

# Research of neural network algorithm based on factor analysis and cluster analysis

Shifei Ding · Weikuan Jia · Chunyang Su ·  
Liwen Zhang · Lili Liu

Received: 16 September 2009 / Accepted: 23 June 2010 / Published online: 7 July 2010  
© Springer-Verlag London Limited 2010

**Abstract** Aiming at the large sample with high feature dimension, this paper proposes a back-propagation (BP) neural network algorithm based on factor analysis (FA) and cluster analysis (CA), which is combined with the principles of FA and CA, and the architecture of BP neural network. The new algorithm reduces the feature dimensionality of the initial data through FA to simplify the network architecture; then divides the samples into different sub-categories through CA, trains the network so as to improve the adaptability of the network. In application, it is first to classify the new samples, then using the corresponding network to predict. By an experiment, the new algorithm is significantly improved at the aspect of its prediction precision. In order to test and verify the validity of the new algorithm, we compare it with BP algorithms based on FA and CA.

**Keywords** Artificial neural network (ANN) · Factor analysis (FA) · Cluster analysis (CA) · FA-CA-BP network

## 1 Introduction

Artificial neural network (ANN) is a kind of cross-subject, which combines with Brain Science, Neuroscience,

Cognitive Science, Psychology, Computer Science, and Mathematics [1]. It has many important applications in nature science, such as Earth Science [2], Environmental Science [3], and Physical Science [4]. Artificial neural network simulates the structure of the human brain neural network and some working mechanism to establish one kind of computing model. Artificial neural network has some characteristics such as self-adaption, self-organization and real-time learning, and powerful ability in dealing with processing non-linear problem and large-scale computation. Neural network has been more than 60 years until now. During these years, hundreds of network algorithm models have been proposed [5], and back-propagation (BP) neural network is one of the most mature and most widespread algorithms [6]. Artificial neural network is convenient for people to solve the problems, but it is not perfect for the feature of the input samples and the properties of the network's structure. For example, a large number of original samples can be used to provide available information, while also increase the difficulty to deal with these data for the neural network, there is some related, or even repeated information which exists in the features of the samples. If we take all of its data as the network input, it will be detrimental to the design of the network, and will occupy a lot of storage space and computing time. Too many feature inputs and repeated training samples will lead to time-consuming work and hinder the convergence of the network, finally affect the recognition precision of the network. So it is necessary to pre-process the original data, analyze and extract useful variable features from a large amount of data, excluding the influences of the related or duplicate factors. It is also important to reduce the feature dimensionality as far as possible under the premise of not affecting the solution of the problems and then classify the similar samples in order to simplify the network structure.

---

S. Ding (✉) · W. Jia · C. Su · L. Zhang · L. Liu  
School of Computer Science and Technology,  
China University of Mining and Technology,  
Xuzhou 221008, China  
e-mail: dingsf@cumt.edu.cn

S. Ding  
Key Laboratory of Intelligent Information Processing,  
Institute of Computing Technology, Chinese Academy  
of Sciences, Beijing 100080, China

These works will improve the training speed, convergence, and generalization ability of the network, so as to improve the recognition precision of the network.

Integrating the original data process and neural network organically, the improved algorithm of the neural network is one hot spot at present. By integrating principal component analysis (PCA) with neural network, Zhang [7] used it in clinical diagnosis, diagnosed the swelling through measuring the amount of nucleoside in urine. Lewis et al. [8] used to analyze complex optical spectroscopy and time-resolved signal, to measure the food quality; He [9] made the comprehensive evaluation analysis on the development level of economic. These examples all reduce the dimensionality by PCA, and eliminate the influence of noise, so as to simplify the training process. Melchiorre combined cluster analysis (CA) and NN to carry on the landslide sensitivity analysis, imported distance measure to classify the samples, thus to exclude the influence of relevant factors to choose significant category, so as to improve the prediction ability of the neural network [10]; Lin et al. [11] used the neural network based on CA and FCA to test the blast furnace data. Combined independent component analysis (ICA) and NN, Gopi [12] used ICA to extract the features of forged images, to identify the forged images through NN; Song et al. [13] analyzed the high-spectral remote sensing data. By the integration of factor analysis (FA) and NN, Pan [14] carried through spectral calibration, extracted real data of the ultraviolet spectrum, and tested the components of the mixture; Janes [15] has established forecasting model through analyzing the farm smell, thus to find out the components of smell and the occurrence process. Huo used in molecular sieve data analysis, FA can successfully compress the input signal data, make a small amount of information loss, and speed up the neural network computing. These improvements mostly reduce dimensionality of the features, reduce the network input, simplify the network structure, and then enhance the computing speed of the network. Relatively speaking, these improvements have not been able to consider the characteristics of the samples [16].

Factor analysis is a kind of analysis method, which changes many variables into the several integrated variables [17, 18]; it concentrates the information of the system's numerous original indexes and saves to the factors, and it can also adjust the amount of the information by controlling the number of the factors, according to the precision that the actual problems need. FA can be seen as a further promotion of the PCA, but the data after dimensionality reduction by FA include more original information than by PCA. Cluster analysis is to compare the nature of the samples, eliminate the relevant factors of the samples and the influence of the noise, and cluster the similar samples, so as to analyze to the point.

Reducing the feature dimensionality of the original data by FA, to reduce the neural network input; classifying the data that have been reduced dimensionality by CA, one category as a small sample, respectively to train each small sample to improve the forecasting precision. Based on FA and CA, the new BP neural network algorithm is set up and verified the practicality through the example analysis.

## 2 The principle of the algorithm

### 2.1 The basic principle of FA

The basic idea of FA is to classify the observation variables, and make the higher correlation that is more closely linked variables in the same category, while the lower correlation among variables in the different categories. In fact, each category represents a basic structure, which is a public factor. The research problem is trying to use the least number of unpredictable public factors of the sum of the linear function and the special factors to describe each component of the original observation. The basic problem of FA is to determine the load factor by the related coefficient between variables.

Suppose  $X$  is the processing samples, where  $n$  is the number of the samples, and  $m$  is the number of the variables, that is  $X = (x_{ij})_{n \times m}$ .

In FA model, we suppose the observation random vector  $X_i = x_1, x_2, \dots, x_m$  and unobservable vector  $F_j = F_1, F_2, \dots, F_p$ ,

$$X_i = \sum_{j=1}^m a_{ij}F_j + c_i\varepsilon_i \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, p) \quad (1)$$

where  $m > p$ ,  $a_{ij}$  is the factor loading, which represents the related coefficient of the  $i$ -variable and the  $j$ -factor. It reflects the importance of the  $i$ -variable in the  $j$ -factor.  $F$  is the public factor, which appears together in the expression of each original observation variable. They are independent unobservable theoretical variables.  $c_i$  represents the load of the unique factor.  $\varepsilon_i$  affects the unique factor of  $X_i$ .

We suppose  $A = (a_{ij})$ ,  $A$  is known as loading factor matrix. When the structure of  $A$  is inconvenient to express the main factor, it is need to carry on a series of rotations on  $A$ , to make  $A$  subject to "the most simple structure rule", and make projection of each test vector on the new coordinate axis to polarize to 1 and 0 as much as possible. Then, we find out the best subset of each variable from numerous factors, finally compute the integrated scoring. These works can achieve the purpose of dimensionality reduction and reduce the input of the network. It is favorable for the design of the network.

### 2.2 System cluster theory

Cluster analysis is one method to study “birds of a feather flock together”. It classifies the samples according to the different relations of their characteristics. The different relations have two kinds of representations, one is distance coefficient, which is to regard each sample as a point in the  $m$ -dimensional (the number of the variable) space, then to define some distance between points and points in the  $m$ -dimensional coordinate axis; the other is similarity coefficient. Here, we use hierarchical cluster analysis (HCA). HCA is a kind of the most widespread method in the cluster analysis. The basic idea of the HCA is that there are  $n$  samples, each sample is one category, first, calculate  $C_n^2$  similar measures, and combine the two smallest measure samples into a category of two elements, and then to calculate the distance between this category and other  $n - 2$  samples according to some HCA method. In the process of combining categories, each step must make the combined categories to keep measure smallest in the system, reduce one category every time, until all the samples are classified as one category. In order to overcome the related influence between variables, the measure uses Mahalanobis distance and uses the commonly used method category average to cluster.

According to sample matrix  $X$ , the Mahalanobis distance between the sample  $x_i$  and  $x_j$  is denoted by  $d_{ij}$ , that is

$$d_{ij} = (x_i - x_j)' V^{-1} (x_i - x_j) \quad (i, j = 1, 2, \dots, n) \tag{2}$$

where  $x_i$  represents the  $i$ -vector, and  $V^{-1}$  is the inverse matrix of the sample covariance matrix. The covariance matrix of the sample is

$$V_{ij} = \frac{\sum_{k=1}^n (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{n - 1} \tag{3}$$

Method of category average calculates the distance between samples first, gains the distance matrix, finds out the smallest element from the distance matrix, and combines the two categories into a new category. Then, it calculates the category average distance between the two categories, denoted by  $D_{pq}$ , that is

$$D_{pq} = \frac{1}{n_p n_q} \sum_{i \in G_p, j \in G_q} d_{ij}^2 \tag{4}$$

where  $G_r = \{G_p, G_q\}$ ,  $n_r = n_p + n_q$ ,  $n$  represents the number of the samples in the categories. The distance between the  $G_k$  and  $G_r$  is

$$D_{kr} = \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2 \tag{5}$$

A new distance matrix can be gained. Then, combine the two categories that have smallest distance, by parity of

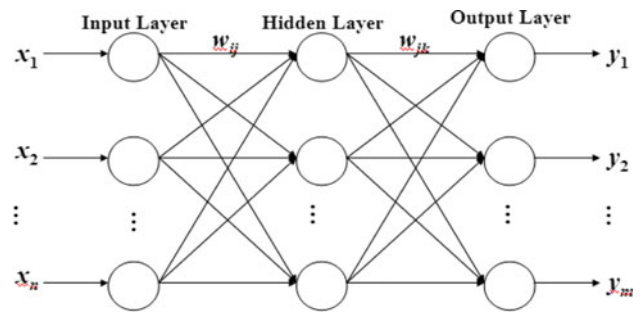


Fig. 1 The topology of three-layer BP neural network

reasoning, until combine to one category. Method of category average accords to spatial conservation, and also has monotonic, it is a widely used, effective method.

We should control the number of the categories by  $\lambda$ -cut value, for it will increase the complexity of the processing problems if there are too many categories, while it will not achieve the expected purpose if there're too few categories. According to the actual problems, generally we cluster the data after dimensionality reduction into 3–6 categories. Each category is seen as a new sample and regarded as the network input.

### 2.3 The principle of BP neural network

Artificial neural network is based on the intelligent computation, which uses the computer network system to simulate the biological neural network, is a non-linear, self-adapting information process system, which is composed by massive processing units, to process the information by simulating the way of processing and remembering the information by the cerebrum neural network. BP algorithm is composed by the forward spread of the data stream and the reverse spread of the error signal. The forward-propagating direction is input layer  $\rightarrow$  hidden layer  $\rightarrow$  output layer, the state of neurons of each layer only affect the neurons of next layer, if it can not obtain the expected output, then turns to the process of reverse spread of the error signal. Figure 1 shows the topology structure of the BP network [19].

Suppose that the network has  $n$  inputs,  $m$  outputs,  $s$  neurons on the hidden layer, the output of the middle layer is  $b_j$ , the threshold of the middle layer is  $\theta_j$ , the threshold of the output layer is  $\theta_k$ , the transfer functions of the middle layer is  $f_1$ , the transfer functions of the middle layer is  $f_2$ , the weight from input layer to middle layer is  $w_{ij}$ , the weight from middle layer to output layer is  $w_{jk}$ , then we can obtain the network output  $y_k$  through a series of relations, the expected output of the network is  $t_k$ , the output of  $j$ th unit of the middle layer is

$$b_j = f_1 \left( \sum_{i=1}^n w_{ij}x_i - \theta_j \right) \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, s). \tag{6}$$

To calculate the output  $y_k$  of the output layer through the output of the middle layer, that is

$$y_k = f_2 \left( \sum_{j=1}^s w_{jk}b_j - \theta_k \right) \quad (j = 1, 2, \dots, n; k = 1, 2, \dots, m). \tag{7}$$

Define the error function by the network actual output as follows

$$e = \sum_{k=1}^m (t_k - y_k)^2 \tag{8}$$

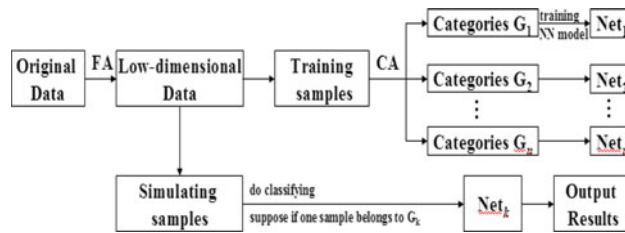
The network training is a continual readjustment process for the weight and the threshold value; the purpose is to make the network error reduce to a pre-set minimum or stop at a pre-set training step. Then, we input the forecasting samples to the trained network and obtain the forecasting result.

### 2.4 The new algorithm based on CA and FA

It has the superiority to use BP neural network to deal with the problems, and it plays an important role in practical application, but it has some limitations, for example, the speed of the convergence is slow, it is difficult to determine the number of hidden layers and hidden units, and the original data is singular, etc. In order to improve the recognition precision of the network, simply the network structure, and improve the network efficiency, the network algorithms are improved through pre-processing the original data. First of all, we reduce the dimensionality of original data by FA, and divide the treated data into the training samples and the simulation samples; then cluster the selected training samples by CA, regard the small samples after clustering as the training samples of the networks, train the networks separately with small category of samples, and determine the network structure, here the number of the neural networks is as same as the number of the clustering; finally classify the simulation samples one by one, and identify with trained network. Combining the neural network with CA and FA organically, we establish the BP algorithm based on FA and CA (FA-CA-BP). Figure 2 shows the flow chart of the new algorithm as follows.

The basic steps of the new algorithm is as follows:

*Step 1* Standardize the original data, and compute its correlation coefficient matrix  $R$ ;



**Fig. 2** The flow chart of the neural network algorithm based on FA and CA

*Step 2* Solve the eigenvalues of  $R$ , denoted by  $\lambda_1, \lambda_2, \dots, \lambda_n$ , arrange them as follows

$$\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_d \leq \dots \leq \lambda_n \tag{9}$$

and then solve the eigenvectors of  $R$ ;

*Step 3* Calculate variance contribution rate (VCR) of the  $i$ th eigenvalue and the cumulative contribution rate (CCR) expressed as follows

$$VCR(\lambda_i) = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \dots + \lambda_n} \tag{10}$$

$$CCR(d) = VCR(\lambda_1) + VCR(\lambda_2) + \dots + VCR(\lambda_d) = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_d}{\lambda_1 + \dots + \lambda_d + \dots + \lambda_n} \tag{11}$$

When  $CCR(d) \geq 80\%$  or  $90\%$ , we can determine the principal factor;

*Step 4* Rotate the factor loading matrix, calculate the factor score, regards it as a new sample, and divide the new sample into the training sample and the simulation sample;

*Step 5* Calculate the Mahalanobis distance between the training samples, gain a distance matrix  $D$ , and combine the two categories that are corresponding by the smallest element of  $D$  as a new category.

*Step 6* Calculate the average distance between the new category and others, gain a matrix  $D1$ , repeat *Step 5*, by parity of reasoning, determine the number of the clustering categories according to the question needs, and regard the category as a new small sample;

*Step 7* Design BP network structure according to the number of the extracted principal factors and the output of the actual problem, train the networks toward the new small samples, to determine the different networks;

*Step 8* Calculate the distance between the sample in the simulation sample and the category, classifies the sample to the category;

*Step 9* Use the trained network to test the simulation sample, and count the test results.

The main purpose for the original data is to reduce dimensionality by FA, reduce the network input, simplify the network structure, improve the training speed, save training time; the main purpose for clustering the samples

is to exclude the relativity and similarity between samples, so as to improve the recognition accuracy of the network.

In order to verify the usability of the new algorithm better, according to the above steps, we establish the BP algorithm based on FA (FA-BP) and the BP algorithm based on CA (CA-BP).

### 3 Experiments

Known data [20], we use it to forecast the occurrence degree of the wheat blossom midge in GuanZhong area. The weather condition has close relationship with the occurrence of the wheat blossom midge; use the weather factors to forecast the occurrence degree. Here, we choose the data of 60 samples from 1941 to 2000 as the study object, use  $x_1 - x_{14}$  to represent the 14 feature variables (weather factors) of the original data;  $Y$  represents the occurrence degree of the wheat blossom midge during that year. Use the standardized methods to process the original data (the processed data is still called  $X$ ). The forecast of the pest occurrence system, in essence can be seen as an input–output system, the transformation relations include the data fitting, the fuzzy transformation and the logic reasoning, these can all be represented by the artificial neural network.

In order to illustrate the problem, we use the BP neural network, BP neural network bases on FA and CA to test, respectively. The experiments choose the 40 samples from 1941 to 1980 as the training samples and use the 20 samples from 1981 to 2000 as the simulation samples. The processed results are listed in Table 1 (the error square sum is the square sum of the difference between the predicted value and actual value).

Using the MATLAB toolbox, we take  $X$  as the network input,  $Y$  as the network output, first establish BP neural network with 14 neurons in the input layer, 1 neuron in the output layer, and simulate the problem.

Using MATLAB to carry on factor analysis, according to FA-BP algorithm, we can extract 6 principle factors, then design the BP network, and get the forecasting results of the FA-BP algorithm. According to CA-BP algorithm, we cluster 40 samples to 3 categories, then take each small

category as a sample, train the network to determine the network structure, then classify the 20 simulation samples; according to the category to which the simulation samples belong, use the trained network to forecast, finally integrate the test results, gain the forecasting result of CA-BP algorithm; at last simulate refer to the step of FA-CA-BP, gain the forecasting results of the new algorithm.

Table 1 shows that, comparing the forecasting result of FA-BP algorithm with the BP algorithm, the precision of the forecast is not reduced, but the step of the convergence is reduced, the error sum of the forecasting result also has been reduced. It means that the data dimensionality is reduced by FA, the network input has been decreased, it is easy to design the network, the network structure is simplified, and the training speed of the network is improved. The forecasting precision of CA-BP algorithm has improved much greater than BP algorithm and FA-BP algorithm. Because of training the neural network to each category, it is not convenient to count the training steps, and the main purpose of the CA is to train more accurate network, improve the forecasting precision. Although FA-CA-BP algorithm increases the complexity, the forecasting precision, the network design and the convergence are satisfactory.

### 4 Results and discussion

This paper establishes the FA-CA-BP neural network algorithm that is based on the BP algorithm, integrated with FA and CA. The original data after processed by FA will have some loss of information, but it can reduce the data redundancy, excluding the influence of the related and repeated data, and reduce the feature dimensionality under the premise of without affecting the solution of problem, decrease the network input, and simplify the network structure. The case indicates that FA-BP algorithm has not reduced the forecasting accuracy but decreased the training steps, improved the speed of the convergence, saved the running time, and relatively reduced the error square sum of the results. Cluster the training samples by CA, classify the similar samples as one category, train the network toward each category separately, that is, each category determines a neural network. Then analyze the simulation samples by CA, classify the samples one by one, use the corresponding network to predict, finally to make the statistics to the forecasting results, the case indicates that the forecasting accuracy of FA-CA-BP algorithm has improved drastically, and the forecasting error is more reduced relatively.

Factor analysis is mainly to exclude the relevant factors of the data features, reduce the feature dimensionality, and optimize the network structure. Cluster analysis is to

**Table 1** The comparison of the performance of BP algorithm, FA-BP algorithm and FA-CA-BP algorithm

NN algorithm	Precision (%)	Training steps	Error sum squares
BP	75	2503	14.7593
FA-BP	75	837	8.5579
CA-BP	90	–	6.5128
FA-CA-BP	90	–	3.7511



remove the relevant factors of the samples, classify the similar samples into one category, and improve the recognition accuracy. Neural network has strong ability to deal with the non-linear problems. The new model that organically combines the advantages of these three algorithms can better fit the more complicated non-linear problems. Through the case analysis, the new model increases the complexity to the certain extent, but the forecasting precision can be improved drastically. Its validity has been confirmed.

Factor analysis is mainly to reduce the data dimensionality. Cluster analysis is mainly to classify the samples. While using the neural network to predict, it has some difficulty to design the network structure. The network astringency and the generalization ability are not very ideal. It is difficult to make full use of the superiority for the neural network to deal with the problems. If we design the network structure after reducing the dimensionality by FA, and train network after clustering by CA, it can make more use of the superiority of the new algorithm, and improve the forecasting precision and the efficiency in dealing with the problems. So the new algorithm is more suitable to deal with the big samples with higher feature dimensionality.

In the course of the study, we found that the theory organically combines the multi-statistical analysis with the neural network and improves the efficiency of the neural network in dealing with the problems to some extent. According to the characteristics of the problems, we can choose suitable statistical theory that combined with the ideal neural network model, to establish a new algorithm that integrates with other methods. It is possible to find a new discovery. Scholars pay more attention to these aspects such as improving the network structure and the algorithms, and integrating with other theories. They apply the neural network and its improvements in a wider field.

**Acknowledgments** This work is supported by the Basic Research Program (Natural Science Foundation) of Jiangsu Province of China under grant no. BK2009093, and the National Nature Science Foundation of China under grant no. 60975039.

## References

1. McCulloch W S, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 10(5):115–133

2. Pradhan B, Lee S (2007) Utilization of optical remote sensing data and GIS tools for regional landslide hazard analysis by using an artificial neural network model. *Earth Sci Frontier* 14(6):143–152
3. Pradhan B, Lee S (2010) Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. *Environ Model Softw* 25(6):747–759
4. Pradhan B, Lee S (2009) Landslide risk analysis using artificial neural network model focusing on different training sites. *Int J Phys Sci* 3(11):1–15
5. Cheng H, Cai X, Min R (2009) A novel approach to color normalization using neural network. *Neural Comput Appl* 18(3): 237–247
6. Rumelhart D, Hinton G, Williams R (1986) Learning representation by back-propagating errors. *Nature* 3(6):533–536
7. Zhang Y (2007) Artificial neural networks based on principal component analysis input selection for clinical pattern recognition analysis. *Talanta* 73(1):68–75
8. Lewis E, Sheridan C, Farrell M et al (2007) Principal component analysis and artificial neural network based approach to analysing optical fibre sensors signals. *Sens Actuators A Phys* 136(1):28–38
9. He F, Qi H (2007) Nonlinear evaluation model based on principal component analysis and neural network. *J Wuhan Uni Techn* 29(8):183–186
10. Melchiorre C, Matteucci M, Azzoni A et al (2008) Artificial neural networks and cluster analysis in landslide susceptibility zonation. *Geomorphology* 94(3–4):379–400
11. Lin S, Zhang D, Li W et al (2005) Neural network forecasting model based on clustering and principle components analysis. *Mini-Micro Syst* 26(12):2160–2163
12. Gopi E (2007) Digital image forgery detection using artificial neural network and independent component analysis. *Appl Math Comput* 194(2):540–543
13. Song J, Feng Y (2006) Hyperspectral data classification by independent component analysis and neural network. *Remote Sens Techn Appl* 21(2):115–119
14. Pan Z, Pan D, Sun P et al (1997) Spectroscopic quantitation of amino acids by using artificial neural networks combined with factor analysis. *Spectrochim Acta Part A Mol Biomol Spectrosc* 53(10):1629–1632
15. Janes K, Yang S, Hacker R (2005) Pork farm odour modelling using multiple-component multiple-factor analysis and neural networks. *Appl Soft Comput* 6(1):53–61
16. Huo W (2007) Research on BP neural network based on factor analysis and its application in rational synthesis of microporous materials. Jilin University, Jilin
17. Anderson T (1984) An introduction to multivariate statistical analysis, 2nd edn. Wiley, New York
18. Speraman C (1904) General intelligence objectively determined and measured. *Am J Psychol* 15:201–293
19. Shi Z (2009) Neural network. Higher Education Press, Beijing
20. Zhang Y (2003) The application of artificial neural network in the forecasting of wheat midge. Northwest A&F University, Xianyang