Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Water quality prediction based on improved BP neural network

Wenming Xue, Xiru Yuan

School of Southwest Petroleum University, Cheng Du 610500, China.

Abstract: Taking the drinking water quality standard as an example, this paper selects five indicators in the water quality monitoring data of drinking water to form a data sample, and analyzes the sample. Finally, a three-layer neural network with two hidden layers is established. The program is written in MATLAB language and the neural network model is implemented on the MATLAB platform. The processed sample data is used for the learning training of the established BP neural network. A good network predicts water quality indicators. Comparing the predicted results with the actual monitoring results, the results are objective and reasonable.
Keywords: BP neural network; Water quality evaluation; MATLAB; prediction.

---

1. INTRODUCTION

Water environmental quality assessment is the basis of all work in water environmental management. Traditional evaluation methods such as single factor evaluation and comprehensive pollution method are questioned because of the limitations of their application. Therefore, it is particularly important to seek an objective and versatile water quality assessment method. In recent years, the outstanding performance of BP neural network in pattern recognition has made it possible. Applying BP neural network to water quality assessment can overcome the shortcomings of traditional evaluation methods and provide a possibility for vertical comparison of river water quality categories.

However, in order to achieve the specified error requirements or achieve a certain number of training times, BP neural network needs to carry out considerable learning and training times, constantly modify the weight and threshold of each network layer, making the BP network water quality assessment model face two major problems - - Work efficiency and identification accuracy issues. This paper explores these two major issues and conducts in-depth research on improving the application of BP neural network in water quality assessment to evaluate water quality grades.

2. BP NEURAL NETWORK

2.1 Structure of BP neural network

The BP network is a multi-layer feedforward neural network based on the error back propagation algorithm. It is mainly composed of the input layer, the hidden layer and the output layer. The input signal is transmitted in the front-to-back direction. In order to explain the structure and working principle of the neural network more vividly, the following figure depicts a typical three-layer neural network with input layer, hidden layer and output layer, which will be organically interconnected

between layers. Connected, neurons in the same layer do not have connections, but there can be one or even multiple layers of hidden layers, with arrows indicating the flow of information.

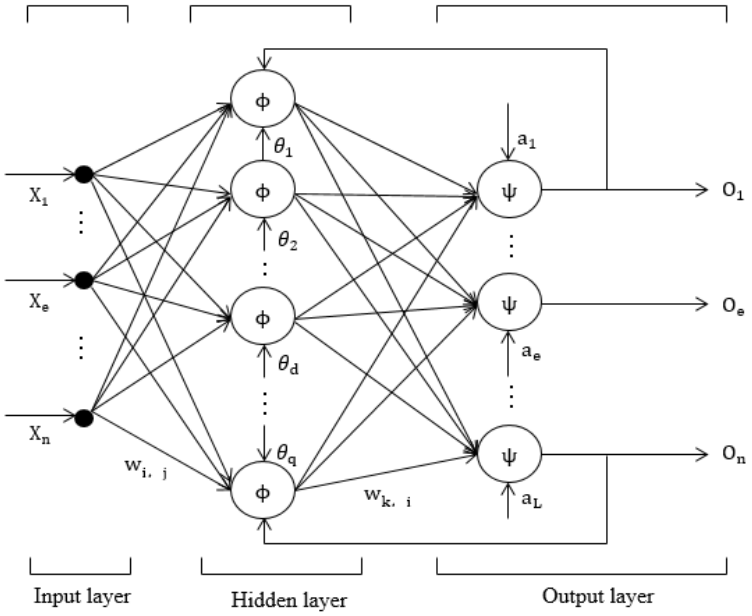

Figure 1 Typical three-layer BP neural network structure

Among them, the external input is represented by the letter x, and the subscript indicates the order of the nodes. The weights of the layers in the network are always represented by the letter W, and the subscripts are also the order of the nodes. i, j in the above figure represents the input layer node j and the hidden layer node i, and k, i represents the output layer node k and the hidden layer node i. The thresholds of the hidden layer and the output layer are represented by the letters θ and a, respectively, and the subscripts are used to indicate the order of the nodes. The excitation function of the hidden layer is represented by the letter φ; The excitation function of the output layer is represented by the letter ψ; The output part is represented by the letter o.

2.2 BP network algorithm

According to the parameters shown in the figure above, the error back propagation calculation process of the neural network algorithm is as follows:

The adjustment formula of the output layer weight is

$$\Delta w_{ki} = -\eta \frac{\partial E}{\partial w_{ki}} = -\eta \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial w_{ki}} = -\eta \frac{\partial E}{\partial o_k} \frac{do_k}{dnet_k} \frac{\partial net_k}{\partial w_{ki}} \tag{1}$$

The output layer threshold adjustment formula is

$$\Delta a_k = -\eta \frac{\partial E}{\partial a_k} = -\eta \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial a_k} = -\eta \frac{\partial E}{\partial o_k} \frac{do_k}{dnet_k} \frac{\partial net_k}{\partial a_k} \tag{2}$$

The hidden layer weight adjustment formula is

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} = -\eta \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial w_{ij}} = -\eta \frac{\partial E}{\partial y_i} \frac{dy_i}{dnet_i} \frac{\partial net_i}{\partial w_{ij}} \tag{3}$$

The hidden layer threshold adjustment formula is

$$\Delta \theta_i = -\eta \frac{\partial E}{\partial \theta_i} = -\eta \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial \theta_i} = -\eta \frac{\partial E}{\partial y_i} \frac{dy_i}{dnet_i} \frac{\partial net_i}{\partial \theta_i} \tag{4}$$

Because

$$\frac{\partial \mathrm{E}}{\partial o_k} = -\sum_{p=1}^{P} \sum_{k=1}^{L} \left(T_k^p - o_k^p\right) \tag{5}$$

$$\frac{\partial net_k}{\partial w_{ki}} = y_i \,; \frac{\partial net_k}{\partial a_k} = 1 \,; \frac{\partial net_i}{\partial w_{ij}} = x_j \,; \frac{\partial net_i}{\partial \theta_i} = 1 \tag{6}$$

$$\frac{\partial \mathrm{E}}{\partial y_i} = -\sum_{p=1}^{P}\sum_{k=1}^{L}\left(T_k^p - o_k^p\right)\cdot\psi'(net_k)\cdot w_{ki} \tag{7}$$

$$\frac{\partial y_i}{\partial net_i} = \phi'(net_i) \tag{8}$$

$$\frac{\partial o_k}{\partial net_i} = \psi'(net_k) \tag{9}$$

So finally get the following formula:

$$\Delta w_{ki} = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}\left(T_k^p - o_k^p\right)\cdot\psi'(net_k)\cdot y_i \tag{10}$$

$$\Delta a_k = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}\left(T_k^p - o_k^p\right)\cdot\psi'(net_k) \tag{11}$$

$$\Delta w_{ij} = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}\left(T_k^p - o_k^p\right)\cdot\psi'(net_k)\cdot w_{ki}\cdot\phi'\cdot(net_i)\cdot x_j \tag{12}$$

$$\Delta \theta_i = \eta\sum_{p=1}^{P}\sum_{k=1}^{L}\left(T_k^p - o_k^p\right)\cdot\psi'(net_k)\cdot w_{ki}\cdot\phi'\cdot(net_i) \tag{13}$$

The following figure shows the program flow of the BP network algorithm:

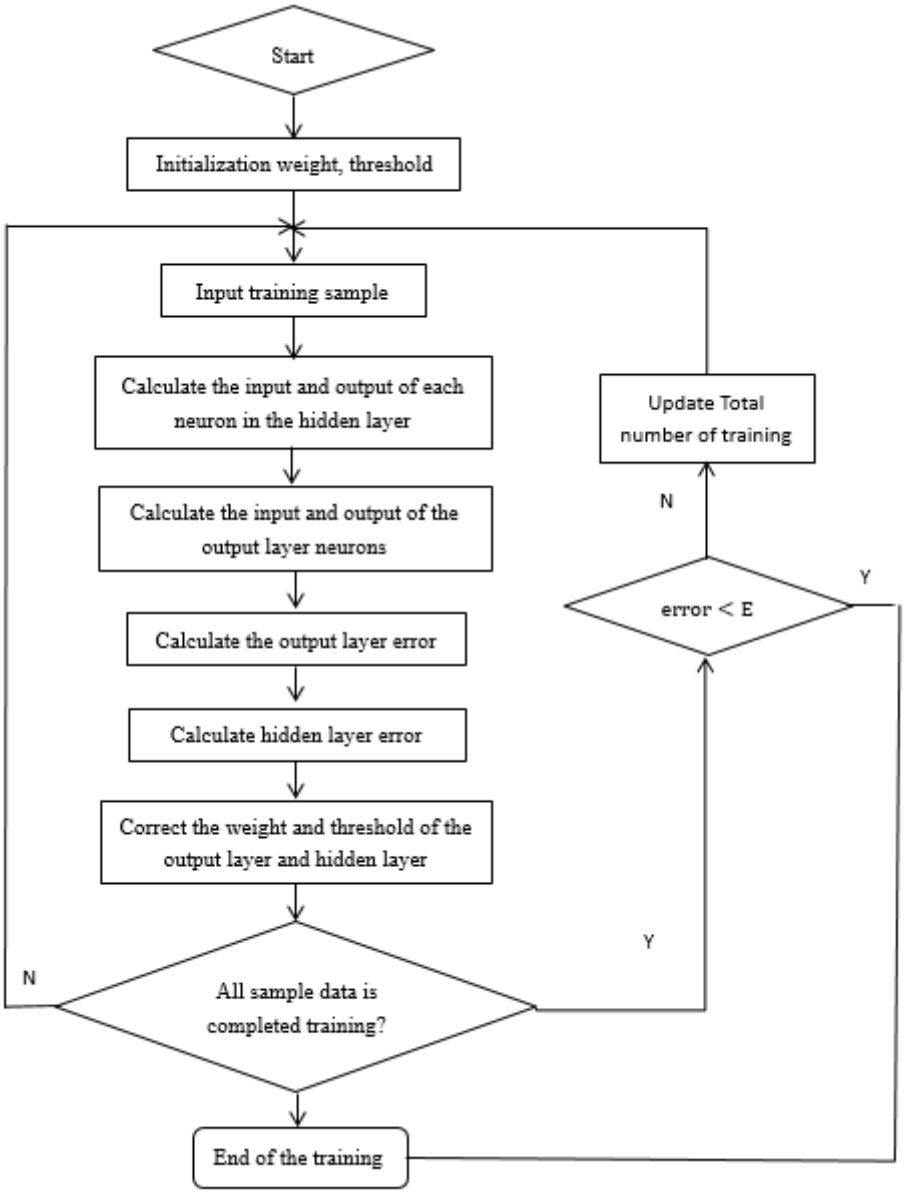

Figure 2 BP algorithm program flow chart

2.3 BP network algorithm improvement

The main disadvantage of the BP network algorithm is: The convergence speed is slow; Easy to fall into local minimum; It is difficult to determine the number of hidden and hidden nodes.

The BP network can implement a nonlinear mapping of input and output, but it does not depend on the model. The association information between the input and the output is distributed and stored in the connection right. The BP network's ability to approximate nonlinear mapping and generalization is an advantage. The slow convergence is a big disadvantage, which affects the practical application of the network in many aspects. Therefore, many people have studied the learning algorithm of BP network extensively and proposed many improved algorithms. The L-M algorithm can shorten the learning time and improve the network accuracy.

The basic idea of L-M algorithm

The traditional BP algorithm uses the steepest gradient descent method to correct the weight. The training process gradually reaches the minimum point from a certain starting point along the slope of the error function to make the error zero. For complex networks, the training process may fall into a local minimum. The change from this point to multiple directions increases the error so that it cannot escape this local minimum point. The basic idea of the LM algorithm is to make it not follow a single negative gradient direction for each iteration, but to allow the error to search along the direction of deterioration, while adapting between the steepest gradient descent method and the Gauss-Newton method. Adjustment to optimize the network weight, so that the network can effectively converge, greatly improving the network convergence speed and generalization ability. The L-M method is also called the damped least squares method, and its weight adjustment formula is:

$$\Delta \mathrm{w} = (J^T J + \mu I)^{-1} J^T e \tag{14}$$

Where J is the Jacobian matrix of the error-weight differential, e is the error vector, and μ is a scalar. When μ is large, the above formula is close to the gradient method with a small learning rate; when μ is small, the above formula It becomes the Gauss-Newton method. Therefore, the L-M method is a smooth blend between the steepest gradient descent method and the Gauss-Newton method. In this method, μ is adaptively adjusted.

Steps of the L-M algorithm

Send all the inputs to the network and calculate the output of the network. Another error function is used to calculate the sum of squared errors of all targets in the training set.

Calculated the Jacobian matrix J of the error versus the weight differential.

The Δw is obtained by the formula (14).

The sum of the squares of the errors is repeated using w + Δ w. When the sum of squared errors is reduced to a certain target error, the algorithm is considered to converge.

3. BUILD PREDICTIVE MODELS

3.1 Select prediction standard

In this paper, the five types of water quality standards in the Standard for Drinking Water Quality (GB5749-2006) are used as training samples, and the water quality index is taken as the water quality assessment model.

The classification criteria of temperature, PH, BOD_5, total hardness (calculated as $CaCO_3$) and conductivity are shown in Table 1.

Table1 Drinking water quality grading standard

Sample index	Grade				
	I	II	III	IV	V
temperature	The artificially caused changes in ambient water temperature should be limited to: the average weekly maximum temperature rise is not large, the maximum temperature drop is not more than 2; the temperature does not participate in water quality evaluation				
PH	6.50～6.90	6.90～7.30	7.30～7.70	7.70～8.10	8.10～8.50
BOD_5	<3	3～4	4～6	6～10	>10
total hardness	<150	150～250	250～350	350～450	450～550
conductivity	50～300	300～500	500～700	700～900	900～1100

Table2 The degree of pollution corresponding to the water quality level□

Water quality level	I	II	III	IV	V
Degree of pollution	No pollution	Slight pollution	Moderately polluted	Severe pollution	Inferior water

3.2 Selection of sample data

The sample data is the test data for drinking water. The specific data is shown in Table 3.

Table3 Drinking water quality index measured value

Serial number	temperature(˚C)	PH	BOD_5 (mg/L)	total hardness(mg/L)	conductivityμS/cm	Grade
1	25.12	7.21	3.29	236	472	II
2	25.36	7.33	5.06	285	570	III
3	26.15	8.16	10.02	505	1010	V
4	24.85	6.89	2.95	126	252	I
5	24.94	7.02	3.75	207	414	II
6	24.62	6.68	1.49	114	228	I
7	25.35	7.41	4.92	273	546	III
8	25.46	7.57	5.25	312	624	III
9	25.72	7.84	7.30	393	786	IV
10	25.18	7.18	3.69	228	456	II
11	25.92	7.96	8.56	413	826	IV
12	26.34	8.28	10.14	513	1026	V
13	25.62	7.56	5.74	272	544	III
14	25.05	7.03	3.36	216	432	II
15	24.98	6.96	3.25	193	386	II
16	24.73	6.61	1.72	96	192	I

17	25.72	7.83	6.97	365	730	Ⅳ
18	25.42	7.39	4.83	279	558	Ⅲ
19	26.36	8.25	10.34	531	1062	Ⅴ
20	25.83	7.92	8.06	373	746	Ⅳ
21	24.96	6.97	3.63	156	312	Ⅱ
22	24.88	6.62	1.21	69	138	Ⅰ
23	24.85	6.75	2.06	103	206	Ⅰ
24	25.32	7.35	4.26	262	524	Ⅲ
25	26.21	8.38	10.10	476	952	Ⅴ
26	26.36	8.42	10.25	494	988	Ⅴ
27	25.97	7.96	9.04	421	842	Ⅳ
28	25.26	7.14	3.55	226	452	Ⅱ
29	24.89	6.78	1.69	132	264	Ⅰ
30	25.42	7.56	4.95	274	548	Ⅲ
31	26.25	8.31	10.08	465	930	Ⅴ
32	25.94	7.83	7.62	383	766	Ⅳ
33	24.86	6.73	2.02	117	234	Ⅰ
34	24.98	7.02	3.18	176	352	Ⅱ
35	25.47	7.56	5.21	302	604	Ⅲ
36	26.20	8.32	10.26	459	918	Ⅴ
37	25.66	7.62	5.69	321	642	Ⅲ
38	24.86	6.76	2.58	134	268	Ⅰ
39	25.15	7.06	3.35	201	402	Ⅱ
40	25.85	7.91	8.74	406	812	Ⅳ
41	26.31	8.47	10.43	535	1070	Ⅴ
42	26.02	8.01	9.45	447	894	Ⅳ
43	25.14	7.02	3.14	182	364	Ⅱ
44	24.64	6.73	1.77	109	218	Ⅰ
45	24.87	6.86	2.25	134	268	Ⅰ
46	25.52	7.49	4.87	276	552	Ⅲ
47	26.32	8.38	10.62	531	1062	Ⅴ
48	25.84	7.76	8.12	396	792	Ⅳ
49	25.97	7.82	7.41	415	830	Ⅳ
50	26.20	8.23	10.41	507	1014	Ⅴ

3.3 The design of network structure

(1)The design of input and output layer

For the input layer, this paper selects the five water quality factors related to it as the input of the sample, namely temperature, PH,BOD_5, total hardness (calculated as $CaCO_3$) and conductivity, ie the neurons of the input layer. The output node is one, and the output result is the water quality evaluation level, which is represented by 1, 2, 3, 4, and 5, which represent Class I, Class II, Class III, Class IV, and Class V, respectively.

(2)The design of hidden layer

In the network design process, the determination of the number of neurons in the hidden layer is very important. Excessive number of neurons in the hidden layer will increase the amount of network calculation, and it is easy to produce over-fitting problems; if the number of neurons is too small, it will affect the network performance and will not achieve the expected results. The number of hidden layer neurons in the network is directly related to the complexity of the actual problem, the number of neurons in the input and output layers, and the setting of the expected error. At present, there is no clear formula for determining the number of neurons in the hidden layer. Only some empirical formulas, the final determination of the number of neurons needs to be determined according to experience and multiple experiments.

In this paper, we refer to the following empirical formulas on the problem of selecting the number of neurons in the hidden layer, and make a preliminary selection according to the formula, then make appropriate adjustment according to the training results, and finally determine the number.

$$L = 2n + 1 \tag{15}$$

Where n represents the number of neurons in the input layer; L represents the number of neurons in the hidden layer.

According to the above empirical formula, the number of neurons in the hidden layer can be calculated. In this experiment, the number of neurons is 11.

(3)Network structure

In summary, the BP neural network structure established in this paper is a 5-11-1 three-layer network. The network structure is as follows:

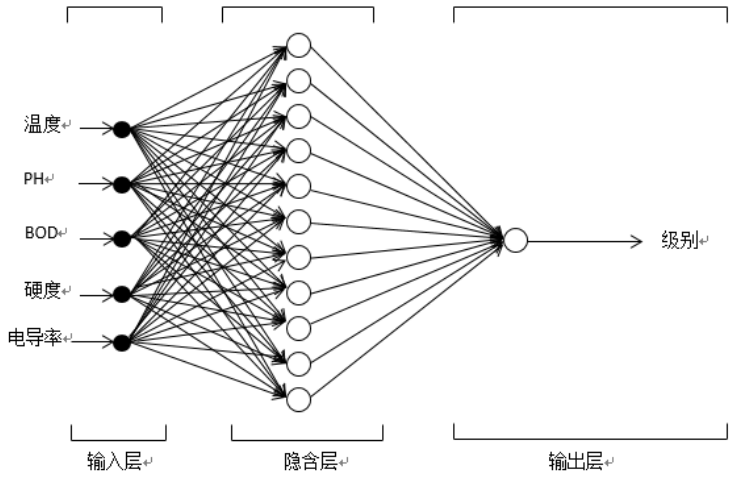

Figure 3 BP neural network structure

4. SIMULATION RESULTS

The improvement of the performance of the neural network during the training process is shown in Figure 4. This performance measurement is based on mse (minimum mean square error) and is displayed in logarithmic log. It can be seen from the figure that the curve is rapidly declining, and the green circle indicates where the performance of the data is best.

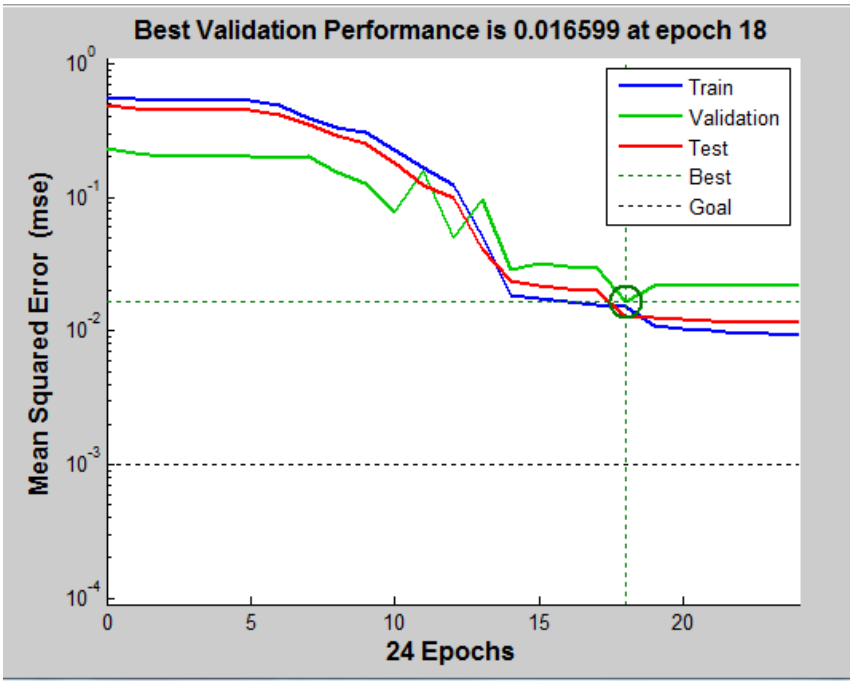

Figure 4 BP network training curve

Figure 5 shows the test results of the training data. It can be clearly seen from the curve that the curve fit between the measured values of the training and the network predictions is very high, which indicates that the predicted values of the training can correctly reflect the actual water quality level.

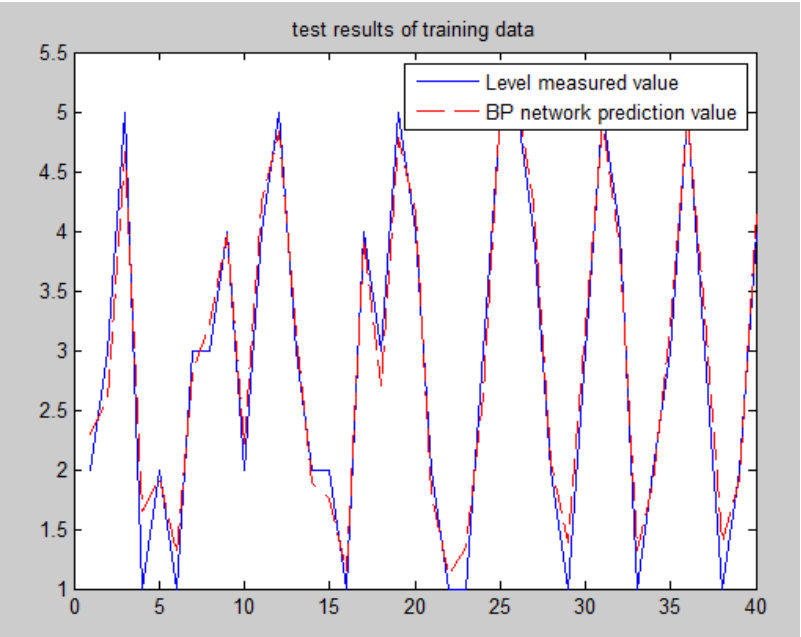

Figure 5 Test results of training data

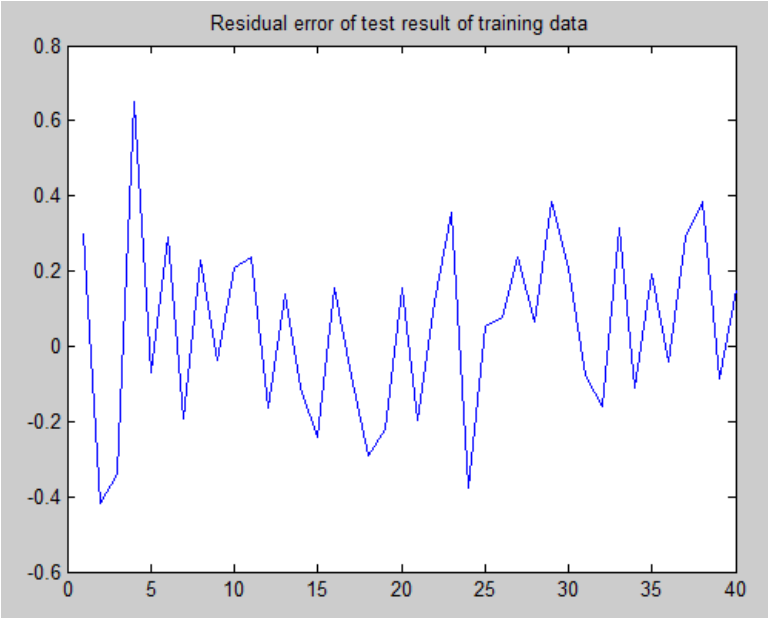

Figure 6 Residual error of test result of training data

Figure 7 shows the test results of the test data. It can be clearly seen from the curve that the curve fit between the measured value and the network prediction value is also high, which indicates that the predicted value of the test can correctly reflect the actual water quality level.

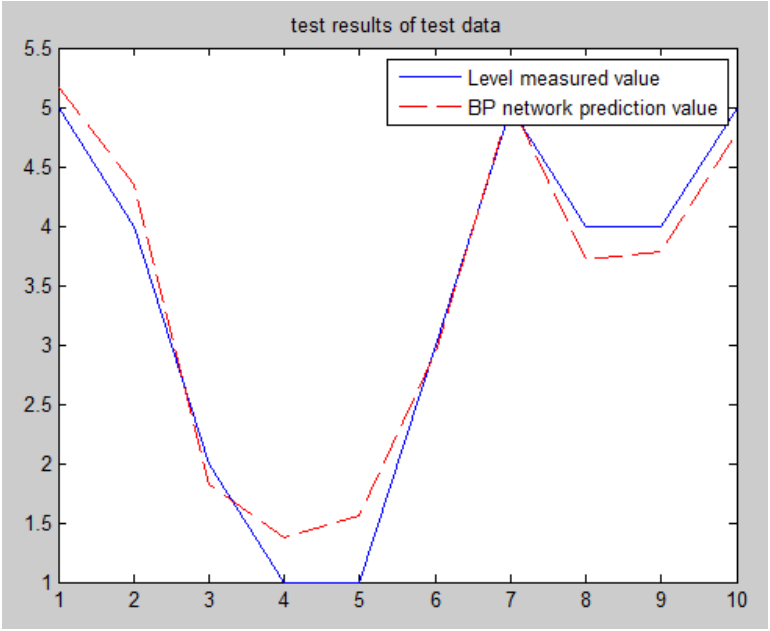

Figure 7 Test result of test data

The mean square error of the training data is 0.060338, and the test data is 0.080920. In order to display the relative error changes more intuitively, the error results are plotted as a line graph, as shown in Figures 10 and 11. It can be clearly seen from the line graph that the relative error of the training data is relatively obvious, and the maximum error reaches 0.6 or more. The reason is that the training data is too small, so the accuracy is relatively poor. The relative error of the test data is relatively stable, and only the relative error of the fifth group is large.

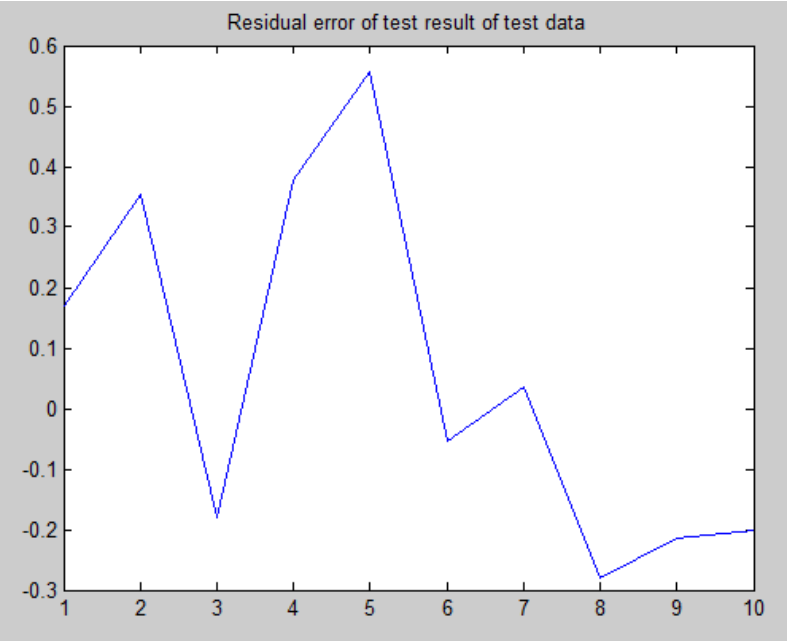

Figure 8 Residual error of test result of test data

```
>> MAILAB
mse=
 0.060338
Relative error:
0.148396-0.138693-0.0677660.649484-0.0359240.291300
mse=
 0.080920
Relative error:
0.0341850.088439-0.0899850.3786960.556868-0.0179000
```

Figure 9 Mean variance and relative error of training results

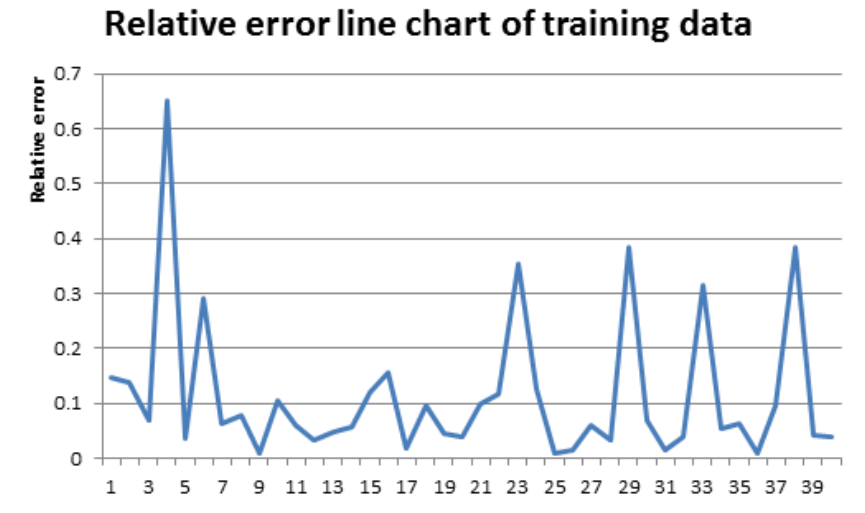

Figure 10 Relative error line chart of training data

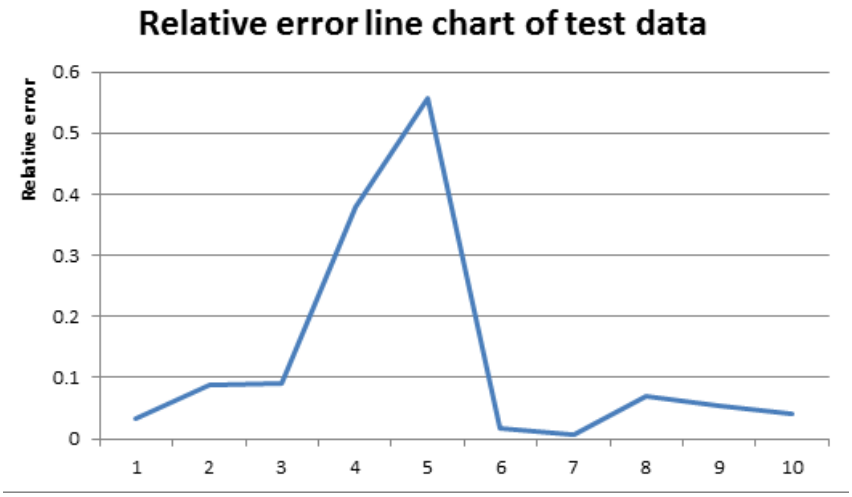

Figure 11 Relative error line chart of test data

Finally, the predicted water quality level of the BP neural network is obtained:

1x40 double

	1	2	3	4	5	6	7	8	9	10	11	12	1
1	2.2968	2.5839	4.6612	1.6495	1.9282	1.2913	2.8086	3.2304	3.9635	2.2087	4.2382	4.8359	
2													
3													
4													
5													
6													
7													

Figure 12 Prediction results of training data

1x10 double

	1	2	3	4	5	6	7	8	9	10
1	5.1709	4.3538	1.8200	1.3787	1.5569	2.9463	5.0343	3.7199	3.7859	4.7985
2										
3										
4										
5										
6										
7										

Figure 13 Prediction results of test data

For example, the first group has a predicted value of 2.2968, indicating that the water quality level of the first group is grade III (moderate pollution); the predicted value of the sixth group is 1.2913, indicating that the water quality grade of the reorganization is grade I (no pollution).

5. SUMMARY

In this paper, the basic principle and derivation process of BP neural network are elaborated. Based on the national drinking water quality standard, the improved BP neural network model, the BP neural network of Levenberg-Marquardt rule training forward algorithm, is established. The model reflecting the water quality level of drinking water analyzes and verifies the rationality and applicability of the method. The modeling process is simple, the simulation results are objective and reasonable, and the application is convenient.

The results show that BP neural network predicts 40 sets of training data and 10 sets of test data respectively. The predicted results are relatively close to the actual values, and the prediction results are more accurate. Although the error is relatively large, the water quality level obtained by rounding

is still very accurate. It is indicated that the BP neural network improved by L-M algorithm is feasible and effective for water quality evaluation. It can accurately evaluate the water quality grade of drinking water, and the prediction accuracy is much higher than the traditional method. In practice, it has the value of further research and development, has a good application prospect, and provides a basis for the protection and prevention of water environment.

REFERENCES

[1] Yuanbin Hou,Jingyi Du,Mei Wang. Neural Networks [M]. Xi'an:Xi'an University of Electronic Science and Technology Press,2007:16-21,55.

[2] Wei Du. Exploration of water quality evaluation and prediction based on neural network [D].Tianjin:Tianjin University,2007.

[3] Qun Miao,Hui Yuan,Changfei Shao.Zhiqiang Liu.Water Quality Prediction of Moshui River in China Based on BP Neural Network[J].Qingdao:Qingdao Technological University,2009.

[4] Hao Zhulin,Zhang Yuanyuan,Feng Minquan.Water Quality Assessment Based on BP Network and Its Application[J].Xi'an:Xi'an University,2007.

[5]Xiaoqing Guo. Water quality monitoring and evaluation system based on neural network model [J]. Chongqing Environmental Science,2003,25(5):8-10.

[6] Guo Jinsong,Li Zhe.Artificial neural network modeling of water quality of the Yangtze River system:a case study in reaches crossing the city of Chongqing[J].Chongqing:Chongqing University,2009.8(1):1-9.

[7] Guohao Song. Application Research of Artificial Neural Network in Water Quality Simulation and Water Quality Evaluation [D].Chongqing:Chongqing University,2008.