2011 International Conference on Information Management and Engineering (ICIME 2011)
IPCSIT vol. 52 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V52.75

BP Neural Network Based On Genetic Algorithm Applied In Text Classification

SUN Ai-xiang [+] and LI Ming-hui

Management institute, Shandong University of Technology, Zibo, China

Abstract. The The BP neural network is one of the most commonly used methods in the field of text classification. BP learning algorithm gained success to some degree, but there are still some drawbacks: the error decreasing slowly, adjusting for a long time, more iterations lead to slow convergence, and training often fall into a local minimum and can not converge to a given error.In order to overcome the shortcomings of the BP neural network, this paper constructed an excellent BP neural network combined with genetic algorithm.In the learning process, the weights are described as chromosomes, then compute the fitness of the chromosomes, and then go on the genetic iteration until the convergence. And in this paper, this algorithm is applied to text classification.The experimental results show that: with the measurement of The BP neural network is one of the most commonly used methods in the field of text classification. BP learning algorithm gained success to some degree, but there are still some drawbacks: the error decreasing slowly, adjusting for a long time, more iterations lead to slow convergence, and training often fall into a local minimum and can not converge to a given error.In order to overcome the shortcomings of the BP neural network, this paper constructed an excellent BP neural network combined with genetic algorithm.In the learning process, the weights are described as chromosomes, then compute the fitness of the chromosomes, and then go on the genetic iteration until the convergence. And in this paper, this algorithm is applied to text classification.The experimental results show that: with the measurement of F1-measure the accuracy of the text classification has been greatly improved.

Keywords: text classification, bp, genetic algorithm, f1-measure.

1. Introduction

Since the 90's of 20th century, With the rapid development of network technology, Information has been expanding in high speed; And information will be growing fast and fast in the future. Now it is very difficult to estimate the amount of information. Among the many information carriers,the text is the most important one. According to statistics, 80% of the information is in the form of text. Relative to other information carriers, the text has been increasing in even more rapid speed. Text Mining has become the most important branch of data mining. It is a rapidly popularizing area of research. Text classification technology is one kind of the most important text mining technologies. Text classification can be applied in many fields;for example:Information filtering, information retrieval and digital libraries.However, the accuracy of text classification will have a direct impact on its applications in various areas. If the accuracy of text classification is low, it will not be helpful; on the contrary,it will bring about negative effect.Classification technology is the core of text classification technology. It plays an important role on the accuracy of text classification.The BP neural network is one of the most commonly used methods in the field of text classification, but the convergence speed of BP algorithm is slow and easily convergence to the local minimum point. The genetic

[+] Corresponding author. Tel.: + 15065339594.
E-mail address: aixiang12@163.com.

algorithm can be transplanted to the BP algorithm to overcome these shortcomings. Genetic algorithms are a developed optimization algorithm based on biological principle -"survival of the fittest". It takes a simple coding techniques to represent complex data structures and uses the genetic operations (selection, mutation, crossover) to improve the adaptability of the genetic population in order to gain satisfactory and optimal solution of the problem. In this paper, the BP neural network based on genetic algorithm is applied to the text classification [1]. The results showed that: F1-measure in the evaluation measure, the text of the classification accuracy has been greatly improved.

2. Text Classification

Text classification is selecting one or more class label to the test document from predefined categories.As the text is not structured data,we need to transform it to structured data which the computer can directly recognize and process, the structured form must can fully reflect the characteristics of the text itself, and can highlight the difference with other texts. Vector Space Model (abbr is VSM), is the most widely used text expression model[2] currently. It is proposed by the G. Salton in the last century, 60 years. In the vector space model, each text are expressed as a vector. And VSM is successfully applied to the SMART text retrieval system.Transforming texts into vectors, need to go through series of pre-processing step such as sub-word, stemming, removing stop words, lowering dimension.

3. BP Neural Network

BP (Back Propagation) neural network [3-4] is the most widely used neural networks currently. The full name of BP neural network is the artificial neural networks based on back propagation algorithm. It is commonly referred to as three-layer feed-forward network or sensor : The three-layers are: input layer, hidden layer and output layer. The features of BP neural network are: the neurons of one layer are fully connected with neurons of its adjacent layers ; the neurons in the same layer are of no connection; the neurons of one layer have not feedback connections with the neurons of other layers.the hierarchical structure of BP neural network is shown in Figure 1:

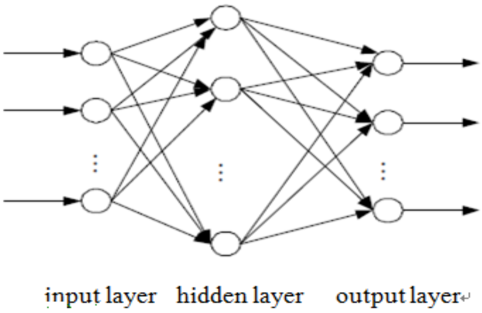

Fig. 1: BP neural network.

The learning process of BP neural network is composed of two processes: the forward dissemination of information and the back propagation of error. The input layer neurons is responsible for receiving input information from the outside world, and pass it to the neurons of the middle layer; The middle layer is the internal information processing layer, responsible for information transformation; According to the requirement of information capacity, the middle layer can be designed as a single hidden layer or more hidden layers;The last hidden layer transfer information to each neuron of the output layer, after further treatment, the neuron network complete one forward propagation study process, the output layer output the study results to the outside world. When the actual output and expected output does not consistent, enter the error back propagation stage. Error propagate from the output layer to the hidden layer, input layer, layer by layer,at the same time correct the weight by the way of gradient descent. Cycle of positive information dissemination and error back-propagation is called the process of continuously adjusting weights and also is called the training process of BP neural network, this process has been carried out to the network output error reduced to an acceptable level, or to the number of learning generation.

BP learning algorithm gains success to some degree, but there are still some drawbacks: the error decreasing slowly, adjusting for a long time, more iterations lead to slow convergence, and training often fall into a local minimum and can not converge to a given error.In order to overcome the shortcomings of the BP neural network, this paper constructed an excellent BP neural network combined with genetic algorithm.In the learning process, the weights are described as chromosomes, and then compute the fitness of the chromosomes, and then go on the genetic iteration until the convergence. In this paper, this algorithm is applied to text classification. The experimental results show that: with the measurement of F1-measure the accuracy of the text classification has been greatly improved

4. BP Neural Network Based on Genetic Algorithm

In the learning process, the weights described as chromosome, and select the appropriate fitness function, then go on the genetic iteration, until the convergence [5, 6].

4.1. The representation form of neural network and genetic factors

To facilitate the genetic operations, a string can be used to represent BP neural network topology, the string is composed of the weights, the weight of BP neural network is represented as,, where k is the number of the layer in BP neural network , W_{ij}^{k} is the connected weight between the i-th neuron in the k laye with the jth neuron in the k +1 layer; W_{ij}^{k} is in the [0,1] range, then the BP neural network can be expressed as:

$W_{11}^{1}\ W_{12}^{1} \dots W_{21}^{1}\ W_{22}^{1} \dots W_{11}^{2}\ W_{12}^{2} \dots W_{21}^{2}\ W_{22}^{2} \dots$

Through this encoding, the topology of the neural network represent a specific structure of genetic factor ,the network information is stored in them ,so we can go on genetic manipulation in order to gain the best chromosome.

4.2. The learning process of BP neural network based on GA

The learning process of BP neural network based on GA is shown in Figure 2 [7, 8].

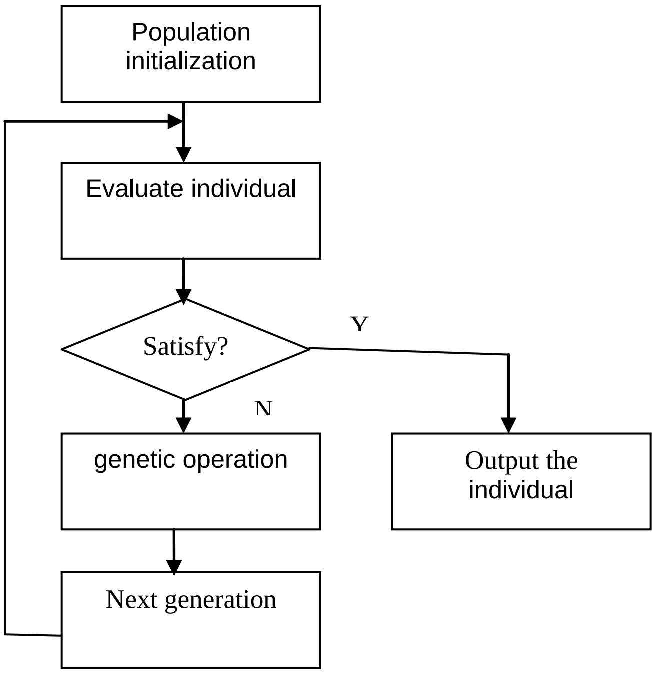

Fig.1: The process of BP neural network based on GA

4.2.1. The initialization of the population

As the conventional optimization algorithms, the initial point must be given before the iteration , the difference is that only one initial point is given before the iteration in the conventional optimization algorithms, but more than one initial point are given before the iteration in genetic algorithms,here,the initial points is the initial population.

4.2.2. The evaluation function

In the start stage that will reduce the input vector's selection to the connection vectors,then that will increase the winning chance of each connection vector.and then that will reduce the deviation between input vectors and the connection vectors as quickly as possible.This method can highten the speed of convergence; but the convergence speed is still low. Any individual in the population must be evaluated by the evaluation function

Assuming the evaluation function is formula (1):

$$f = 1/\sum(y_i - y_t) \tag{1}$$

Where, y_i is the desired output, it comes from samples (x_i,y_i), y_t is the actual output of neural networks.when the input is x_i

4.2.3. Genetic operations

Because of high dimensional property of text data, feature selection and feature extraction must be carried out to reduce the dimension of text data before clustering, After feature extraction, a word may be mapped to more than one dimension of input space*;* this method to determine the initial connection weights becomes very difficult.

Select Operation

Selecting which individual is by the value of the fitness,The larger the individual fitness is, the greater the chance that can be selected to participating in reproduction. And accordingly, the smaller the fitness of individuals is , the greater the chance of being eliminated is;The number of individuals is reflected by the elimination probability p_s.Assuming the number of individual groups is pop, it means that ps.pop individuals of poor fitness will be eliminated and can not enter the next iteration, in order to maintain a fixed number of individuals in the next iteration, the individuals that are eliminated by select operation will be replaced by the individuals having large fitness which are accquied by mutation operation

The mutation operation

Taking into account the chromosome (neural network representation) characteristics, there is not a cross operator, only a mutation operator.A sudden mutation rate p_m is an important parameter in genetic algorithm.The mutation operation is randomly selecting pm.pop individuals from preserved individuals for each individual selected, select a random number W_{ij}^k to change, the change is adding an increment $\vartriangle W_{ij}^k$ as formula (2)(3):

$$\vartriangle W_{ij}^k = N(0, b_m) \tag{2}$$

$$b_m = \lambda\ (1 - f/f_{max}) \tag{3}$$

Where, λ is the coefficient, $N(0, b_m)$ is the Gaussian function which mean is 0, variance is b_m. The greater the ith neuron fitness is, the greater the b_m is;so the probability that $\vartriangle W_{ij}^k$ off the mean 0 will be greater.

3 GA termination conditions

As,GA termination conditions, we can take different forms. In this paper, When the individual reaches the user's accuracy requirements, we can terminate all operations. Usually chosen the g formula (4) as convergence criterion:

$$| f (k+1)-f (k) | \le \varepsilon \tag{4}$$

Where f (k) is the best individual fitness of k generation.

5. Experimental Data And Analysis

5.1. Test corpus

This experiment used the Chinese text classification corpus [9] that Tan Song-bo, WANG Yue-fen filed. 200 texts (40 texts of finance and economics, 40 texts of computer, 40 texts of sports, 40 texts of health, 40 texts of real estate)are extracted from the corpus for training. At the same time, 200 texts (40 texts of finance and economics, 40 texts of computer, 40 texts of sports, 40 texts of health, 40 texts of real estate)are extracted from the corpus for testing. After preprocessing, each of this 400 text is converted into a 100-dimensional vector.

5.2. The criteria to evaluate the effectiveness of textclassing

F-measure is used in evaluating the text retrieval system [10]. F-measure combines two kinds evaluation criteria of text retrieval: Precesion (abbreviated as P; also known as purity) and recall (Recall, abbreviated as R).

Their meaning in evaluating text classification accuracy are as follows:

A category i, precision, recall the definition of the following equation:

The precision, recall of a category i, are defined as formula (5):

$$P(i) = N_1/N_2 \quad R(i) = N_1/N_3 \tag{5}$$

Which - N_1 the number of correct text divided into categories i by classifier

N_2 - the number of all texts in the category i

N_3 - the number of all texts divided into categories i by classifier

the F-measure of Category i are defined as formula (6):

$$F_\beta(i) = \frac{(\beta^2+1)\times P(i)\times R(i)}{\beta^2\times P(i)+R(i)} \tag{6}$$

Usually take $\beta = 1$, the recall and precision take the same weight, as formula (7):

$$F_1(i) = \frac{2\times P(i)\times R(i)}{P(i)+R(i)} \tag{7}$$

For the classification results, the overall F1-measure is the the weighted average value of each category ,s F1 , as shown in formula (8):

$$F_1 - \text{measure} = \frac{\sum_i (|i| F_1(i))}{|i|} \tag{8}$$

Where | i | is the number of all the text in i category.

5.3. The implementation specific parameters of GABP algorithm:

Group size: pop = 100

Bp neural network: the number of nodes in input layer is 100, the number of nodes in output layer is 5, the number of hidden layer nodes is 20

W_{ij}^k: the initial value obtained randomly from[0,1]

Out probability ps = 0.1

Mutation probability pm = 0.2

ε = 0.001;

5.4. The experimental results

Results of the comparison as shown in Figure 3:

The figure shows that: for the same test corpus, the classification results of BP is general, the classification results of GABP has been greatly improved

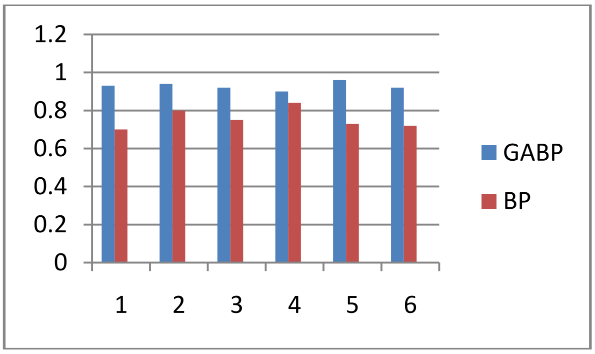

Fig. 2: Compare of converge generation.

6. Conclusion

This paper describes a new algorithm-GABP applied in text classification. In the learning process, the weights are described as chromosomes, and then compute the fitness of the chromosomes, and then go on the genetic iteration until the convergence. GABP overcomes the shortcoming of the original BP : the error decreasing slowly, adjusting for a long time, more iterations lead to slow convergence, and training often fall into a local minimum and can not converge to a given error., so it is an excellent text classification algorithm.The results show that: at the evaluation measurement--F1-measure, the GABP has greatly increased the accuracy of text classing.

7. References

[1] Wang An-lin. complex system,s analysis and modeling. Shanghai: Shanghai Jiaotong University Press .2004

[2] Salton G．Automatic Text Processing[M]．Addison-wesley Publishing Company,1988

[3] Zhang Liming. artificial neural network model and its application [M]. Shanghai: Fudan University Press, 1992.

[4] Wu Jiantong, Wang Jian hua. neural network technology and its application [M]. Harbin: Harbin Institute of Technology Press, 1998.

[5] Wang Chongjun. A genetic algorithm based on BP neural network algorithm and its application [J]. Nanjing University, 2003, 39 (5): 459 - 466.

[6] YEN G G. LU Haiming. Hierachical Genetic Algorithm Based on Neural Network Design[C] / / IEEE Symposiumon Combinations of Evolutionary Computation and Neura lNetwork. 2000.

[7] Liu Xu, Xue Fuzhen,Tanglei. Adaptive genetic algorithm based on multi-variable system design method of approximate model [J]. Chemicals and Instruments, 2009, 36 (1): 27 - 30.

[8] YANG guo jun , Cui Ping yuan,LI Lin-lin. Genetic Algorithm in Neural Network Control and Implementation [J]. System Simulation, 2001, 13 (5): 567 - 570.

[9] Tansong Bo, WANG Yue-fen. Chinese text classification corpus -TanCorpV1.0. http://www.searchforum.org.cn/tansongbo/corpus1.php

[10] David H, Heikki M, Padhraic S. Principles of data mining [M]. Zhang Yinkui, Liao Li, Song Jun and so on. Machinery Industry Press. Beijing, 2003