Journal of Environmental Sciences Vol.17, No.6, pp.946—952, 2005

ISSN 1001—0742
CN11—2629/X

Article ID: 1001-0742(2005)06-0946-07 **CLC number**: X131; Q178.1 **Document code**: A

Application of Bayesian regularized BP neural network model for analysis of aquatic ecological data—A case study of chlorophyll-*a* prediction in Nanzui water area of Dongting Lake

XU Min[1,*], ZENG Guang-ming[1,2,*], XU Xin-yi[1], HUANG Guo-he[1,2], SUN Wei[1], JIANG Xiao-yun[1]

(1. College of Environmental Science and Engineering, Hunan University, Changsha 410082, China. E-mail: ykxumin@21cn.com; zgming@hnu.cn; 2. Sino-Canadian Center of Energy and Environment Research, University of Regina, Regina, SK, S4S 0A2, Canada)

Abstract: Bayesian regularized BP neural network(BRBPNN) technique was applied in the chlorophyll-*a* prediction of Nanzui water area in Dongting Lake. Through BP network interpolation method, the input and output samples of the network were obtained. After the selection of input variables using stepwise/multiple linear regression method in SPSS 11.0 software, the BRBPNN model was established between chlorophyll-*a* and environmental parameters, biological parameters. The achieved optimal network structure was 3-11-1 with the correlation coefficients and the mean square errors for the training set and the test set as 0.999 and 0.00078426, 0.981 and 0.0216 respectively. The sum of square weights between each input neuron and the hidden layer of optimal BRBPNN models of different structures indicated that the effect of individual input parameter on chlorophyll-*a* declined in the order of alga amount > secchi disc depth(SD) > electrical conductivity (EC). Additionally, it also demonstrated that the contributions of these three factors were the maximal for the change of chlorophyll-*a* concentration, total phosphorus(TP) and total nitrogen(TN) were the minimal. All the results showed that BRBPNN model was capable of automated regularization parameter selection and thus it may ensure the excellent generation ability and robustness. Thus, this study laid the foundation for the application of BRBPNN model in the analysis of aquatic ecological data(chlorophyll-*a* prediction) and the explanation about the effective eutrophication treatment measures for Nanzui water area in Dongting Lake.

Keywords: Dongting Lake; chlorophyll-*a*; Bayesian regularized BP neural network model; sum of square weights

Introduction

Over the last decade, many efforts on ANN (Artificial neural network) modeling of chlorophyll-*a* prediction with environmental parameters have been carried out (Chen, 2001; Recknagel, 1997; Lek, 1999; Whitehead, 1997). Despite the successful attempts to look at the problem from different aspects, these efforts are of a number of limitations and some problems may not have received sufficient attention and discussion. The common characteristics of the main relevant literature are summarized as follows: (1) When ANN is employed in chlorophyll-*a* prediction of the lake aquatic ecological system, generally, the disadvantage is that it can not ensure global optimization and overfitting problem. Based on the satisfactory achievement of BRBPNN (Bayesian regularized BP neural network) in the QSAR (quantitative structure-activity relationships) model (Burden, 1999; 2000), the global optimization can recur to the repeated running to acquire feasible results and the impact of overfitting problem on generalization ability can be avoided using regularization method. However, up till now, BRBPNN model has been seldom applied to the lake aquatic ecological system. (2) There are always interactions among parameters in the non-linear aquatic ecological system, but during the past application of ANN to chlorophyll-*a* prediction, most of the work has concentrated on environmental parameters as the input parameters of ANN (Recknagel, 1997; Whitehead, 1997; French, 1998; Karul, 2000) and ignored the impacts of biological parameters (alga amounts, alga biomass and dominant species etc.) on the output variables. Here we think that it is necessary to consider the non-linear relationships between chlorophyll-*a* and biological parameters and this can essentially supplement ANN model which only considers environmental parameters. (3) Most of the work did not based on the optimal choice of the input variables of ANN. Many use almost all possible environmental parameters as inputs (Joseph, 2003). For example, Karul *et al*. (Karul, 2000) included environmental variables of: PO_4-phosphorus, NO_3 and NH_3-nitrogen, alkalinity, suspended solids concentration (SS), pH, water temperature, EC, dissolved oxygen, solar radiation and SD, and the output variable is chlorophyll-*a*. Since the effects of some input nodes may be duplicated (e. g. generally SS strongly correlates with SD), it possibly results in the duplicate impacts of the inputs on the outputs, and the greater errors may be induced. (4) Not enough attention has been paid on the performance interpretation of the trained network. The issue of evaluating neural network performance goes far beyond assessing the ability of the network to generalize to a validation or testing data set. Causal analysis is clearly a weakness of neural network analysis in most of the discussed work. It is generally believed that it is difficult to interpret ANN by its nature. However, it is worthwhile looking at the trained model in depth in order to extract some knowledge from the learning process, which may be possible in some cases (Joseph, 2003).

In this paper, we used the BRBPNN to predict the chlorophyll-*a* trend in Nanzui water area. The advantage of this model is that it can automatically select the regularization parameters and integrate the characteristics of high convergent rate of traditional BPNN and prior information of Bayesian statistics. Then we attempted to realize the following purposes: (1) to obtain the network with good robustness, fitting and generalization ability; (2) to investigate the feasibility of BRBPNN to predict chlorophyll-*a* trend; (3) to attempt to discover some inherent information of the network;

Foundation item: The Natural Science Foundation of China for Distinguished Young Scholars(No. 50225926; 50425927), the Teaching and Research Award Program for Outstanding Youth Teachers in Higher Education Institutions of MOE, P.R.China(TRAPOYT) in 2000, the Doctoral Foundation of Ministry of Education of China (No. 20020532017) and the Hi-Tech Research and Development Program(863) of China(No. 2001AA644020; 2003AA644010; 2004AA649370); * Corresponding author

(4) to provide some bases for the effective eutrophication treatment of Nanzui water area in Dongting Lake.

1 Theories and methods

1.1 Study area

Dongting Lake is located to the south of Jingjiang River in the lower reaches of Yangtze River, and lies in the North of Hunan Province. Xiangjiang River, Zijiang River, Yuanjiang River and Lishui River, known as the "Four Rivers", lie to its west, south and east. As Dongting Lake is of fertile soil, sufficient rainfall and rich natural resources, it provides good habitat for aquatic animals and plants (Bu, 2002; Jin, 1995; He, 2003). Thus, it is of great significance to promote the research of aquatic ecology for the synthetical renovation and development of Dongting Lake.

The studied region of our paper is Nanzui water area situated in the western Dongting Lake (Fig.1), which is the entrance and converging place of Yangtze River and Lishui River into Dongting Lake. Owing to the special geographical characteristics of Nanzui water area, it is of much more complex characteristics of uncertainty and non-linearity (MacKay, 1992) in the ways of hydrological conditions, water quality and natural or human factors, which leads to a lot of aquatic ecological response to environmental conditions and makes the prediction of ecological dynamics (chlorophyll-a prediction) remain a very difficult problem in this area.

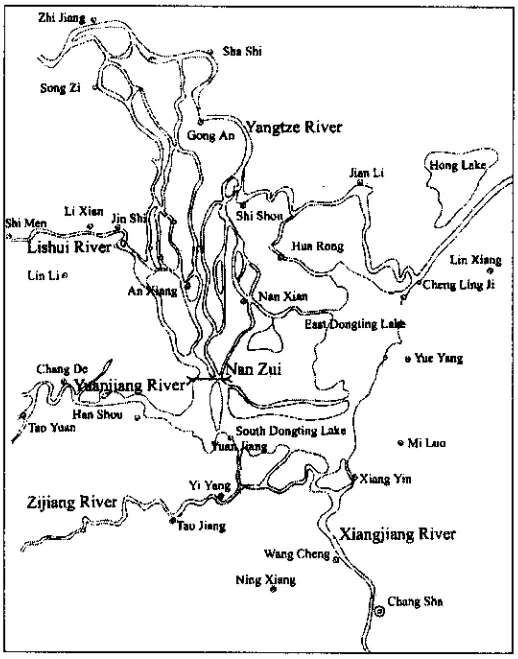

Fig.1 Nanzui water area of Dongting Lake

1.2 Theory of Bayesian neural network

In this paper, the regularization method can be used to improve generalization ability of the network and the training objective function F is denoted as:

$$F = \alpha E_W + \beta E_D, \tag{1}$$

where E_W is the square sum of weight of the network, E_D is the square sum of residual between network response values and objective values, α and β are the parameters of objective function (regularization parameter), whose values demonstrate that the emphasis of the network training depends on the decrease of output residual or the network volume. It is well known that the crux of regularization method is how to select and optimize the parameters of objective functions through Bayesian statistics. If α and β are regarded as stochastic variables, the formula is given according to the Bayesian rules:

$$P(\alpha,\beta \mid D,M) = \frac{P(D \mid \alpha,\beta,M)P(\alpha,\beta \mid M)}{P(D \mid M)}, \tag{2}$$

where D is the training data, M is the used network model, and w is the weight of the network. According to Bayesian rule, if α and β are assumed to satisfy uniform distribution, then when the likelihood of $P(D \mid \alpha, \beta, M)$ is maximized, the probability of posterior distribution of α and β in Eq.(2) will be up to the maximal value. Assuming the residual and the weight are stochastic variables and based on the Bayesian rule, Eq.(3) is obtained:

$$P(w \mid D,\alpha,\beta,M) = \frac{P(D \mid w,\beta,M)P(w \mid \alpha,M)}{P(D \mid \alpha,\beta,M)}. \tag{3}$$

Supposing the residual and the weight accord with Gaussian distribution, then

$$P(D \mid w,\beta,M) = \frac{\exp(-\beta E_D)}{Z_D(\beta)}$$

and

$$P(w \mid \alpha,M) = \frac{\exp(-\alpha E_W)}{Z_W(\alpha)}. \tag{4a}$$

To ensure that $P(D \mid \alpha, \beta, M)$ becomes regularization factor of Eq.(3), Eq.(4b) can be denoted:

$$P(w \mid D,\alpha,w\beta,M) = \frac{\exp(-F(w))}{Z_F(\alpha,\beta)}. \tag{4b}$$

If Eq.(4a) and Eq.(4b) are taken into Eq.(3), Eq.(5) is achieved as follows.

$$P(D \mid \alpha,\beta,M) = \frac{Z_F(\alpha,\beta)}{Z_W(\alpha)Z_D(\beta)}, \tag{5}$$

where $Z_W(\alpha) = (\pi/\alpha)^{N/2}$, $Z_D(\beta) = (\pi/\beta)^{n/2}$, $Z_F(\alpha,\beta) = (2\pi)^{N/2}\det^{-1/2}(H)\exp(-F(w^{MP}))$, and $H = \beta\nabla^2 E_D + \alpha\nabla^2 E_W$ is Hessian matrix of the objective function (F). When logarithm method and derivation method are respectively applied to Eq.(5), and supposing differentiation equation is equal to 0, then $P(\alpha, \beta \mid D, M)$ is maximized and posterior probability of weight is minimized. At this moment, the formulas of α and β are expressed:

$$\alpha^{MP} = \frac{\gamma}{2E_W(w^{MP})}, \quad \beta^{MP} = \frac{n-\gamma}{2E_D(w^{MP})},$$

$$\gamma = N - \alpha^{MP}\mathrm{trace}^{-1}(H^{MP}), \tag{6}$$

where n is the number of sample set, N is the total amount of network parameters, γ is the number of effective parameters, which relatively have more impacts on the reduction of error function. Initially assuming α and β according with Levenberg-Marquardt algorithm, the minimal value of $F(w)$ can be obtained by iterative training of BRBPNN. Updating α and β based on Eq.(6), then obtain the optimal value of posterior distribution, search the minimal value of the new $F(w)$, and finally train iteratively until convergence. The detailed steps of BRBPNN are referred to the reference (Foresee, 1997).

1.3 Data sets and methods

1.3.1 Normalization of data

As original data of the lake aquatic ecological system have different units and magnitudes for different indicators, it is necessary to process the data by normalization method. The normalized method of this paper is expressed as:

$$x_i' = (x_i - u)/S_d, \qquad (7)$$

where x_i' is the normalized data, x_i is the original data; u and S_d are the average value and the standard error, respectively. After the normalization of the data, x_i' ranges from -1 to 1 which is favorable for network training. At the same time, the calculated results can be reconverted through inverse function of Eq.(7).

1.3.2 Network variables selection

It is well known that chlorophyll-*a* is an integrative indicator, which describes the biomass of phytoplankton (eutrophication) in the aquatic ecological system. So it is comparatively easy for us to select chlorophyll-*a* as the output variable of the network.

Based on the past work, all the possible impacted factors of chlorophyll-*a*, such as environmental parameters (SD, SS, TP, TN, pH, water temperature and DO) and biological parameters (alga amount, biomass of phytoplankton), may be selected as the initial input variables (Jin, 1995). But in this paper, in the selection process of the input variables of the network, stepwise/multiple linear regression method in SPSS 11.0 software should be employed. And at the same time, it is of great importance that the selection must take the followings into account: (1) whether there are strong relevant relationships between input variables and output variable; (2) whether there are duplicative characteristics among input variables (namely, whether there are close correlations among input variables in SPSS 11.0 software). For example, we should notice whether there are close correlations between SS and SD or between alga amount and alga biomass; (3) because of the particular geographical location of Nanzui water area, it is also necessary to consider whether the sediment, industrial wastewater and sewage wastewater from Yangtze River and Lishui River have impacts on the lake aquatic ecological system.

1.3.3 BP network interpolation

Generally, a large amount of samples for each parameter are needed for BPNN, but in fact, it is difficult to frequently monitor each parameter because of the restriction of experimental conditions and experimental expenditure, especially in poorer regions. Therefore, our aim is to obtain such data as much as possible to meet the network.

This paper will select one week as the period of short-term prediction, namely, to predict chlorophyll-*a* concentration of this week using the data one week ago. Now the routine monitoring frequency is once every month in Nanzui water area. Although these data can reflect the dynamic trend of aquatic ecological characteristics, it is not suitable for short-term prediction. Therefore, on the basis of keeping the characteristics of the original data, three values are interpolated between two adjoint monitoring data, so that Nanzui water area have the data every week for each parameter(Pei, 2004).

The detailed interpolation method is emulating function of the network in MATLAB 6.5(Pei, 2004). Where t(t = 1, 1.25, 1.5, 1.75, 2, ⋯, 11.75, 12) is time variable of normalization, and t' (t' = 1, 2, 3, ⋯, 12) is the original time variables denoted by u and S_d of t, in which integers sequentially represent actual monitoring data from January to December and decimals sequentially represent interpolated data. The network is trained regarding t' as input variables and regarding corresponding chlorophyll-*a* concentration as output variables, where the iterative step error is limited to 10^{-3} of SSE (sum square error). After changing input variables into t, regarding t as lateral axis, and regarding the content of each variable as longitudinal axis, at last the surveyed data (illustrated by " * ") and interpolated data (illustrated by " + ") are shown in the Fig.2 using emulating function. With repeated iteration, we can find out the optimal interpolation data, which are considered as the source of training set of the network.

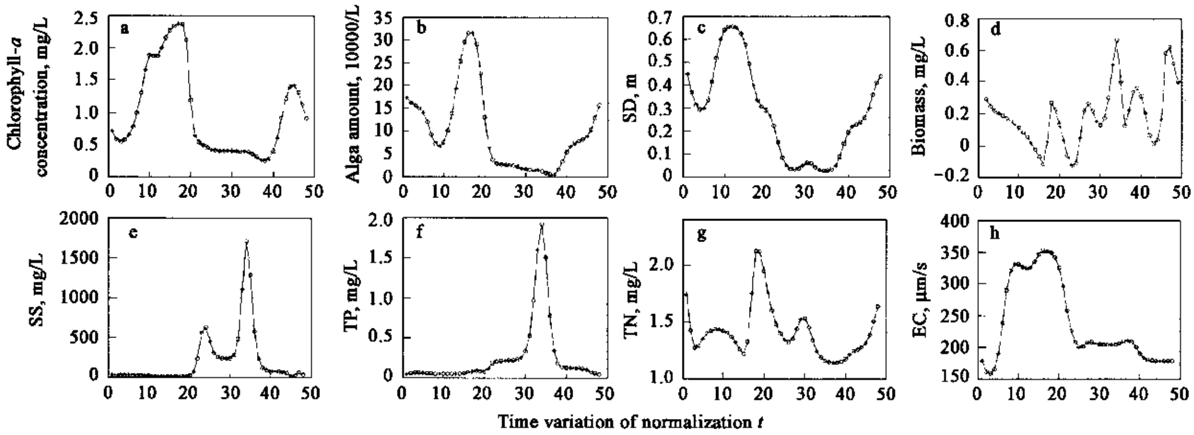

Fig.2 Fluctuation of chlorophyll-*a*, environmental and biological parameters
a. Chlorophyll-*a*; b. alga amount; c. SD; d. biomass; e. SS; f. TP; g. TN; h. EC

1.3.4 Selection of training set and test set

The training set of Nanzui water area is taken from the actual monitoring concentration and their interpolated data in 1999. And the test set comes from the actual monitoring data of each month in 2000(about 20% of total amount of data).

1.3.5 Weight of the network

It has been always one of the difficult research topics of the network on how to obtain effective information from BPNN model. To a certain extent, the performance of optimal training model reflects the complex non-linear relationships between input variables and output variable, which can be explained through network weight and bias information. When output layer only involves one neurode, the influences of input variables on output variable are directly presented in the influences of input parameters on the network. Simultaneously, in case of the connection along the paths from the input layer to the hidden layer and along the paths from the hidden layer to the output layer, it is attempted to study how input variables react to the hidden layer, which can be considered as the impacts of input variables on output variable.

According to the reference(Joseph, 2003), the relative importance of individual variable between input variables and output variable can be expressed as:

$$I = \frac{\sum_{j=1}^{H} ABS(w_{ji})}{\sum_{i=1}^{N}\sum_{j=1}^{H} ABS(w_{ji})}, \qquad (8)$$

where w_{ji} is the connection weight from i neurode in the input layer to j neurode in the hidden layer, ABS is an absolute function, N, H are the number of input variables and neurodes in the hidden layer, respectively.

1.3.6 Software and method

Stepwise/multiple linear regression method is realized through SPSS 11.0 software. And BRBPNN is debugged in neural network toolbox of MATLAB 6.5.

2 Results and discussion

2.1 Input selection

Based on the past research of Nanzui water area (Environmental protection monitoring station of Dongting Lake in Hunan Province, 2001), we selected alga amount, alga biomass, SD, SS, TN, TP and EC as the initial input variables of BRBPNN. In order to prevent duplicative training or prediction error (Joseph, 2003), it must be firstly considered whether there are close relationships among the input variables through stepwise/multiple linear regression method. From Table 1, when all the coefficients are equal to 0, significance probability of t test in statistics is lower than 0.05. This shows that alga biomass increases with the increase of alga amount and there is a significantly positive correlation between them. In addition, SD decreases with the increase of SS and there is a significantly negative correlation between them. So, this paper only selects alga amount and SD as input variables of the network among the four variables of alga biomass, alga amount, SD and SS.

Then stepwise/multiple linear regression method is also used to study the relationship between the input variables (alga amount, SD, TN, TP and EC) and the output variable (chlorophyll-a). Table 2 illustrates the detailed results about regression coefficients and statistical test and shows the three input variables (alga amount, SD and EC) are the main parameters for chlorophyll-a. And it also indicates that when all the coefficients are equal to 0, significance probability of t test is smaller than 0.05, which shows that these input variables have significant correlation with the output. Additionally, the tolerance value of each coefficient is greater than 0.1, which shows that there are no colinear characteristics among alga amount, SD and EC.

Table 1 Coefficients of linear regression model based on standardized training data

Parameters		Unstandardized coefficients		Standardized coefficients	t	Sig.	95% confidence interval for B		Colinearity statistics	
		B	S_d				Lower value	Upper value	Tolerance	VIF
Relation of alga	Constant	0.097	0.040		2.438	0.020	0.016	0.178		
amount and alga biomass	Alga amount	0.012	0.003	0.593	4.357	0.000	0.006	0.017	1.000	1.000
Relation of SS	Constant	0.361	0.059		6.105	0.000	0.229	0.493		
SD	SS	0.000	0.000	-0.633	-2.588	0.027	-0.001	0.000	1.000	1.000

Notes: VIF: Variance inflation factor; Sig.: Significance

Talbe 2 Coefficients of stepwise linear regression model based on standardized training data

Parameters	Unstandardized coefficients		Standardized coefficients	t	Sig.	95% confidence interval for B		Colinearity statistics	
	B	S_d				Lower value	Upper value	Tolerance	VIF
Constant	-0.565	0.161		-3.501	0.001	-0.890	-0.240		
Alga amount	0.033	0.006	0.419	5.170	0.000	0.020	0.046	0.535	1.868
EC	0.004	0.001	0.382	4.992	0.000	0.002	0.006	0.601	1.665
SD	0.935	0.292	0.273	3.204	0.003	0.347	1.523	0.485	2.063

According to the above analysis, this paper selects alga amount, SD and EC as input variables of the network. In order to verify the feasibility of this selection assumption based on stepwise/multiple linear regression method, the other two variables (TP, TN) are additionally regarded as inputs. By analyzing the network weights of five variables, the rationality of the assumption will be explained in the following chapter.

2.2 Determination of BRBPNN structure

Theoretically, it can fit all the continuous or limited disconnected functions using three-layer connected network with sigmoid transfer function in the hidden layer and linear transfer function in the output layer(The MathWorks, www. mathworks. com). Here we selected tangsig and pureline functions of MATLAB to improve efficiency. The number of neurodes in the input layer is determined based on the number of the selected input variables and the neurode in the output layer only includes chlorophyll-a. Generally, the number in the hidden layer of traditional BPNN is roughly confirmed through investigating the effects of the repeatedly tested network. But here, BRBPNN can automatically search the optimal value in posterior distribution (MacKay, 1992; Foresee, 1997). In order to show how to determine the number in the hidden layer, Fig.3 describes the trend of the neurode number S in the hidden layer under the condition of optimal network with different combination of input variables, 15 times of parallel training and 2000 epochs of maximal stopping. From Fig.3, when S is greater than 8, under the conditions of three BRBPNN model with different input variables, MSE of the training set and effective neurode number roughly tend towards stability. Therefore, we can determine that when S is greater than 8, three BRBPNN models can achieve satisfied results, but they might not be the optimal ones.

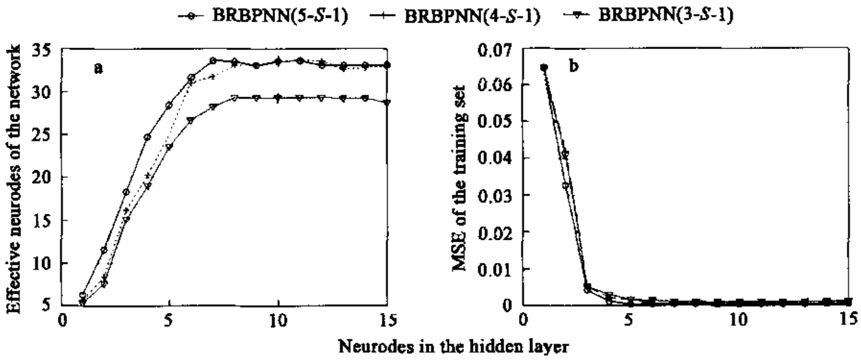

Fig.3 Changes of optimal BRBPNNs along with the hidden layer neuron number by parallel training for 15 times

2.3 Prediction results comparison

Fig.3 and Table 3 illustrate the calculation results of BRBPNN models with different combination of input variables and different number of neurodes in the hidden layer. Both the correlation coefficients and MSE fully indicate that there are strong non-linear relationships between chlorophyll-a and environmental parameters or biological ones. Simultaneously, lower training or prediction errors can be expressed.

However, among all the models, based on the results of Table 3 using trial method, the correlation coefficients and MSE of the test set of BRBPNN(3-S-1) are superior to both BRBPNN(4-S-1) and BRBPNN(5-S-1). This shows that TP and TN of input parameters may have few impacts on the chlorophyll-a and interfere the prediction precisions. Thus, BRBPNN(3-S-1) is found to be the optimal model.

In order to express the prediction results of BRBPNN(3-S-1) more clearly, Fig.4 illustrates the comparisons between experimental values and calculated values about the training set and the test set. It is found that calculated values are very close to experimental values and all of them are almost assembled near the isoline, which shows that BRBPNN(3-11-1) has advantageous abilities of good fitting and preventing overfitting problem.

Table 3 Results comparison in different models

Model*	Training set of normalization		Training set		Test set		Sum of neurode	Effective neurode
	E_D	E_W	R	MSE	R	MSE		
BRBPNN(3-8-1)	0.0353608	72.4331	0.999	0.00099263	0.978	0.0253	41	27.6261
BRBPNN(3-11-1)	0.0334685	73.4579	0.999	0.00078426	0.981	0.0216	56	28.0995
BRBPNN(3-14-1)	0.0338583	73.0816	0.999	0.00090410	0.980	0.0239	71	28.6092
BRBPNN(4-8-1)	0.0229142	83.3766	0.999	0.00053694	0.965	0.0387	49	31.5859
BRBPNN(4-12-1)	0.0156097	87.2292	0.999	0.00036578	0.976	0.0311	73	33.5299
BRBPNN(4-15-1)	0.0155977	87.2306	1.000	0.00036550	0.977	0.0306	91	33.5326
BRBPNN(4'-9-1)	0.0412793	48.7014	0.999	0.00096728	0.928	0.1527	55	27.5062
BRBPNN(4'-12-1)	0.0410104	49.0799	0.999	0.00096098	0.924	0.1566	73	27.2922
BRBPNN(4'-15-1)	0.0325375	47.8674	0.999	0.00076244	0.935	0.1287	91	28.7297
BRBPNN(5-8-1)	0.0186434	49.9676	1.000	0.00043686	0.861	0.2776	57	33.8115
BRBPNN(5-12-1)	0.0092904	62.3132	1.000	0.0002177	0.907	0.1327	85	36.4404
BRBPNN(5-15-1)	0.0975739	61.1969	1.000	0.00022864	0.899	0.1756	106	35.8623

Notes: * 3.Alga amount, SD, EC; 4.alga amount, SD, EC, TP; 4'.alga amount, SD, EC, TN; 5.alga amount, SD, EC, TP, TN

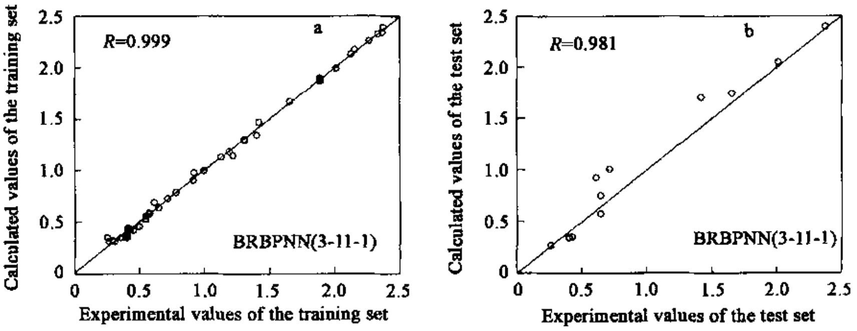

Fig.4 Results of BRBPNN(3 - 11 - 1)

2.4 Robustness of BRBPNN

Generally, initial weight at random always brings great trouble to the stability of traditional BPNN. Fig. 3 indicates that the optimal BRBPNN with different structures show better robustness and keeps the number of effective neurodes steady when S is greater than 8. Comparatively, BRBPNN(3-S-1) possesses the best robustness and keeps MSE of the training set and the test set within a smaller range stably. Additionally, it can be inferred that the worst robustness of BRBPNN(5-S-1) may be induced by the interference of TN and TP on the network training.

2.5 Weight interpretation

Eq. (8) is adopted to evaluate the significance of each input variable. The weight interpretation is carried out for the optimal BRBPNN. And the relative importance of individual input variable is calculated, as shown in Table 4. It is found that the contributions of TP and TN are the minimum; and alga amount, SD and EC are the maximal, which shows that these three factors have greater contributions to the network and also indicates these three factors are of more significance for chlorophyll-a. So, we can confirm that the analysis of network weights accords with the assumption of stepwise/multiple linear regression method in the former chapter and it is feasible to apply stepwise/multiple linear regression method to the selection of input variables. Moreover, among the three factors (alga amount, SD and EC), we can also find that alga amount of biological parameters has the same impacts as environmental parameters on chlorophyll-a. And so it is necessary to select biological parameters as input variables of chlorophyll-a prediction network.

Table 4 Sum of square weights (SSW) and the relative importance (I) from input neuron to hidden layer

Model		Alga amount	SD	EC	TP	TN
BRBPNN(3-11-1)	SSW	17.6823	14.8254	11.9473	-	-
	I, %	39.8	33.3	26.9	-	-
BRBPNN(4-15-1)	SSW	10.0920	9.9449	8.6869	6.1440	-
	I, %	31.8	31.4	27.8	9.0	-
BRBPNN(4'-15-1)	SSW	12.1549	10.8452	7.8818	-	2.7574
	I, %	35.0	31.3	25.6	-	8.1
BRBPNN(5-12-1)	SSW	9.9955	9.4471	7.2739	2.8569	2.3381
	I, %	30.4	28.7	22.1	8.7	7.1

2.6 Weight explanation for eutrophication treatment

According to the network performance analysis of Table 2 and Table 4, because chlorophyll-a has significantly relevant correlations with alga amount, SD and EC, it can be inferred that the three factors are limiting factors affecting chlorophyll-a content but TP, TN are not. And they can be explained to be linked with the practical entrophication situations of Dongting Lake as follows.

(1) Chlorophyll-a is one of the most important indicators for lake eutrophication and its concentration increases with the increase of alga amount. In Table 4, because the relative importance of alga amount is the maximal, it indicates that alga amount has the largest impacts on the chlorophyll-a. So we can deduce that it is the most important measure to prevent the propagation of alga from eutrophication taking place in Dongting Lake.

(2) In Table 4, SD is the second most important factor and this can be analyzed as follows: Nanzui water area locates at the converging place where Yangtze River and Lishui River enter into Dongting Lake, which leads to serious soil erosion, and a large amount of sediments results in the decrease of SD. Thus, we can conclude that the treatment of soil erosion (including Yangtze River, Lishui River and Dongting Lake) is an important way to control eutrophication of Dongting Lake.

(3) EC generally increases with the increase of conductible/electric impurity or ion concentration, which mainly results from discharged wastewater of Yangtze River and Lishui River. This also shows that the eutrophication of Dongting Lake can be effectively alleviated if we control the pollution of industrial wastewater and sewage wastewater in the two rivers.

(4) As well-known at present, there are no serious entrophication problems in Nanzui water area even if the concentration of TP and TN is relatively high. Furthermore, Table 4 also illustrates that the relative importance of TP and TN is the minimal and they don't belong to limiting factors. The above can be explained that Nanzui water area has an unique hydrological and hydraulic characteristics and water exchanging rate of Dongting Lake is smaller than 20 d. So, although the concentrations of TP and TN are relatively high, it is still difficult for the propagation of phytoplankton. Therefore, to protect the hydrologic conditions of Dongting Lake is also a significant action for the lake eutrophication prevention.

3 Conclusions

Using stepwise/multiple linear regression method in SPSS 11.0 software, the selection of input variables of BRBPNN model is realized and the variables with close interactive correlation among the input variables are successfully rejected, which prevents the inputs from duplicately affecting output results of the model and effectively keeps the errors of the network from increasing. Thus, with this method we can effectively solve the crux of the selection of input variables at present in the neural network of the lake.

On the basis of BRBPNN model, the aquatic ecological chlorophyll-a model of Nanzui water area is established with environmental parameters and biological parameters, and good prediction results are achieved. The achieved optimal network structure is 3-11-1 with the correlation coefficients and the mean square error for the training set and the test set as 0.999 and 0.00078426, 0.981 and 0.0216 respectively. Thus, this model is greatly helpful for chlorophyll-a prediction and can offer scientific basis for the eutrophication trend analysis of Nanzui water area.

The simple-effective quantitative method (including sum of square weights and the relative importance of individual input variable) is proposed. The sum of square weights between individual input neuron and the hidden layer of optimal BRBPNN models of different structure investigate the impact of individual parameter on chlorophyll-a declines in the order of alga amount > SD > EC. And the results also demonstrate that these three factors are limiting ones for the change of chlorophyll-a content, and TP and TN are not limiting ones, which also shows that alga amount of biological parameters has the same important influences as environmental parameters upon chlorophyll-a concentration based on the BRBPNN weight. So, this method is meaningful for discovering the inherent information of neural network and provides bases for the eutrophication treatment of Nanzui water area.

Bayesian regularization method automatically obtains the maximal values of posterior distribution. This conveniently deals with the selection of regularization parameter and avoid overfitting problem with good robustness. So, it can be concluded that BRBPNN model of chlorophyll-a prediction must be extensively applied in the future.

References:

Aguilera P A, Garrido Franich A, Torres J A *et al*., 2001. Application of the Kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality[J]. Water Research, 35(17): 4053—4062.

Barciela R M, Garcia E, Fernandez E, 1999. Modelling primary production in a coastal embayment affected by upwelling using dynamic ecosystem models and artificial neural networks[J]. Ecological Modelling, 120: 199—211.

Burden F R, Winkler D A, 1999. Robust QSAR models using Bayesian regularized neural networks[J]. J Med Chem, 42(16): 3183—3187.

Burden F R, Winkler D A, 2000. A quantitative structure-activity relationships model for the acute toxicity of substituted benzenes to tetrahymena pyriformis using Bayesian-regularized neural networks[J]. Chem Res Toxicol, 13(6): 436—440.

Bu Y X, Su S M, 2002. Balance analysis of nitrogen and phosphorus detained in Dongting Lake[J]. Yangtze River, 33(3): 23—25.

Chen Q W, 2001. Application of self-organization feature maps to analysis of aquatic data[J]. Journal of Hydraulic Engineering, (6): 8—13.

Conrad Lamon III E, Craig A S, 2004. Bayesian methods for regional-scale eutrophication models[J]. Water Research, 38: 2764—2774.

Environmental Protection Monitoring Station of Dongting Lake in Hunan Province, 2001. Research of aquatic ecological system in Nanzui water area of Dongting Lake[Z]. 1—55.

Foresee F D, Hagan M T, 1997. Gauss-newton approximation to Bayesian regularization[C]. In: Proceedings of the 1997 International Joint Conference on Neural Networks, Houston. 1930—1935.

French M, Recknagel F, Jarrett G L, 1998. Scaling issues in artificial neural network modelling and forecasting of algal bloom dynamics[C]. In: Proceedings of the International Water Resources Engineering Conference, ASCE(Abt S. R., Young-Pezeshk J., Watson C. C. ed.). Memphis, Tennessee, August 3—7.

He P, Wang B Z, 2003. Landscape ecological assessment and eco-tourism development in the South Dongting Lake Wetland [J]. Journal of Environmental Science, 15(2): 271—278.

Jin X C, 1995. Lakes in China(Research of their environment)[M]. China Beijing: China Ocean Press.

Jouko Lampinen, Aki Vehtari, 2001. Bayesian approach for neural networks-review and case studies[J]. Neural Networks, 14: 257—274.

Joseph H W L, Huang Y, Dickman M *et al*., 2003. Neural network modelling of coastal algal blooms[J]. Ecological Modelling, 159: 179—201.

Karul C, Soyupak S, Cilesiz A F *et al*., 2000. Case studies on the use of neural networks in eutrophication modelling[J]. Ecological Modelling, 134: 145—152.

Lek S, Guegan J F, 1999. Artificial neural networks as a tool in ecological modelling, an introduction[J]. Ecological Modelling, 120: 65—73.

MacKay D J C, 1992. Practical Bayesian framework for backprop networks[J]. Neural Computation, 4(3): 448—472.

Mark E B, Craig A S, Kenneth H R, 2004. A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis[J]. Ecological Modelling, 173: 219—239.

Pei H P, Luo N, Jiang Y, 2004. Application of back propagation neural network for predicting the concentration of chlorophyll-a in West Lake[J]. Acta Ecological Sinica, 24: 246—251.

Recknagel F, French M, Harkonen P *et al*., 1997. Artificial neural network approach for modelling and prediction of algal blooms [J]. Ecological Modelling, 96: 11—28.

The MathWorks, Inc. http://www.mathworks.com/access/helpdesk/help/pdf-doc/nnet/nnet.pdf. Natick, MA, USA[EB].

Whitehead P G, Howard A, Arulmani C, 1997. Modelling algal growth and transport in rivers: a comparison of time series analysis, dynamic mass balance and neural network techniques[J]. Hydrobiologia, 349: 39—46.

(Received for review March 10, 2005. Accepted April 25, 2005)