Presentation a Neural Network with Gradual–Clustering Performance for Text Classification

Fahim Salimi
Department of Computer Engineering,
Islamic Azad University,
Bandar Abbas, Iran

Aazam Zarei
Department of medical,
Islamic Azad University,
Bandar Abbas, Iran

ABSTRACT
So far, various methods have been used to classify text. One of the methods of text classification is using Artificial Neural Network (ANN). In this article, we have proposed and examined text classification with the proposed method of clustering neural network. The method of ANN is that, this network is composed of several sub networks (in this method, we consider each sub-network as a cluster which contains nodes and edges) with specific examples and unique models and they are interconnected step-by-step in order that the network be completed. For finding the pattern of sub-networks, the relationship between the inputs and outputs are put into consideration and the resulting pattern is generalized in sun-networks. When sub-networks are compounded together, regarding rules that they have learned, they have found the ability to create a similar output from the same inputs. The proposed system includes two phases: Learning and Test. The system in learning phase considers a set of training texts for extracting sub-network properties as to obtain the main features of each sub-network, while it uses these specific features of sub-network for classifying the uncategorized text in test phase. We have utilized two sets of data for our experiments: 1) 20-newsgroup; 2) Reuters 21578. The experimental obtained results show that our proposed method can extend text classification, at its best to 92%.

Keywords
Artificial Neural Network, Unsupervised Learning, Text Classification, Machine Learning, Clustering Neural Network.

1. INTRODUCTION
Among the applications of classification it can be pointed out that the classification of webpages, indexing news items at different internet sources, and thematic categorization at the areas of commerce, medical and bioinformatics. Machine learning method with making the rules of text classification can be used as one of the good options for classifying. ANNs, however, in comparison with other classification method, have been used due to the long time for learning. At machine learning method, a number of manual categorized educational documentation is given to the system. Automatically, common deductive procedure makes rules from the sets of pore-labeled documents by means of learning features of categories. Text classification contributes natural language texts to one of or more pre-introduced categories based on the content. The processes of classifying are that at first Natural Language Processing is done, and then these results applying automatic learning method are interpreted. Functions such as text retrieval and text classification use this interpretation. The final aim is to classify texts at fixed number of pre-introduced categories.

In this article, we have attempted to model human's neural system and learning mechanism and regarding that clustering neural system has a similar mechanism with human learning mechanism, we use it in text classification procedure. Gradually, CNNs are constructed by means of basic component of sub-networks with observing examples. At first, for each new pilot model, we create a sub-network in the form of graph in which we consider inputs and outputs as nodes and the relationship between them as edges. This sub-network at learning procedure examines the relationship between inputs and outputs and considers a pattern for the same data. The produced sub-network can produce similar outputs for the same inputs. So, the considered sub-network for learning is able to identify the pattern of relevant input features. To obtain the neural network with multiple abilities, we compound the produced sub-networks together and we address to identifying patterns from inputs features.

We utilized two sets of data for examining the effects of text classification using clustering neural network: 1- 20-newsgroup; 2) Reuters 21578 that showed the accuracy of the proposed method on the collected data at its best 92% and demonstrates the above linear growth.

2. IMPLEMENTED WORKS
Both artificial neural networks and text classification have been widely examined. Neural networks have been stabilized in many tasks, e.g. in text classification [1] – recognizing the pattern [2] and optimization [3]. The most popular artificial neural network that has the most applications is multilayer perceptron network with propagation. [4] Applying neural networks for text classification is also quite common. For example, Lee et al. [5] proposed to change the objective function to improve the efficiency of feed-forward neural networks. Zhang and Zhou. [6] Proposed a model called Back Propagation for Multi-Label Learning (BP-MLL), which improved the training time for back propagation neural networks. Ghiassi et al. [7]

3. PROPOSED METHOD
The proposed method for text classification is based on machine learning method. In this method, there are two phases: training and test. In training phase, predetermined categories are used for machine learning and the meaning of each class for machine learning system is specified. In test phase, unknown Documents are given to the system. The system contributes that text to the class which is more similar automatically. The main rule of machine learning is to search an extensive space from possible assumptions and the kept previous knowledge by the learner and obtain the best possible class for the text of the test. The task of the learner in this space is for constituting assumptions that have the maximum compatibility with the existing training examples.

We want to investigate text classification procedure with clustering neural network method. Incremental neural network consists of three parts: constructing sub-network, compounding sub-network, and unsupervised learning.

3.1 Constructing Sub-network

A neural sub-network consists of inputs and outputs of nodes and edges with weight. The simplest sub-network assuming two features of A,B and response X is figure 1.

Next time, when neural network confronts with input properties of A,B will react and produces output X, since it already has learned this pattern based on the previous experience.

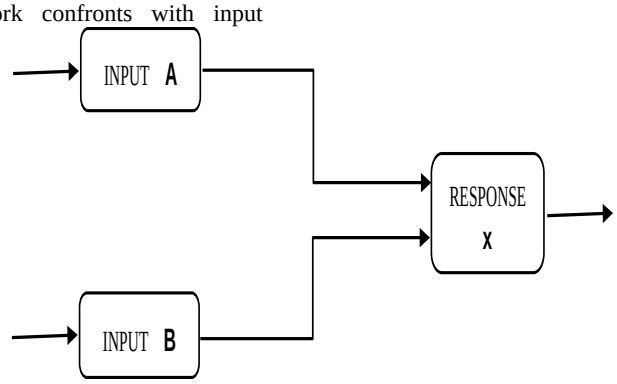

Figure 1. An example of a subnet with two input nodes and one output node

3.2 Compounding sub-network

For each pilot model, the similar sub-network can be constructed like the previous section. The compound of two sub-networks can be explained as follows. At first, input and output nodes of the neural network and then edges of the neural network are compounded, and it can be added in case that there is no edge. A compound of two simple sub-networks is explained in figure 2.

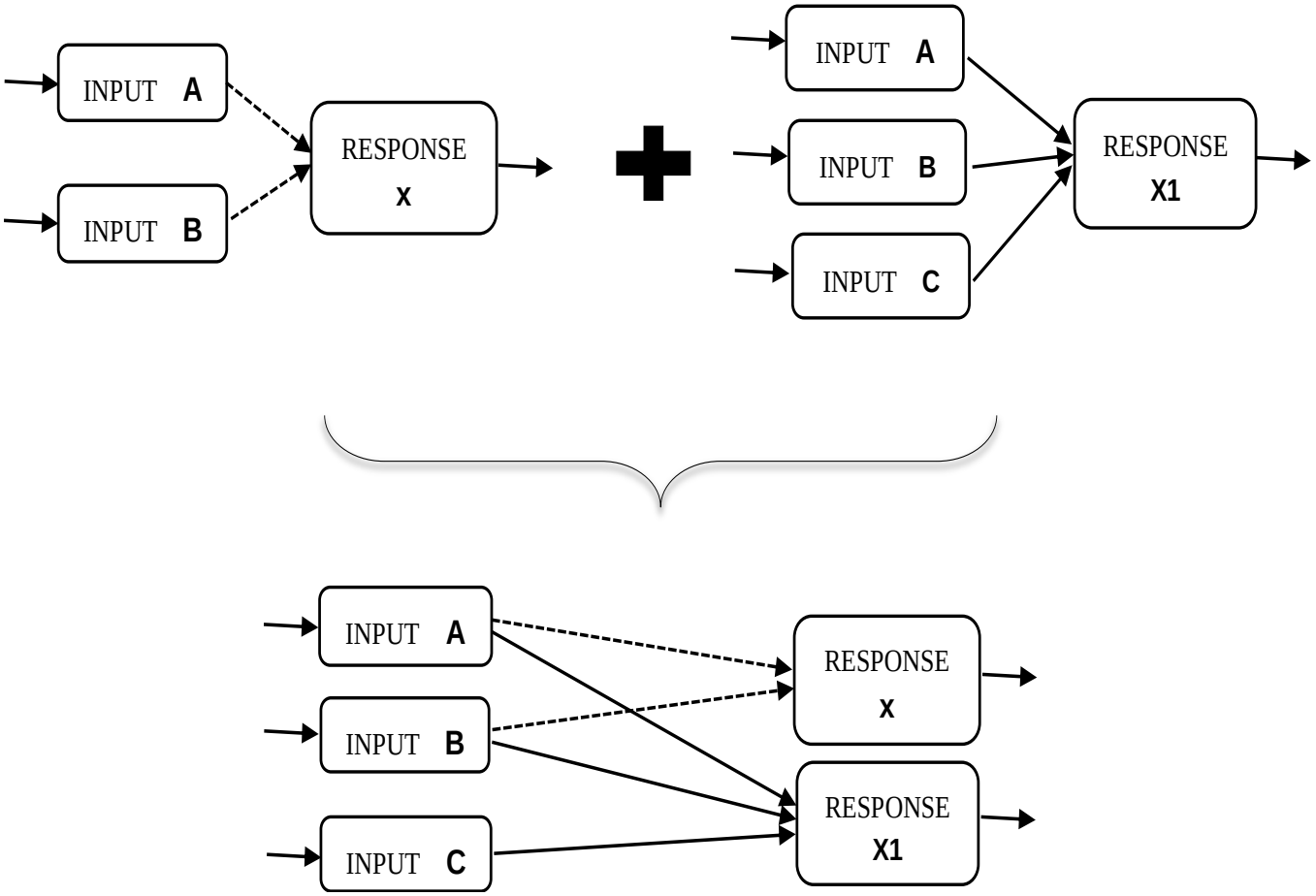

Figure 2. An example of the compound of 2 sub-networks

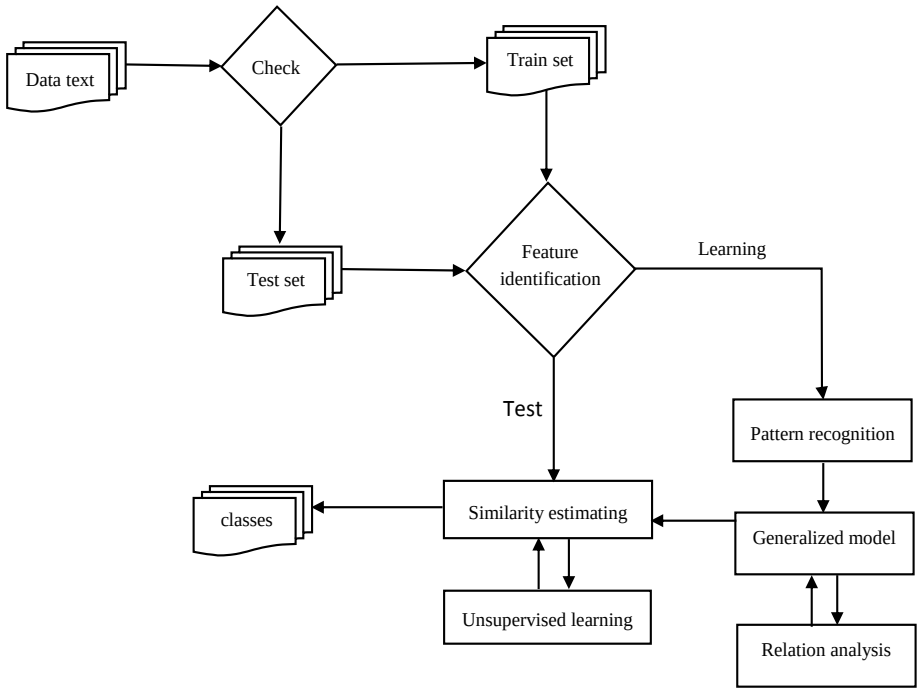

Figure 3- the architecture of clustering neural network system

For examples whose edges are available, the weight of the same edge is adjusted with an amount named learning coefficient.

3.3 Unsupervised Learning

For making a neural system, we give the certain pilot models to the system separately. This process will continue until the system is able enough to identify the input patterns. Although the system is similar to educational and experimental steps in a supervised learning procedure, it can be implemented without real distinguishing of these two steps.

The learning of this system is similar to human learning. Remembering new contents is a type of learning process. In human learning mechanism when we face with a new content, environment and areas can affect our memory thus, similar concepts modify the same content.

4. CNN ARCHITECTURE

The architecture of Clustering artificial networks system for text classification consists of 4 main sections: pattern recognition, generalizing feature for nodes, relation analysis, and unsupervised learning.

This architecture has been shown in figure 3

5. EXPERIMENTS

We used two standard benchmarks for evaluating text classification with clustering neural network: 1) 20-newsgroup; 2) Reuters 21578. For conducting the experiments, we used 70% of the data for learning randomly and the remaining 30% for test.

5.1 20-Newsgroup

We used 20-Newsgroup for evaluating text classification with clustering neural network. In table 1, 20-newsgroup has been shown.

Table 1. 20-newsgroup Data Set

Talk, politics, misc.	For sale, space, hardware com
Rec, car rec, sport	Volley ball, rec
Software, comp	Misc, politics
Religion, talk	Comp , graphics
Misc, comp, Mideast	Sys, mac, comp
Sport, football, bicycle	Asia, politics
Talk, misc, religion	Hockey, motorcycles rec
Comp, ibm, sys	Talk, polotics, med
Ms-windows, graphics	Electronics, windows
Politics, med, talk	Autos, Hardware, sci

5.2 Reuter-21578

Also, we used Reuter-21578 as a set of data. Reuter consists of 135 categories and 21578 items. That in this experiment we utilized 10 Reuter categories consist of 9976 documents that are shown in table 2.

Table 2. Reuter Top 10 Data Set

Class name	No. of document
Acq	2369
Corn	237
Crude	578
Wheat	283
Trade	486
Money	717
Ship	282

Interest	478
Grain	582
Earn	3964

6. RESULTS

For better conclusion we experimented 1000 documents from data set and compared their results with SVM and BAYES method that can be observed in table 3.

Table 3. Accuracy of Bayes, SVM and CNN

Exp	Vocab size	BAYES	SVM	CNN
1	26273	87.65%	89%	91.9%
2	49865	82.29%	88.98%	91.02%
3	46561	86.2%	88.2%	90.5%

Also, we utilized 10 Reuter categories consist of 9976 documents .The results for Reuter data set in three models of CNN, SVM, Bayes are shown in table 4.

Table 4. Accuracy of Bayes, SVM and CNN

Total vocabulary size	9187
Total validation records	2988
Accuracy of initial naïve BAYES	80.65%
Accuracy of initial SVM	82.27%
Improved accuracy using CNN	88.75 %

7. CONCLUSION

Artificial neural systems have various applications. In this article, we attempted to improve the efficiency of neural networks in text classification representing a classification method. The results of the experiments showed text classification methods using CNN that the efficient coefficient in its best reaches to 92% ant this represents a linear growth. That can be observed in chart 1.

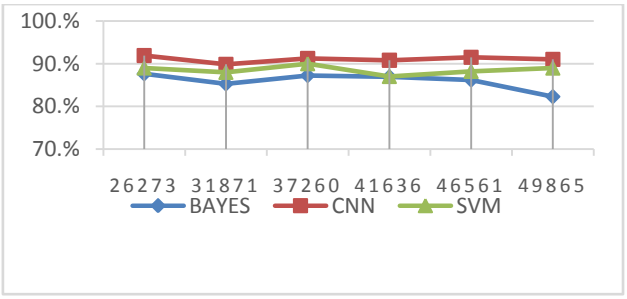

Chart 1. ComparisonChartaccuracyBAYES,SVMandCNN

8. REFERENCES

[1] C. Apte, F. Damerau, and S. M. Weiss "Automated Learning of Decision Rules for Text Categorization," ACM Transactions on Information Systems (TOIS), vol. 12, pp. 233–251, Jul. 1994.

[2] C. M. Bishop, Neural Networks for Pattern Recognition, 1st ed.Oxford University Press, 1996.

[3] A.Cochocki and R. Unbehauan, Neural Networks for Optimization and Signal Processing, 1st ed., Wiley, 1993.

[4] F. Rosenblatt, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Spartan Books, 1961.

[5] H. M. Lee, C. M. Chen, and C. W. Hwang, "A Neural Network Document Classifier with Linguistic Feature Selection," in Proc. of the International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert System, Vol. 13, pp. 555-560, 2000.

[6] M. L. Zhang and Z. H. Zhou, "Multi-Label Neural Networks with Applications to Functional Genomics and Text Categorization", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, pp. 1338- 1351, Oct. 2006.

[7] M. Ghiassi, M. Olschimke, B. Moon, and P. Arnaudo, "Automated Text Classification using a Dynamic Artificial Neural Network Model", Expert Systems with Applications, Vol. 39, pp. 10967-10976, 2012