International Journal of Database Theory and Application
Vol.9, No.6 (2016), pp.275-284
http://dx.doi.org/10.14257/ijdta.2016.9.6.28

Application of Teaching Data Mining Based on Cloud Computing in the Prediction of Learning Achievement

Jianfen Liu and Junhui Zheng*

*Collage of Computer Science and Technology, Pingdingshan University,
Pingdingshan, Henan, China
2094991521@qq.com*

Abstract

Under the highly development of the information society, only by carrying out education reform can be cultivate more innovative talents. Because the traditional education idea has already taken root in the hearts of people, education informatization is a necessary way to change this kind of thought. The generation of cloud computing technology leads to the revolution of data processing technology. It can make use of a small amount of resources to effectively deal with the big data in the information system of educational institutions. Neural network is one of the important technologies of educational data mining in cloud computing environment. BP neural network is a typical multi-layer forward network, which is composed of input layer, hidden layer and output layer. It can be used to predict the data through the training model. In this paper, based on the characteristics of the distribution of education resources, we put forward the method to analyze big data of education by using Hadoop technology. This method uses the MapReduce programming model to manage the data, so as to improve the speed and efficiency of data analysis. Secondly, in Hadoop platform, this paper puts forward the method of parallel BP neural network in education data processing. The method consists of the following main steps: firstly, input data and set up a three layer parallel neural network. Secondly, according to the location of each node to block the data, and transfer M separate blocks to the Map function for processing. Thirdly, through the gradient descent method, the Map function finds the weight distribution of each block by iterative algorithm. Fourthly, we transfer the key-vlaue to the Reduce function, and update the statistics. Finally, repeat the update the calculation process of weight. After several iterations, the optimal solution of the objective function is found, and the weight distribution of the network is obtained. Finally, we simulate the parallel BP neural network algorithm based on education cloud platform, in order to prove that it is suitable for the prediction of learning achievement of the network teaching system.

***Keywords**: Machine learning, cloud computing, neural network, education*

1. Introduction

Under the highly development of the information society, only by carrying out education reform can be cultivate more innovative talents. Because the traditional education idea has already taken root in the hearts of people, education informatization is a necessary way to change this kind of thought. The generation of cloud computing technology leads to the revolution of data processing technology. It can make use of a small amount of resources to effectively deal with the big data in the information system of educational institutions. The concept of cloud computing has been widely concerned by people after a few years of development. Google, IBM, Microsoft, Amazon and other technology companies have developed cloud computing and achieved their own cloud computing platform. In addition to the cloud computing technology in the business sector, the open source community also has a deep study of cloud computing technology.

Hadoop is a typical representative of the open source cloud computing technology [1]. Google published the paper on the GFS and MapReduce in 2003 and 2004 [2-3], and the distributed file system and parallel programming model were introduced. Because it is very fit for the demand of mass information processing, it has caused great repercussions in the academic and industrial circles. At the same time, the research and development of cloud computing platform involves a variety of core technologies. Among them, the big data statistical analysis and the probability forecast technology is the key point and the hot spot of the research. The value of information contained in the big data is constantly being excavated [4]; Cloud computing information security technology is also in constant development [5,6]; Distributed computing [7] and distributed storage [8-10] are closely related with the resource scheduling technology.

At present, neural network is one of the important technologies of educational data mining in cloud computing environment. After decades of development, researchers have done a lot of research work in many aspects. Such as the adjustment of learning rate, the improvement of error function and the excitation function, the optimization of the network structure and the optimization of the algorithm. In the aspect of improving the learning rate, the paper proposes the step size adaptive method that uses the as a part of the network parameters, so as to have their own learning step [11]. Based on the same principle, the three stage learning step is realized in literature [12]. In literature [13], the learning rate is adjusted dynamically by the correction of the weights. In the aspect of optimization algorithm, in order to speed up the convergence rate of the standard LMBP algorithm, a variable step size LMBP algorithm is proposed in literature [14]. Based on the momentum term, the literature [15] adds a term that is proportional to the error, thus forming the three terms BP algorithm. Based on the standard Sigmoid function, the paper presents a model of activation function with four adjustable parameters [16]. However, these methods do not have the ability to deal with the large data of education, and cannot adapt to the modern education system.

In this paper, based on the characteristics of the distribution of education resources, we put forward the method to analyze big data of education by using Hadoop technology. This method uses the MapReduce programming model to manage the data, so as to improve the speed and efficiency of data analysis. Secondly, in Hadoop platform, this paper puts forward the method of parallel BP neural network in education data processing. The method consists of the following main steps: firstly, input data and set up a three layer parallel neural network. Secondly, according to the location of each node to block the data, and transfer M separate blocks to the Map function for processing. Thirdly, through the gradient descent method, the Map function finds the weight distribution of each block by iterative algorithm. Fourthly, we transfer the key-value to the Reduce function, and update the statistics. Finally, repeat the update the calculation process of weight. After several iterations, the optimal solution of the objective function is found, and the weight distribution of the network is obtained. Finally, we simulate the parallel BP neural network algorithm based on education cloud platform, in order to prove that it is suitable for the prediction of learning achievement of the network teaching system.

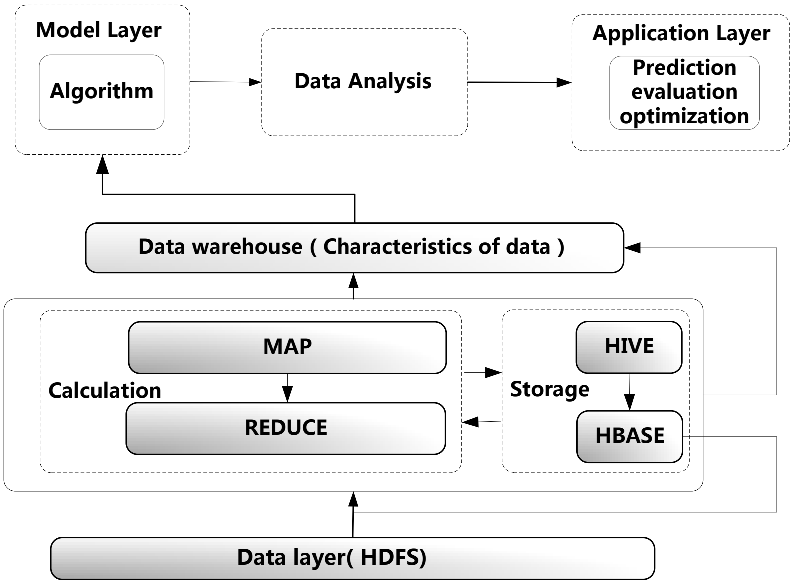

Figure 1. The Architecture Diagram of Big Data in Cloud Platform

2. Cloud Computing Platform Based on Hadoop

The application of cloud computing in the field of education can provide effective services for teaching, experiment, self-learning and tutoring. With the application of cloud computing in the field of education is gradually mature, the field of education is undergoing a transformation from the primary, simple computer assisted instruction to the complex, interactive cloud computing assisted instruction. Cloud teaching platform not only can improve the teaching efficiency, promote the interactive learning of students, but also can improve the students' thinking ability and promote the development of students' intelligence, so as to improve the overall teaching quality.

With the characteristics of college teaching, we use the distributed and hierarchical design structure in the teaching big data. In this paper, the establishment of the cloud computing platform as shown in Figure 2. From Figure 2, it can be divided into 3 layers which are data layer, model layer, and application layer.

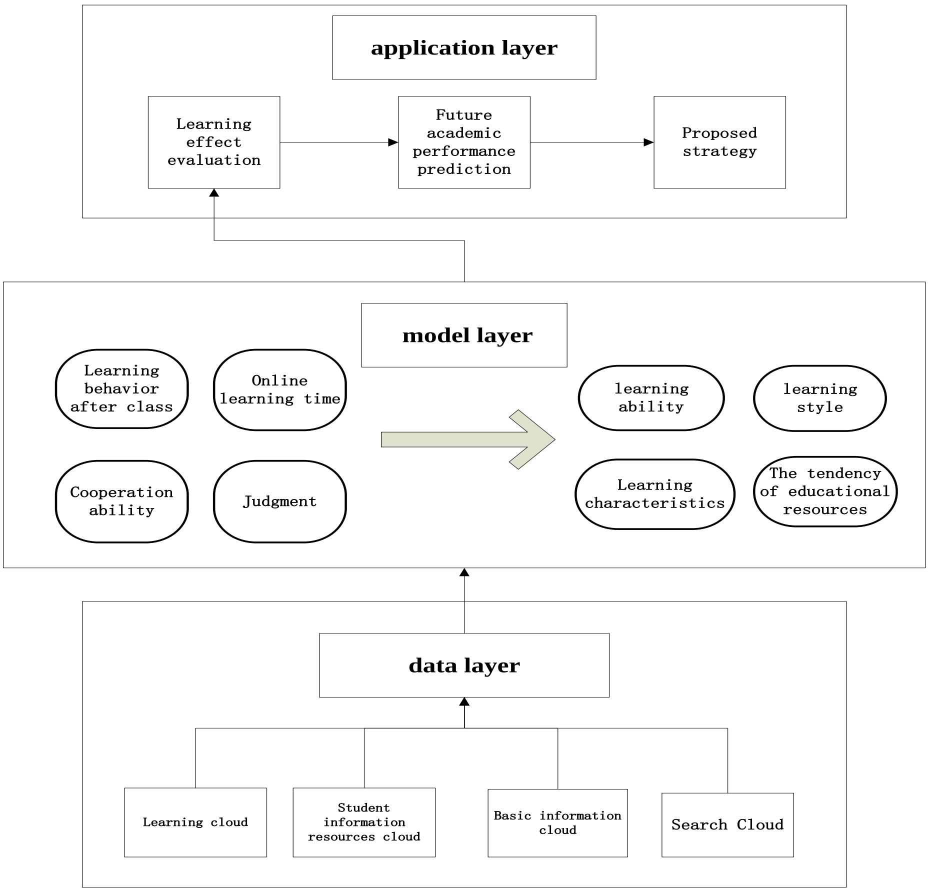

Figure 2. The Education Platform of Cloud Computing

(1) Data layer:

The information of the data layer mainly comes from the cloud server data, such as learning cloud, student information resources cloud and other educational resources. It is through the HDFS Hadoop technology to store the data information, and uses ZooKeeper, Hbase and other data processing and management tools to dynamically generate MapReduce tasks, so as to calculate them.

(2) Model layer:

In the model layer, Hadoop uses mathematical model to process the information which is stored in the data layer, and outputs the summary information. It includes the analysis of students' learning behavior, cooperative behavior, and other basic behavior. Through the analysis, we can grasp the basic information such as characteristics of students, learning ability, resource satisfaction, so as to put forward the suitable suggestions for their learning.

(3) Application layer:

The information of students and other educational resources are analyzed in the model layer. Then the results are obtained. These results include students' learning ability, students' personality traits and so on. Teachers can make use of these results to evaluate the current teaching effect and students' learning effect, and predict the effect of teaching and learning in the future. Through the analysis of the forecast results, the best teaching strategies are given to the teachers.

3. Parallels BPNN Model Based on Map-Reduce

BP is a kind of multilayer feedforward neural network, and its main training algorithm is the error back propagation algorithm. In 1986, the United States artificial intelligence expert for the first time proposed the concept of BP neural network. After more than 20 years of development and application, BP neural network model has a certain representation. It is widely used in many fields, such as the data compression, classification, pattern recognition and function approximation. It consists of three parts, namely the input layer, the hidden layer and the output layer. Its network structure as shown in Figure 3.

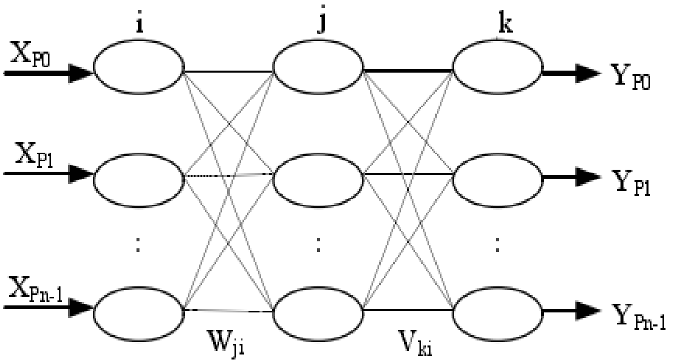

Figure. 3 The Network Structure of the BPNN

The number of nodes in the hidden layer of the BP network with 3 layers is p. w^1 and w^2 are the weight matrix of neural networks from the input layer to the hidden layer and from the hidden layer to the output layer. The excitation function of hidden layer is $\varphi(x)$, and the excitation function of output layer is $g(x)$.

The input of the hidden layer is:

$$y_j = \sum_{i}^{m} w_{ji}^{1} u_i \text{ , } (j = 1, 2, \cdots, n) \qquad (1)$$

The output of the hidden layer is:

$$y_i = \varphi(\sum_{i}^{m} w_{ji}^{1} u_i) \qquad (2)$$

The input of the output layer is:

$$o_s = \sum_{j=0}^{p} w_{sj}^{2} u_j \text{ , } (s = 1, 2, \cdots, n) \qquad (3)$$

The output of the neural network is:

$$o_s = g(\sum_{j=0}^{p} w_{sj}^{2} u_j y_i) \qquad (4)$$

The error of the output neuron is:

$$e_s = d_s - o_s$$

Among them, d_s is the expected output value of the s th neurons.

Define performance index function:

$$E = \frac{1}{2}\sum_{s=1}^{n} e_s = \frac{1}{2}\sum_{s=1}^{n} (d_s - o_s) \qquad (5)$$

According to the gradient descent method, we get the gradient of the performance index function on the threshold value of each weight.

$$\begin{cases} \dfrac{\partial E}{\partial w_{sj}^{2}} = -e_s g^{'}(o_s) y_i \\ \dfrac{\partial E}{\partial w_{ji}^{1}} = \sum_{s=1}^{n} [-e_s g^{'}(o_s) w_{sj}^{2} \varphi^{'}(y_j) u_i] \end{cases} \tag{6}$$

The basic idea of parallel neural network is divided into two parts. Firstly, we divide the blocks according to the location of the input data in the network, and use the gradient descent algorithm to calculate an iteration result of weight of the blocks. Then, we update the results in the whole network, and carry on the next iteration and update until we find the optimal solution of the network weights. In the Hadoop platform, the parallel method which is proposed in this paper has three main steps. Firstly, input data and set up a three layer parallel neural network. Secondly, according to the location of each node to block the data, and transfer M separate blocks to the Map function for processing. Thirdly, through the gradient descent method, the Map function find the weight distribution of each block by iterative algorithm. Then, we set the calculation result to (key, value), and save them. Among them, the key represents the weight of the corresponding position coordinates, and the value represents the size of the weight. Fourthly, we transfer the key-vlaue to the Reduce function, and update the statistics. Then, the Reduce function updates the weights of the input, so as to update the whole network. Finally, repeat the update process of weight computation. Through several iterations, the optimal solution of the objective function is found, and the weight distribution of the network is obtained.

4. Simulation Experiment and Result Analysis

As the input data set of the BP network, the training sample plays an important role in the training of the network. The data set is selected scientifically will determine whether the BP network can be used for prediction, as well as forecast results. In order to understand the students from various aspects, the teaching system needs to track the learning status of students, and obtains some factors related to learning achievement from the traditional teaching experience. The main factors are the students' online learning time, students' learning ability, homework, test scores and so on. We take these state information as the input data of BP network. First, we give a list of the original data in Table 1. Because of the large amount of data, we only give a part of the original data.

Table 1. The Original Data

Nember	online learning time	learning ability	homework	test score A	test score B	test score C	The total learning achievement
1	40	strong	90	95	85	80	excellent
2	35	normal	72	79	83	79	good
3	30	normal	70	59	72	63	normal
4	15	weak	60	52	43	45	fail
5	33	normal	70	75	68	72	normal
6	26	normal	72	59	68	77	normal
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

As the dimensions of the various indicators are different, so we cannot make a direct comparison. In order to make the index have comparability, and to speed up the convergence rate of the neural network, this paper has carried on the normalized processing to each index. Then, we train the parallel BPNN model. After many tests, the system has a minimum of fitting residuals when the nodes number of hidden layers is 3. Therefore, this paper is a 6-3-1 model. Training results are shown in Figure 4.

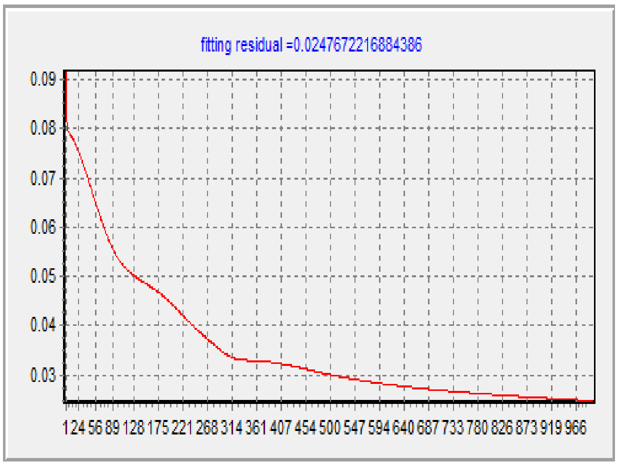

Figure 4. The Number of Hidden Layer Nodes is 3 in Neural Network Training (N=1000)

In order to verify the performance of the high performance computing system under the background of big data, this paper uses the Hadoop cluster environment is made up of 8 SYSTEM X3850 IBM server, each server is a quad core PC, each core as a Hadoop node. One of them is also used as NameNode, JobTracker and DataNode. The rest of nodes are DataNode and TaskTracker. After that, we use the fast Fourier-Transform algorithm to generate the test data sets of 3 different sizes: small data sets (S, 100MB), medium data sets (M, 300 MB) and big data sets (L, 500MB). We take different data sets as the sample data set of the parallel BPNN algorithm, and compare with the serial BPNN and parallel BPNN algorithm. Experimental results are shown in Figure 5.

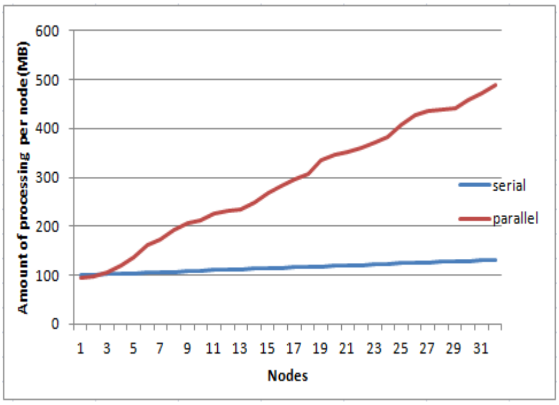

Figure 5. Data Processing Capability with Different Nodes

As can be seen from Figure 5, the amount of data processed by the two BPNN algorithms in a same node dimension is different. The amount of data processed by parallel BPNN algorithm is less than traditional BPNN algorithm. However, with the gradual increase of nodes, the amount of data processed by the two algorithms have gradually produced a difference. Although with the increase of nodes, the two algorithms can deal with more data. But the parallel BPNN algorithm is more obvious and more outstanding.

The speedup of the system is defined as the ratio of the execution time of serial algorithm to the parallel algorithm to solve the same problem on the same system. If the serial execution time of a parallel program is $T(1)$, the parallel execution time is $T(N)$ on the N nodes, the speedup of the program is $S(N) = T(1)/T(N)$. In fact, $T(1)$ consists of two parts, namely serial execution part T_S and parallel execution part T_p. Therefore, the speedup can be expressed as $S(N)=\frac{T(1)}{T(N)}=\frac{T_S+T_P}{T_S+T_P/N}$.

Next, we give the effect chart of the speedup of the parallel BPNN algorithm and parallel SVM algorithm.

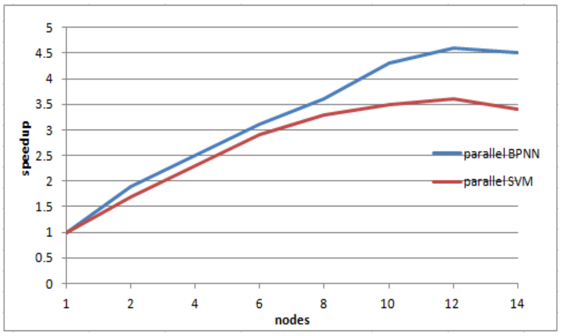

Figure 6. The Speedup

As is known to all, the speedup of the algorithm for the perfect parallel system is close to 1. But in practical applications, with the increase of the number of nodes, the consumption of the network transmission between nodes increases, so the linear speedup is very difficult to achieve. From Figure 6, we can see that with the increase of data set, the speedup of the algorithm of this article is close to the linear increase, especially for big data sets. In practice, the greater the amount of data, the higher the speedup of the parallel BPNN algorithm. That is to say, the algorithm of this paper can meet the demand of the forecast of the high dimensional data of the teaching system.

Finally, we use the parallel neural network model which is established in this paper to predict the learning achievement of students with different sample sizes. The experimental results show that the prediction model established in this paper has a better prediction accuracy. In addition, with the increase in the amount of data, the accuracy of the prediction is gradually increased.

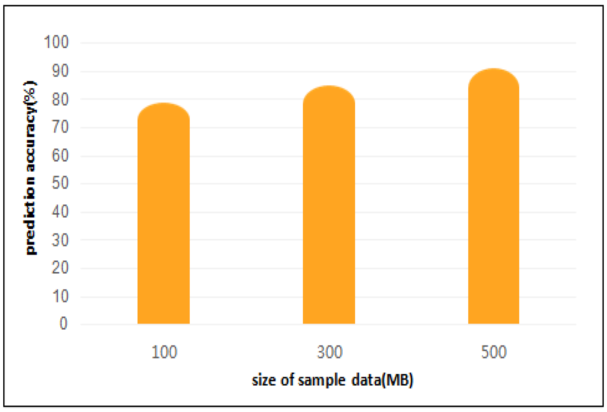

Figure 7. Accuracy of the Prediction

5. Conclusions

In this paper, based on the characteristics of the distribution of education resources, we put forward the method to analyze big data of education by using Hadoop technology. This method uses the MapReduce programming model to manage the data, so as to improve the speed and efficiency of data analysis. Secondly, in Hadoop platform, this paper puts forward the method of parallel BP neural network in education data processing. The method consists of the following main steps: firstly, input data and set up a three layer parallel neural network. Secondly, according to the location of each node to block the data, and transfer M separate blocks to the Map function for processing. Thirdly, through the gradient descent method, the Map function finds the weight distribution of each block by iterative algorithm. Fourthly, we transfer the key-value to the Reduce function, and update the statistics. Finally, repeat the update the calculation process of weight. After several iterations, the optimal solution of the objective function is found, and the weight distribution of the network is obtained. Finally, we simulate the parallel BP neural network algorithm based on education cloud platform, in order to prove that it is suitable for the prediction of learning achievement of the network teaching system.

References

[1] Hadoop. http:// hadoop.apache.org/

[2] S. Ghemawat, H. Gobioff and S. T. Leung, "The google file system. In SOSP '03: Proceedings of the nineteenth ACM symposium on Operating systems principles", ACM, **(2003)**, pp. 29-43.

[3] J. Dean and. S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", In Proceeding of the 6th Symposium on Operating Systems Design and Implementation, San Francisco CA, **(2004)**, pp. 12.

[4] K. Michael and K. W. Miller, "Big Data: New Opportunities and New Challenges", IEEE Computer, no. 46, no. 6, pp. 22-24.

[5] H. Takabi, J. B. D. Joshi and A. G. Joon, "Security and Privacy Challenges in Cloud Computing Environments", IEEE Security & Privacy, vol. 8, no. 6, **(2010)**, pp. 24-31.

[6] F. Dengguo, "Research on security of cloud computing", Journal of software, no. 1, **(2011)**, pp. 71-83.

[7] M. Paprzycki, "Education: Integrating Parallel and Distributed Computing in Computer Science Curricula", IEEE Distributed Systems Online, vol. 7, no. 2, **(2006)**, pp. 6.

[8] T. C. Jepson, "The basics of reliable distributed storage networks", IEEE IT Professional, vol. 6, no. 3, **(2004)**, pp. 18-24.

[9] A. Dimakis, P. B. Godfrey and Y. Wu, "Network Coding for Distributed Storage Systems", IEEE Transactions on Information Theory, vol. 56, no. 9, **(2010)**, pp. 4539-4551.

[10] Q. Xiulei, Z. Wenbo and W. Jun, "The status and challenges of distributed cache technology in cloud computing", Journal of software, no. 1, **(2013)**, pp. 50-66.

[11] L. Zhenqing, "Improvement of BP Training Algorithm for Artificial Neural Network and its Application in NDT", Measurement and control technology, vol. 20, no. 3, **(2001)**, pp. 56-58.

[12] H. Heng and W. Ruixiang, "Improved BP neural network in design of aircraft antiskid braking system", Journal of Beijing University of Aeronautics and Astronautics, vol. 30, no. 6, **(2004)**, pp. 561-564.

[13] H. C. Hsin, and M. Sun, "An adaptive training algorithm for back propagation neural networks", IEEE Trans on System s, Man, and Cybernetics, vol. 25, no. 3, **(1995)**.

[14] S. B. Hai and Z. X. Feng, "On Improved Algorithm of LMBP Neural Networks", Control Engineering of China, vol. 15, no. 2, **(2008)**, pp. 164-167.

[15] Y. H. Zweiri, J. F. Whidborne and L. D. Seneviratne, "A three-term back-propagation algorithm", Neurocomputing, vol. 12, no. 3, **(2003)**, pp. 305-318.

[16] L. E. Yu, Y. P. Xian and S. X. Bo, "Improved Algorithm of BP Neural Networks Based on the Activation Function with Four Adjustable Parameters", Microelectronics & Computer, vol. 11, no. 9, **(2008)**, pp. 89-93.