Teaching Neural Networks to Classify the Authors of Texts

Mehmet Can
Kanita Karađuzovic Hadžiabdić
Nesibe Merve Demir
mcan@ius.edu.ba kanita@ius.edu.ba ndemir@ius.edu.ba
International University of Sarajevo
Faculty of Engineering and Natural Sciences
Hrasnicka Cesta 15, 71000 Sarajevo
Bosnia and Herzegovina

Abstract: A lot of research has been done on author classification using various methodologies. One of them is using artificial neural networks. It is common that the number of descriptors used for author classification exceeds two. In this paper we propose a means of using artificial neural network to classify the authors of texts using only two descriptors: the number of words in a paragraph and a number of characters per word in a paragraph. The approach taken uses committee machines based on ensemble averaging. The basic idea is to solve the complex computational task by dividing it into a number of computationally simple tasks and then combining the solution of these tasks. The high performance achieved is because the committee is much better than the single best constituent in the isolation. Our results show that with the above approach we succeeded to correctly classify the works of Leo Tolstoy and George Orwell.

Keywords: Neural networks, committee machine, author classification, stylometry

Introduction

Author identification is the task of identifying the author of a given text; therefore, it can be formulated as a typical classification problem, which depends on discriminator features to represent the style of an author.

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]. Stylometry is mostly used for detecting plagiarism, finding authors of anonymously published texts, for disputed authorship of literature, etc.

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 [3].

Another important issue is choosing analytical technique. Different strategies were used until now 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 [4]. 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 [5].

Neural networks were also used in some researches [6]. The other machine learning approaches used are case based reasoning, support vector machines, etc.

In this research we use artificial neural networks for author classification. The neural network designed is composed of committee machines where each machine uses the feed-forward multilayer perceptron trained by the backpropagation algorithm.

We first introduce artificial neural networks and more specifically, multilayer perceptrons. We also mention basic idea behind committee machines. Then we introduce our design and experiment by presenting the results of categorization performance. Finally, we conclude with some general conclusion and future directions.

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: (1) store knowledge, (2) apply the knowledge stored to solve problems and (3) acquire new knowledge through experience [7].

Artificial neural network (ANN) is a software simulation of a “brain” [8]. Neural network (NN) is a machine that is designed to model the way in which the brain performs a particular task or function [7]. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well [14].

A neuron forms the basis for designing NN. A neuron has 5 fundamental elements [9].

1. Inputs: the examples that NN needs to learn.
2. Weights: show the effect of data in the neuron. Weights can be fixed or vary and may lie in a range that includes negative and positive values.
3. Adder: used for summing the input signals, which are weighted by the respective neuron.
4. Activation function: used for limiting the amplitude of a neuron's output. Activation function can also be expressed that it converts a neuron's weighted input to its output activation. Three basic types of functions are;
   - Threshold Function,
   - Piecewise-Linear Function,
   - Sigmoid Function.
5. Output: the value that is referred by the activation function.

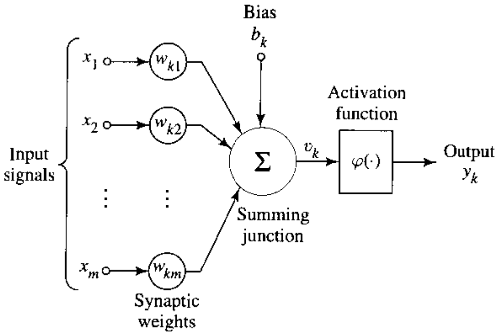

Figure 1.1 Nonlinear model of a neuron [7]

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 [10].

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 [11]. Learning rate, the constant of proportionality, is an important factor of this algorithm. Too high value of the learning rate causes oscillations around local minima of the error function and when too low results in slow convergence.

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 [12].

Design

In this research the neural network was designed to classify the books of Leo Tolstoy and George Orwell.

Choosing the proper features that will successfully perform the classification is not an easy task. A lot of work has been done in the past in author classification and identification using more than two descriptors. In this research we show that two descriptors:

1. the number of words per paragraph
2. the number of characters per word

were enough to successfully classify the works of Leo Tolstoy and George Orwell.

In order to classify the works of Leo Tolstoy and George Orwell using only two descriptors the principle of divide and conquer method was used, where a complex computational task is solved by dividing it into a number of computationally simple tasks and then combing the solution of these tasks. In neural networks using supervised learning, computational simplicity is achieved by distributing the learning task among a number of experts, that divide the input space into a set of subspaces, and fusing the knowledge acquired by experts to arrive at an overall decision that is supposedly superior to that attainable by any one of them acting alone. We used ensemble averaging, which linearly combines the outputs of different predictors producing an overall output [7].

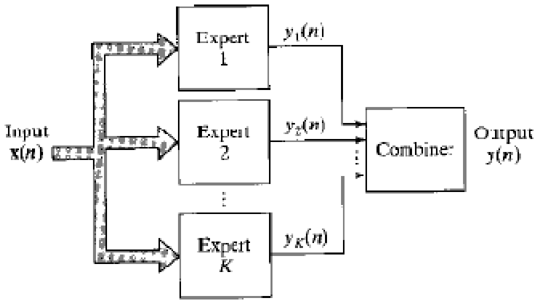

Figure 1.2 committee-machine based on ensemble averaging [7]

Only two of the above mentioned set of descriptors was used to show that it is possible to use neural networks to classify (possibly with an error of up to 5%) the text of two different authors, in our case the work of Leo Tolstoy and George Orwell.

The neural network designed is composed of ten committee machines. As shown in Figure 1.2 these committee machines share common input and combine individual outputs to obtain the overall output *y*. The architecture of each individual committee machine is as follows:

- the use of the feed-forward multilayer perceptron trained by the backpropagation algorithm,
- number of inputs is two (the mentioned descriptors),
- one hidden layer with three neurons,
- each committee machine produces an output of one vector, which after a

threshold which is experimentally chosen to be 0 classifies the two authors. A vector produced contains values of +1 and -1. +1 indicating the input represents a text written by Leo Tolstoy and values of -1 indicating a text written by George Orwell.

- size of the training set is 120 (sets of 60 data for each author), and 100 for the testing (sets of 50 data for each author). The data used in testing is different than the data used in the training,
- the data was normalized,
- activation function: antisymmetric sigmoid tangent hyperbolic activation function
- the optimum learning rate was found to be 0.00001

All of the parameters in the above stated topology of the neural network except for the initial values of weights in each committee machine remain the same, including the training and test data. Thus each committee machine starts with a different set of initial weights, everything else remains the same. Each committee machine adjusts its own set of weights. Once the network has been trained, its hidden layer activations are recorded as a representation of the number of words and number of characters per word in a paragraph. During testing this stored information is used to classify the text of Leo Tolstoy and George Orwell. On average 300 iterations were performed by each committee machine.

All of the experiments were done in *Wolfram Mathematica7.*

Results and discussion

Initial approach was to use the feed-forward multilayer perceptron trained by the backpropagation algorithm with the sigmoid tangent hyperbolic activation function, without the use of the committee machines. The number of inputs was set to the two descriptors chosen as explained previously.

The success of the training was first tested by varying the number of neurons in the hidden layer and it was found that more than 3 neurons in the hidden layer did not improve the efficiency of the neural net. When designing neural networks it is well known that the number of hidden units is determined during training. Too little hidden neurons leads to inaccuracy and too many to a failure to generalize. Initial weights were randomly chosen by the neural network, and threshold value was chosen to be 0.

The graph below shows the distribution of the descriptors used. The horizontal axis displays the first descriptor, i.e. the number of words per paragraph and the vertical axis displays the second descriptor, the number of characters per word in paragraph.

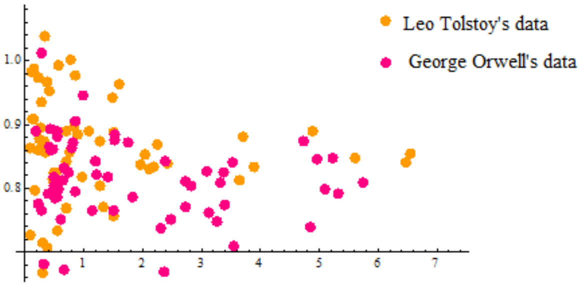

Figure 1.3 Distribution of a training set including both author's data

As depicted in the graph there is some presence of overlapping. Since no more than 72% correct classification of the two authors was found during training, and only at most 63% of correct classification in testing. It was decided to use a different approach in an attempt for a more successful classification, the committee machines.

The data in the second column displayed in the table below is taken as the first input (after normalization) and the data in the fourth column (also after normalization) is taken as the second input.

Paragraph	**Words**	Characters	**Char/Word**
1	**80**	424	**5.30**
2	**7**	38	**5.43**
3	**4**	16	**4.00**
4	**74**	383	**5.18**
5	**98**	451	**4.60**
6	**35**	171	**4.89**
7	**35**	162	**4.63**
8	**16**	77	**4.81**
9	**8**	35	**4.38**
10	**26**	118	**4.54**
11	**15**	59	**3.93**

The training data used was the data collected from Tolstoy's novel "Death of Ivan Ilych" and data collected from George Orwell's "Animal Farm". Sixty sets of descriptors were taken from each novel, and thus 120 data was used for training. Each committee machine was trained with this data. The test data consisted of 50 sets of *different* descriptors from each of the novels (the *same* novels were used,). The final output using the committee-machine based on ensemble averaging produced a success of 100% for the testing. However, no higher than 63% of successful classification was found by an individual committee machine.

The two graphs below Figure 1.4 and Figure 1.5 depict *separate* results belonging to the test data from Tolstoy's novel "Death of Ivan Ilych" (success of 80%) and Orwell's "Animal Farm" (success of 46%) respectively, together producing 63% of the classification success. The data points represent the inputs of the test data from each of the authors respectively and those that are encircled represent the correctly classified data as belonging to the text of Leo Tolstoy in the first graph and George Orwell's in the second graph. The results displayed were obtained by only *one* committee machine.

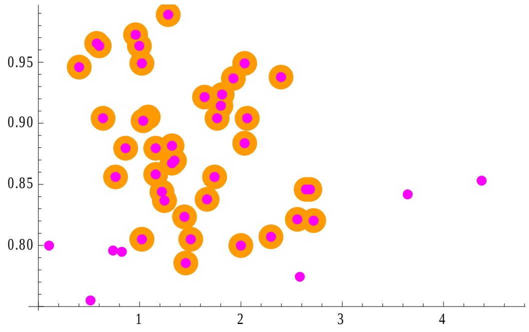

Figure 1.4 The test results belonging to the test data from Tolstoy's "Death of Ivan Ilych"

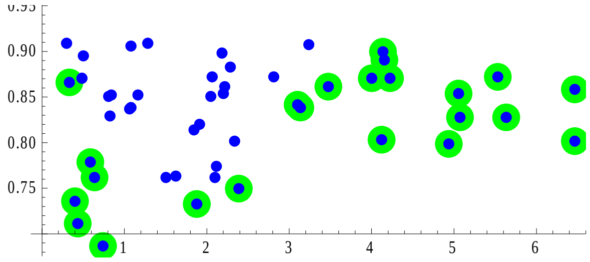

Figure 1.5 The test results belonging to the test data from Orwell's "Animal Farm".

Depending on the weights produced by a committee machine, the results displayed in the above figures vary.

Committee Results:

As before in the graphs below, the data points represent the inputs of the test data from each of the authors, those that are encircled represent the correctly classified data as belonging to the text of Leo Tolstoy in Figure 1.6 and George Orwell in Figure 1.7, (both test and training data are taken from same novel)

This time, the results are obtained by the committee machines based on ensemble averaging. It can clearly be seen that significant improvement of classification has been achieved. The correct classification of both authors is performed. This high performance has been achieved since the committee is much better than the single best constituent in the isolation.

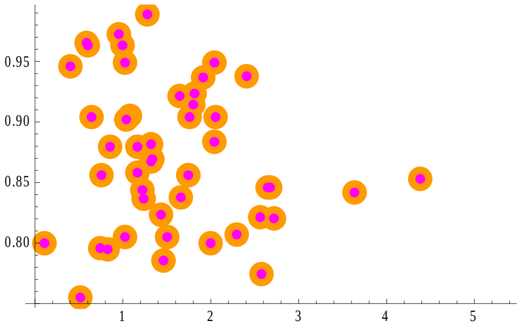

Figure 1.6 The committee-machine ensemble averaging for Leo Tolstoy's "Death of Ivan Ilych" test data

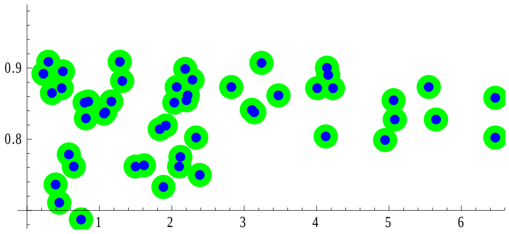

Figure 1.7 The committee-machine ensemble averaging for George Orwell's "Animal Farm" test data

We also tried testing our neural network with different novels from the *same* authors, such as Tolstoy's "War and Peace" and Orwell's "1984", thus completely different novels than those used during training. This time on average over 95 out of 100 paragraphs were correctly attributed.

Conclusion

In this paper as analytical technique feed-forward neural network using back propagation algorithm is chosen. This involves the training of neural networks to classify data even in the presence of noise and non-linear interactions within data sets [13]. Multilayer perceptrons using back propagation have been used to classify authors of works by Tolstoy and Orwell. Ten committee machines are trained by multilayer perceptron having different initial weights. Character/word ratios of each paragraph and number of words in each paragraph were used as descriptors.

The network is trained to attribute text from Leo Tolstoy's and George Orwell's novels. When test data is taken from the same novel as the training data, committee-machine based on ensemble averaging produced a success of 100%. However, when the training and test data are taken from different novels (authors remain the same), on average we achieved a success of over 95%.

Artificial Neural Networks is efficiently used for successful classification of Leo Tolstoy's and George Orwell's text. Despite providing only rough data as inputs to the neural network, successful classification has been achieved.

Future research

It remains to be investigated do the two mentioned descriptors capture the style of Leo Tolstoy and George Orwell or are they simply enough only to correctly classify their novels? Our experiments involved the use of static structure of the committee machines using ensemble averaging. What would happen if boosting is used, where weak learners are combined to give a strong learner? Would the neural network described also work if more than two authors are used?

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

[5]. 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.

[6]. Anthony Pasqualoni.(2006) Author attribution using Neural Networks.

[7]. Haykin S. (1999). Neural networks and Learning Machines. Second Ed., Pearson. New Jersey.
[8]. Niamh McCombe(2002). Methods of Author Identification. pg19

[9]. Oztemel,Ercan.(2003)Yapay Sinir Ağları. Papatya Yayıncılık, Istanbul. pg48
[10]. Zoubin Ghahramani(2004). Unsupervised Learning. Gatsby Computational Neuroscience Unit,UK

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

[12]. Lionel Tarassenko(2004). A guide to neural computing applications. John Wiley & Sons Inc., New York,Toronto.

[13]. Thomas V.N Merriam and Robert A.J. Matthews(1994). Neural computation in stylometry: An application to the works of Shakespeare and Marlowe. Literary and Linguistic Computing, 9(1).
[14]. Christos Stergiou, Dimitrios Siganos. Neural Networks