman (1997), "The state of authorship attribution studies: Some problems and solutions", Computers and the Humanities, 31(4):351–365

[6] P.Makvandi, Jassbi et al.(2005),"Application of Genetic Algorithm and Neural Network in Forecasting with Good Data", WSEAS Transaction on Systems, pg 337-342.

[7] R. S. Forsyth and D. I. Holmes (1996), "Feature finding for text classification", Literary and Linguistic Computing, 11(4):163–174.

[8]Pasqualoni A. (2006), "Author Attribution Using Neural Networks".

[9] Haykin S. (1999), "Neural networks and Learning Machines", Second Ed., Pearson, New Jersey.

[10] O. de Vel, A. Anderson, M. Corney and G. Mohay (), "Mining Email Content for Author Identification Forensics"

[11] Oztemel, E. (2003),"Yapay Sinir Ağları, Papatya Yayıncılık", Istanbul. pg48

[12] Ghahramani Z. (2004), "Unsupervised Learning", Gatsby Computational Neuroscience Unit, UK

[13] Kishan Mehrotra- Chilukuri K. Mohan and Sanjay Ranka (1996),"Elements of Artificial Neural Networks", MIT Press

[14]Tarassenko L. (2004), "A Guide to Neural Computing Applications", John Wiley & Sons Inc., New York, Toronto.

[11] McCombe N. (2002), "Methods of Author Identification", pg19

[15] Taner,T. (1997), Kohonen's Self Organizing Networks with Conscience.

[16] Safa, Peyami(2002) , "Bir Aksamdi", Otuken Nesriyat.

[17] Safa, Peyami(2000) , "Yalniziz", Otuken Nesriyat.

[18] Gazzah1 S. and Ben Amara N. (2008), “Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script”, The International Arab Journal of Information Technology, Vol. 5

[19] Sepetcioglu, Mustafa Necati , “Bir Buyulu Dunya ki”, Akran Yayincilik.

[20] Sepetcioglu, Mustafa Necati(1988) , “Anahtar”, Irfan Yayinevi.

[21] Merriam T.V.N. and Matthews R.A.J. (1994), “Neural computation in stylometry: An application to the works of Shakespeare and Marlowe”, Literary and Linguistic Computing, 9(1).

[22] Zhou N., Zhang W., Lee C., Xu L. (2008), “Lexical Tone Recognition with an Artificial Neural Network”, Ear Hear

[23] Pamuk, Orhan(2002), “Kar”, Iletisim Yayinlari.

[24] Pamuk, Orhan(2008), “Benim Adim Kirmizi”,Iletisim Yayinlari.

[25] Jamak, A., Savatic, A., Can M. (2012), “Principal Component Analysis for Authorship Attribution”, Business System Research, Vol.3 No.2 pp. 49-56