VOL.2 NO.2September 2013 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online:www.scjournal.com.ba

Artificial Neural Network Techniques in Authorship Attribution

Nesibe Merve Demir

International University of Sarajevo, Faculty of Engineering and Natural Sciences, HrasnickaCesta 15, Ilidža71210 Sarajevo, Bosnia and Herzegovina

Article Info

Article history:
Received 17 Sep.2013
Received in revised form 17 Oct 2013

Keywords:
Neural Network, Kohonen's Method, Nearest neighbor Method, Committee Machines, Authorship attribution

Abstract

This paper covers a text classification problem: the identification of the author of a text. It is necessary to find author of a text with given information from a set of candidates whose sample texts were provided. Attempting to solve authorship problems by choosing only features that exist in the anonymous texts did not yield good results. In contrast to other classification tasks, it is not clear which features of a text should be used to classify an author. Also methods that are chosen for classification are important .In this paper, methods for the machine learning of an authorship attribution classifier are investigated using 2 books from each 3 writers as the data set.

1. INTRODUCTION

Quite rapid development is seen on artificial neural networks technology which implements learning of computers since 1990s. This technology in a short time achieved to be discipline which takes attention of researchers and studies started to be a part of daily life by taking place out of laboratories. Artificial neural networks were making important contributions on researches related to human brain whose way of working is unknown and also they were causing to new developments because they are computer systems which are developed with the aim of carrying out capabilities automatically without taking any kind of help as producing new ideas with learning way which is among characteristics of human brain and producing and discovering new information. A big number of artificial neural network models are developed in a short time and uncountable application arose.

Problems whose solution is not possible and cannot set formulation mathematically would be solved by computers through intuitive methods. Studies which equip computers with those characteristics and implement development of those capabilities are known as artificial intelligence studies.

The most fundamental characteristics of intelligent systems are those that they have characteristic of giving decision as based on information while working or producing solutions for problems and giving decisions on following events by learning the inputs with available information.

Authorship categorization is a subset of the more general problem called "authorship analysis" [1]. Authorship analysis includes other distinct fields such as author characterization and similarity detection. Specific author features such as unusual diction, 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. There are many different techniques using these features for author identification. These include statistical approaches, neural networks, feed-forward neural networks, cascade correlation, genetic algorithms and Markov chains.

2. GENERAL CHARACTERISTICS OF ARTIFICIAL NEURAL NETWORK

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 [2].

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" [3]. Neural network (NN) is a machine that is designed to model the way in which the brain performs a particular task or function [2]. 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.

General characteristics of artificial neural networks change according to applied network model. The features of each model are explained as detailed with respect to related models. There is given general characteristics which are functional for all models. They can be summarized ad following:

Artificial neural networks carry out machine learning. Basic function of artificial neural networks is to make computers learnt. They try to give similar decisions on similar events by learning the events.

Their programs are not similar to the programming methods of which working styles are known. They have data processing method which is exactly different from data processing methods in which conventional programming and artificial intelligence are applied.

Storage of information: Information on artificial neural networks is measured with the values of networks' connection and is stored on the connections. Data is not located inside of a data base or a program as it is on the other programs. Information is stored on the network and it is difficult both to detect and interpret it.

Artificial neural networks learn by using the samples. To make artificial neural networks learnt event, it is necessary to determinate samples related to the event. They gain the ability of making generalization about related example by using samples (adaptive learning). It is not possible to educate artificial neural network if there is not sample or it cannot be found. Samples are examples which are experienced. For example, a doctor asks some questions to the patient and gives medicine by diagnosing according to the answers. The questions which are asked, the answers which are replied and the diagnose which is given can be qualified as a sample. Artificial neural network can give similar diagnose to the similar sickness if consultation of a doctor with his patients in a specified time and diagnoses of him is considered as a sample by noting them. It is quite important that samples which are obtained to show the event as completely. Successful results cannot be achieved if the event cannot be showed to the network in all aspects and related samples are not presented. It is not because the network is failing, but instead the examples are not taught well to the network. Because of that producing and gathering of samples has a special importance on the science of artificial neural network.

Firstly it is necessary to test education and performance of artificial neural networks to make them work securely. Education of artificial neural networks means to show existing samples to the network one by one and determination of relations between events on the sample by making worked mechanisms of the network itself. Present samples to educate each network are separated into two different sets. The first is used to educate network (education set), other is to test performance of network (test set). Each network is firstly educated with education set. It is accepted that education is completed when network starts to give true answers to all samples. Then, the samples as a test set which have never seen by network are taught to the network and answers of it need to be checked. Performance of the network is accepted as good if it gives answers which are acceptable as true to the samples that have never seen and it is taken to be used; if necessary it is used as online. If the performance of the network is not enough, solutions are applied as reeducation or education with new samples. This process continues until performance of the network reaches to the acceptable level.

They can produce information on new samples. Network can produce information on the samples that has never seen before by generalizing the samples which are taught to it.

They can work only with numeric information: Artificial neural networks work only with numeric information. It is necessary to turn information which is showed with symbolic expressions into numeric presentation. It makes difficult for expression of symbolic information with numeric values to interpret information and explain decisions (produced solutions).

3. STRATEGIES OF LEARNING

Different kinds of learning strategies are used on the systems that learn by samples like artificial neural networks. The system which carries out learning and learning algorithm in use changes depending on these strategies. Generally 3 kinds of learning strategies are implemented. These are as following:

a. *Supervised Learning*

A supervisor helps to learning system for learning state in this kind of strategy. The supervisor issues samples related to the state which is required to be learnt as input-output

set. In other words, inputs for each sample and outputs which need to be formed in return to inputs are sent to the system. The duty of system is to map inputs to outputs which are determined by the supervisor. The relation between inputs and outputs of sample is learnt by this means. The network of multilayer perceptron can be given as an example for networks which use this strategy.

b. *Unsupervised Learning*

A supervisor does not exist which helps learning of the system on this kind of strategy. Only input values are shown to the system. It is expected that the system learns relations between parameters on samples by itself. This is the strategy which is commonly used for classification problems. However, after the system learned, user needs to mark the meaning of output. ART networks can be shown as an example as the system which uses this strategy.

c. *Reinforcement Learning*

A supervisor helps to learning system also in this kind of strategy. However, the supervisor watches the system for production of output response to inputs which are shown to it instead of sending output set to the system which is necessary for each input set and products a signal which shows whether the produced output is right or wrong. The system maintains learning process by considering the signal which comes from the supervisor. LVQ network can be given as an example as the system which uses this strategy.

4. EXPERIMENT

In this research, methods for the multi-topic machine learning of an authorship attribution classifier were investigated using texts from novels as the data set. Four different methods of artificial neural network are proposed to classify the texts of authors using a set of lexical descriptors and compared their results. 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, the main problem is choosing descriptors that would recognize authors' style. In this research fourteen textual descriptors are used. The list of features is given in the table below.

Table 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)

Wolfram Mathematica7 was used as an experimental tool. In the research, nearest neighbor method, means as centers, Kohonen's self organizing Map (SOM) and special committee machines were used.

5.COMPARISON OF CLASSIFICATION TECHNIQUES and RESULTS

Firstly, the data is prepared by using Principal Component Analysis (PCA). Instead of using 14 features in the algorithm, 3 features are sent to algorithm after Principal component analysis done. It helped to save the time while there is no change in the success. Methods were used with PCA data and without PCA data.

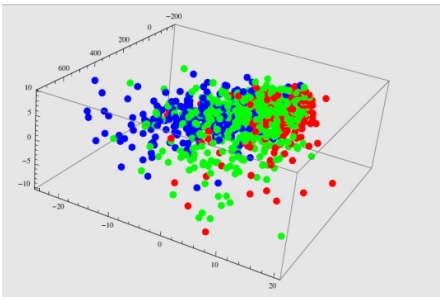

Figure 1. Training data after PCA

First experiments were done with 387 data from first books of three authors. NN method, means as centers and SOM methods were used with rough data.

For NN method, each data was sent to algorithm one by one and the nearest group was chosen for sent data. Books were sent separately in there. However, in c-means method center vector was calculated and this time for each data point, distance was counted between this center and data. The nearest one was chosen. For SOM, basically training was done by choosing average of training sets as weight for deciding output neurons. SOM algorithm continues to iterate till error is less than the epsilon value that was chosen as 10^16, by using the centers. After training, the weights are recorded. During testing and evaluation, that information is used.

Later, NN method and c-means were experimented with PCA data.

The neural network is trained with the aim of classifying paragraphs of three authors. In testing process, mixed data set is sent that contains three authors' same books' paragraphs.

The next step is experience of special committee machine. Firstly pair wise comparison was done. Each authors' first

book was sent to neural network one by one as a pair of other authors' books. 6 neural network mechanisms were worked. Training and testing were done for these networks.

For maestro committee machine, weights of six trained networks were taken. Random samples of 85 paragraphs were chosen. Firstly with training set, network is educated. Then with first validation data, it is checked and continues to train if necessary. As third step, second validation data set was sent for checking and if necessary more training was done. After three steps of training, test data was sent to the machine.

Training and Testing Results

387 paragraphs were chosen from Peyami Safa(P.S.)'s novel "Bir Akşamdı", 387 paragraphs were chosen from Mustafa Necati Sepetçioğlu (M.N.S.)'s novel "Bir Büyülü Dünyaki" and 387 paragraphs were chosen from Orhan Pamuk (O.P.)'s novel "Kar" for training process.

In the table 2 below, the results of classification at the end of training by Nearest Neighbor (NN), means as centers (C-M) and SOM methods for rough data are given.

Table 2. Mean Values of Rough data for Training

Methods	Mean Values (Rough Data)
NN	49.09%
C-M	44.53%
SOM	37.55%

In the table 3 below, the results of classification at the end of training by Nearest Neighbor (NN) and means as centers (C-M) for PCA data are given.

Table 3. Correct classifications of the *training* data with PCA

Methods	Mean Values (PCA Data)
NN	45.47%
C-M	42.90%

If we consult results one by one for each writer, we can say that features are best fits to PeyamiSafa's style. In the table below, results of each author for rough data can be seen.

Table 4. Correct classification results of each author for rough data

	P.Safa	M.N.Sepetcioglu	O.Pamuk
NN	57.36%	45.22%	44.70%
C-M	86.30%	22.74%	24.55%

In the table below, results for PCA data were given for each writer. We can say that PCA data did not help to improve results but just helped to gain time and performance.

Table 5. Correct classification results of each author for PCA data

	P.Safa	M.N.Sepetcioglu	O.Pamuk
NN	52.45%	41.86%	42.20%
C-M	84.50%	22.75%	21.50%

The results of pair wise comparison were given in the table below. Firstly, Peyami Safa's book was compared with M.N. Sepetcioglu's book and it was written in the row '1-2'. Then same data was compared with O.Pamuk's book and it was written in the row '1-3'. So in the table 1 represents P. Safa's book, 2 represents M.N. Sepetcioglu's book and 3 represents O. Pamuk's book.

Table 6.Pair wise comparison results of committee machine

	Validation	Test
1-2	77%	71%
1-3	70%	68.8%
2-1	73.5%	70.5%
2-3	64.7%	63.5%
3-1	73.5%	74%
3-2	70.2%	68.2%

For maestro committee machine, the results are in the table 7. For each step, 85 different paragraphs were sent to machine. So till coming to testing data, machine learned from 3 different data sets. Results show us that PeyamiSafa used a well- established language so results are similar from beginning to end.

Table 7. Maestro committee machine results

	Training	ValidationI	ValidationII	Testing
PeyamiSafa	80%	75.29%	82.35%	85.88%
M.N.Sepetcioglu	55.29%	44.71%	45.88%	40%
OrhanPamuk	43.53%	42.35%	24.71%	28.24%

Evaluating Results

After training and testing, another set of data used for checking success of classification with same authors' different novels. This part is experienced for nearest neighbor method and c-means method. Firstly 374 paragraphs were collected from there book which are Peyami Safa's "Yalnızız", Mustafa Necati Sepetçioğlu's "Anahtar" and Orhan Pamuk's "Benim Adim Kirmizi". This paragraphs as rough data were used in c-means method. Table 8 shows the results in detail.

Table 8. C-means results for second books

	P.Safa	M.N.Sepetcioglu	O.Pamuk	Mean
C-mea	70.87%	43.58%	42.42%	42.5%

Later, 187 PCA data were used from these books. As a final output of classification process, the success was not too high. Table 9 shows the results in detail.

Table 9.Correct classifications of the *evaluating* data for NN and c-means

	P.Safa	M.N.Sepetcioglu	O.Pamuk	Mean
NN	34.22%	37.97%	39.04%	37.07%
C-me	73.8%	48.66%	43.66%	43.13%

4. CONCLUSIONS

An approach is described for writer identification using three Turkish writers and two novels from each one. Proposed method is based on the combination of optimal local and global feature subset.

As analytical technique neural network using Kohonen's self organizing Map (SOM) method, Nearest Neighbor (NN) method, means as centers (C-Means) method and special committee machine are 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.

REFERENCES

O. de Vel, A. Anderson, M. Corney and G. Mohay (2001), Mining Email Content for Author Identification Forensics, ACM SIGMOD Record archive, Volume 30 Issue 4, December, Pages 55 - 64.

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

T.Tas, A. K. Gorur (2007) Author Identification for Turkish Texts, Journal of Arts and Sciences,No.7.

A. Savatic, A. Jamak, and M. Can, (2012) Detecting the Authors of Texts by Neural Network Committee Machines, Southeast Europe Journal of Soft Computing Volume 1. Number 1 March.

M. Can (2012) Authorship Attribution Using Principal Component Analysis and Nearest Neighbor Rule for Neural Networks, Southeast Europe Journal of Soft Computing 1-2 (2012) 36–47

K. Karađuzović-Hadžiabdić, N. M. Demir, (2012) Teaching Neural Networks to Classify the Authors of Texts, Southeast Europe Journal of Soft Computing Volume 1. Number 1 March.