A Comparison among Three Neural Networks for Text Classification

Zhan Wang, Yifan He and Minghu Jiang

Lab of Computational Linguistics, Dept. of Chinese Language, Tsinghua University, Beijing, 100084, China

Email: jiang.mh@tsinghua.edu.cn

Abstract

In this paper the effectiveness of three neural networks, the Competitive, the Backpropagation (BP) and the Radial Basis Function (RBF), in text classification is examined. The Competitive Network is a kind of unsupervised learning which is used in data clustering. The BP network is one of the most widely used models among artificial neural network patterns and the RBF network has also showed its vitality in recent years. All of the three are fit for pattern classification and function approximation. The three networks are independently used automatic text classification. Experimental results show that BP and RBF network outperform Competitive network because of the application of supervised learning. Besides its much shorter training time than BP, the RBF network makes precision and recall rates that are almost at the same level as BP's. Thus RBF network deserves more attention in the use of text classification.

1. Introduction

Text classification is a process to determine the category automatically based on the content of the texts. With the booming of the Internet, there is an urgent demand to classification systems that can help people organize vast information. The system usually consists of two parts: feature selector and text classifier. Feature selector reflects a document to a set of feature, in most cases a feature vector. The classifier classifies the feature vector to the category it belongs. Researches indicate that many techniques can be used in feature selection to improve accuracy as well as reduce the dimension of the feature vector and thus reduce the time of computation [4]. These techniques include using concept frequency instead of original word frequency and Latent Semantic Indexing (LSI) [6]. However, after using these techniques, the feature matrix is still quite complicated and requires a robust classifier to deal with.

Many classifiers have been applied to classify texts, including Vector Space Models, K-Means, Support Vector Machine (SVM), Naïve Bayes and so on [5]. Among many methods applied, several kinds of neural networks have shown attractive abilities. Several kinds of competitive networks are used in text classification [6], including Learning Vector Quantization (LVQ) and Self-Organizing Maps (SOM) network. These two are both variants of the basic unsupervised competitive network. Besides, Backpropagation (BP) network and Radial Basis Function (RBF) network are two successful examples as well [2]. BP network has been used in various areas and is now a proven classifier with satisfying performance. However, the main drawback of BP network is the long duration of training. RBF network is a three-layer network which is characterized by its high speed of training, and has begun to be introduced into the area of text classification.

In this paper we make a performance comparison between the basic competitive network, the BP network and RBF network in text classification. For unsupervised competitive network, we use the evaluation criterion described in [6]. We define positive and negative accuracies and use their average to evaluate the performance of clustering.

Table I. Definition of Positive and Negative Accuracy

	Clustered together	Not clustered together
In same category	A	C
In different category	B	D

Positive Accuracy $A_p=A/(A+C)$, and negative accuracy $A_n=D/(B+D)$. Average Accuracy $A_a=(A_p+A_n)/2$. For supervised classification networks, the factors we take into account are time, precision, recall and F1. Precision and recall are two widely used criteria of evaluation in text classification and text retrieval. Let us assume that P is Precision, R is Recall, A_i is the number of documents in category i, and B_i is the number of the documents classified by the classifier to category i, then

$P_i = \frac{A_i \cap B_i}{B_i}$, and $R_i = \frac{A_i \cap B_i}{A_i}$. F1 is a criterion that combines precision and recall. $F1 = \frac{2P \times R}{P + R}$.

2. Competitive network

A basic competitive network has an input layer and a competitive layer. The nodes in the competitive layer "compete" with each other, and the node that has the largest output becomes the "winning" neuron. The winning neuron is set to 1 and all other neurons are set to 0. The training of the basic competitive network uses the Kohonen learning rule. For each input pattern, the weight vector of the winning node is moved closer to the input vector using the following formula:

$$w_i(q) = w_i(q-1) + \alpha(p(q) - w_i(q-1))$$

where w_i is the weight of the winning neuron, p is the corresponding input vector and α is the Kohonen learning rate. However, a problem of this model is that if the initial weight of a neuron is far from any vector, it will never be trained. So a bias vector is added to the result of the competition. The winning node would cause the bias vector to decrease. Under this mechanism, it is more difficult for a neuron to continue to win. The degree of bias is represented by a factor called conscience rate.

3. Backpropagation network

Backpropagation network is one of the most widely used neural networks. It is a multi-layer network which includes at least one hidden layer. First the input is propagated forward through the network to get the response of the output layer. Then, the sensitivities are propagated backward to reduce the error. During this process, weights in all hidden layers are modified. As the propagation continues, the weights are continuously adjusted and the precision of the output is improved.

4. RBF network

Radial Basis Function (RBF) network has been widely used for pattern classification, function approximation and text classification.

4.1. Structure

The RBF network is a three-layer feed-forward neural network, between the input and the output layers there is a "hidden layer". When training, vectors are input to the first layer and fanned out to the hidden layer. In the latter, a cluster of RBF functions turn the input to output, adjusting the weight of the input to the hidden layer. Then, under the target vector's supervising, the weigh of the output vector of the hidden layer is adjusted. When clustering texts, the Euclidean Distance between the input vectors and the weight vectors, which have been adjusted by training process, is calculated. Each input sample is sorted to a class. Then the output layer collects samples belonging to same classes and organizes an output vector, the final clustering.

4.2. Training

In the hidden nodes, the activation function is usually chosen as Gaussian Function, the input of node *i* is the product of threshold b_i and the Euclidean Distance between weight vector W and input vector X:

$$k_i^q = \sqrt{\sum_j \left(x_j^q - w_{ji}\right)^2} \times b_i$$

Where, x^q_j is the *j*th component of the *q*th input vector, w_{ji} is the weight between the *j*th node in the input layer and the *i*th node in the hidden layer, b_i is a threshold to control the accuracy of the Gaussian Function. The output of the same node is as follows:

$$r_i^q = \exp\left(-\left(k_i^q\right)^2\right) = \exp\left(\sqrt{\sum_j \left(x_j^q - w_{ji}\right)^2} \times b_i\right)$$

Instead of adjusting b_i, we can use the parameter of *spread* in Neural Network Toolbox of Matlab 7.0 to control the performance of the network. The larger *spread* is, the smoother the function approximation will be. The input of the output layer is weight sum of the output of the nodes in hidden layer. The activation function is linear, so the output of the whole network, in response to the *q*th component of the input, is shown as:

$$y^q = \sum_{i=1}^{n} r_i \times v_i$$

where v_i is *i*th component of weight vector from the hidden layer to the output layer.

As is discussed above, RBF has a strong capability of approximation to the kernel vector in a limited part of the whole net. The training of the RBF network should be divided into two processes. The first is unsupervised learning, which adjusts the weight vector between the input and hidden layer. The other is supervised learning, which adjusts the weight vector between the hidden and output layer. Three parameters should be given before training: input vector, target vector and the threshold value, in Matlab, the *spread*.

5. EXPERIMENTS

5.1 Experimental Data

The testing corpus is based on 1204 documents from People's Daily from 1996 to 1998, all of which are first classified manually. These documents are separated into two parts: the training documents and the testing corpus. The training corpus contains 1084 documents in six different categories, and the rest 120 documents are used as testing samples, with 20 documents in each category. Instead of using word frequency, we use concept attributes as features of each document. We first select some key words from the document, and then reflect them to concepts in Hownet using the method described in [4]. Some key words without correspondent concepts in Hownet are preserved in the feature set to improve recall, such as names and addresses. Using the techniques described in [4], the dimension of the feature vector is reduced to 500.

5.2 Classification by the Competitive Network

In our experiment, we set the Kohonen learning rate of the competitive network to 0.01. We find that conscience rate has a large impact on the performance of the competitive network. The network is unable to cluster the documents when conscience rate is larger than 0.001 or too close to zero. We choose the conscience rate 0.000005. The result of clustering is listed in Table II.

Table II. Experimental Result of the Competitive Network

Positive Accuracy	Negative Accuracy	Average Accuracy
0.5438	0.9159	0.7299

We also experiment with other network originated from the basic competitive network, such as SOM. We find that the results are alike.

5.3 Classification by the BP Network

In our experiment, we choose tangent function and logarithm function as the activation function of the BP network and the training goal is set at 0.01.

The experiments are realized in Matlab 7.0, training of BP network costs a lot of time, on most personal computers training may take more than half an hour. After training, BP network can classify documents from the training set almost one hundred percent correct. As of documents from the test set, results are listed in Table III.

Table III. Experimental Result of the BP Network, F1=0.7929

	Precision	Recall
Economics	0.85	0.85
Politics	0.6471	0.55
Computer	0.9412	0.8
Sports	0.9	0.9
Education	0.76	0.95
Law	0.6667	0.7
Average	0.7941	0.7917

5.4 Classification by the RBF Network

The value of *spread* is 1.2. After 1084 iterations, the training completes. This process takes only 3 minutes or so. Then another 120 texts are used for testing. Comparing with the classification that has been made manually, we can get the precision, recall and F1 value of the classification made by RBF network. Our experimental results are listed in the following Table IV.

Table IV. Experimental Result of the RBF network, F1=0.7722

	Precision	Recall
Economics	0.615	0.8
Politics	0.435	0.5
Computer	1	0.8
Sports	1	0.9
Education	0.947	0.9
Law	0.722	0.65
Average	0.7866	0.7583

6. Discussions

From the experimental result above we can see that the classifier based on competitive network can classify document of different categories correctly, represented by a high negative accuracy rate. However, a low positive accuracy rate shows that many documents from different categories are not clustered together. This is partly because it is difficult to have information about the correct category in unsupervised learning. Classifiers based on BP network can classify documents in most categories correctly. However, in some special cases, such as the Politics type above, the performance drops to an unacceptable degree. In RBF experiment, precision and recall are low in some categories, such as Politics. In Computer and Sports categories, the performances are fairly nice. The two networks perform unsatisfactorily in the same category, the Politics, probably due to that we fail to recognize some proper nouns in certain categories and thus lost some important information about the text. Feature words in politics often have vague boundaries with other fields, such as economics, culture, military affaires, and so on. While sports and computer are highly characteristic fields, it is much easier choosing their feature words. Besides, the number of training documents are slightly too small compared to the dimension of the feature vector and the complex nature of text classification.

The RBF net takes less than one-tenth of the time BP takes when training. At the same time, it performs well in the categories in which the results are satisfactory. In the sports category, it even outperforms the BP network. Though in unsatisfactory categories such as economics and politics, it performs much worse than the BP network, we believe that if more effort is made in feature selection, the results will be even better. The fast training of RBF network is partly because it omits the feedback procedure which forms the BP's main characteristic. While a most important problem lies in how to decide the number of nodes in the hidden layer, which is also relative with the

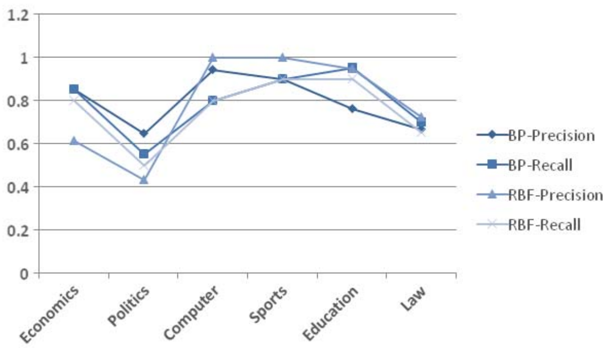

Fig. 1. Comparison between the BP and RBF

training speed of the net. In one of its modified models, RBF has a self-adaptability mechanism in resolving this problem. It can start training by 0 hidden node and in each iteration add hidden nodes according to the training error of the last iteration. Finally, when the error is reduced network to a given level, the hidden nodes number stops adding and remains.

7. Conclusions

Among the three networks we test, the basic competitive network sometime cannot cluster documents of the same category together, because it lacks category information as a method of unsupervised learning. As of the two supervised learning methods, the BP network has been proved effective in text classification for its mature backpropagation mechanism. The RBF network shows its quickness in training and it can even approach the BP's capability, especially being modified in some means. In fact, RBF network has many models. For instance, the network we use in this paper is a combination of RBF nodes and linear nodes. Furthermore, RBF nodes can be joined with competitive nodes to constitute a Probabilistic network which is also used in solving classifying problems. As a classifier, RBF can be a good substitute for the BP network, when the selected features are clear enough for BP network itself to produce satisfactory results.

Acknowledgement

This work was supported by National Natural Science Key Foundation of China (No. 60331010) and National Natural Science Foundation of China; Tsinghua University 985 Research Foundation; the Excellent Young Teacher Program, Ministry of Education, China; and State Key Lab of Pattern Recognition Open Foundation, Chinese Academy of Sciences.

References

[1]. Martin T. Hagan, Howard B. Demuth and Mark Beale. Neural Network Design. PWS Publishing Company, 1996.

[2]. Mark J. L. Orr. Introduction to Radial Basis Function Networks. Technical report, Center for Cognitive Science, University of Edinburgh, 1996.

[3]. Zhang Jianbao, et al. Realization of the RBF Network Classifier and Its Application. Computer Engineering and Research, Vol. 23, No.6, 2001. pp. 106-108.

[4]. Liao S. and Jiang M. "An Improved Method of Feature Selection Based on Concept Attributes in Text Classification". Lecture Notes in Computer Science, Vol. 3610, pp. 1140 – 1149, 2005.

[5]. Pang J., Bu D. and Bai S. "Research and Implementation of Text Categorization System Based on VSM". Application Research of Computers, 2003, No.9. pp. 23-26

[6]. Xu J., Wang Z. "An Eficient M ethod of Text Clustering Based on LSI and SNN". Journal of Tianjin University, Vo1. 37, No.1l, 2004. pp. 1026-1030.