Journal of Computing and Electronic Information Management

ISSN: 2413-1660

Algae Bloom Forecasting Method Based on Multiple BP Network Integration Model

Li Wang, Chong Gao, Xiaoyi Wang, Xuebo Jin, Jiping Xu, Huiyan Zhang, Jiabin Yu, Qian Sun, Tingli Su

School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China

Abstract: With the rapid development of economy and continuous improvement process of industrialization, the eutrophication of water body has become a global environmental problem, so prevention of water eutrophication and cyanobacteria bloom has become an urgent task of water pollution control and ecological restoration. In recent years, the government vigorously promotes the policy of intelligent prediction of water environment, and many experts and scholars have made some research on the artificial neural network and other intelligent methods in water eutrophication assessment and algal bloom prediction, which have gained some results. However, the performance of artificial neural network is affected by sample training algorithm, parameter setting, model selection and other factors. With the complexity of the problem, the training time will increase with the increasing number of hidden layer nodes of a single network, which will cause the problem of the model training. Furthermore, it often leads to poor generalization ability and low accuracy of prediction of the network when training excessively or insufficiently. In order to obtain more accurate forecasting results, this paper proposes an efficient algal bloom forecasting method based on multiple BP network integration model after the investigation and analysis of the mechanism of the occurrence of cyanobacterial bloom. Firstly Bootstrap sampling technique is used to acquire different data samples, then training a number of different BP networks, and finally integrating multiple BP network by Bagging algorithm to establish algal bloom prediction model in Taihu River basin. The ensemble learning based on Bagging algorithm can fully excavate the information contained in the sample, and describe the relationship and change rules of the factors. The experimental results show that the multiple BP network integration model based on Bagging algorithm has a high prediction accuracy compared with the single BP network model, which provides a new idea for the prediction of algal bloom.

Keywords: algal bloom forecast, multiple BP network integration model, ensemble learning, Bagging algorithm

1. Introduction

The occurrence of Lake Eutrophication and cyanobacteria bloom has become one of the major environmental issues in the world at present [1]. Especially since the drinking water crisis caused by cyanobacteria bloom in Taihu Lake in 2007, cyanobacteria bloom has received much attention by different levels of management and community. As for Taihu Lake, the average content of total nitrogen and phosphorus nutrients of the whole lake stays at high levels in spring and summer, and the occurrence of cyanobacteria bloom has become the normal phenomenon. Furthermore, it is difficult to address the pollution problem basically in a short time. So it is particularly important to strengthen the prediction of algal bloom in the Taihu River Basin. In recent years, many scholars have carried out a series of exploration and research on water bloom prediction in Taihu basin, and continuously made new achievements, which deepened the people’s scientific understanding of cyanobacteria bloom and as a theoretical basis to support the prediction of cyanobacteria bloom. Fanxiang Kong et al [1], proposed the "four stage theory" on the formation of cyanobacteria bloom, namely, the process of dormancy, recovery, growth and aggregation, and formulated the concept of the prevention of cyanobacteria bloom in Taihu Lake. Aiming at the difficult situation of algal bloom prediction, Kexin Zhang [2], Shiping Zhu [3], Dagang Li [4], Zaiwen Liu [5] et al separately proposed the BP neural network, gray -BP neural network, the process neural network and the RBF neural network model of algal bloom prediction, which opened up new ideas for the prediction of water bloom. After more than ten years of research, artificial neural network and other intelligent methods have been widely applied to eutrophication assessment and algal bloom prediction, and have obtained achievement to some extent. However, the performance of the artificial neural network is affected by the sample training algorithm, parameter setting, and model selection and so on, so it needs a lot of experience of the designer to select the appropriate network. And when the training is excessive or insufficient, it will lead to poor generalization ability, this process has become one of the reasons why intelligent prediction model can play a role in particular rivers and lakes.

Motivated by the shortcomings of existing research, this paper proposes an efficient algal bloom forecasting method based on multiple BP network integration model, and it is applied to the prediction of water bloom in Taihu Lake Basin. The remainder of this paper is organized as following. Section 2 briefly presents the problem of single neural network, and proposes BP network integration model. Section 3 is the concrete implementation steps of a case study. The experimental results list of Section 4 and the paper conclusions of Section 5.

2. Network ensemble learning

2.1 The problem of single neural network

A lot of theories and practice results show that the following problems existed in the neural network prediction model based on the learning of the sample data: (1) It is difficult to determine the appropriate neural network model and algorithm, as well as setting parameters and time consuming; (2) With increasing complexity of the problem, the number of hidden nodes of a single network and the training time will greatly increase, which will result in difficulty of training; (3) Unlimited data is required to achieve the perfect effect, but in fact the training data is usually very limited. Single network is difficult to fully reflect and mining information, which is easy to cause the low training accuracy, what's more, excessive pursuit of training accuracy would lead to over fitting phenomenon; (4) The neural network is an unstable predictor: if the training data model change a little, the generalization ability of different models can be generated, which will lead to change obviously in the prediction of the unknown data. So the accuracy and reliability of the system cannot be guaranteed. However, the ensemble neural network can improve the performance of neural networks by mining the information on a number of individual networks. Detailed analysis will be discussed in the following.

2.2 Basic concepts of neural network ensemble learning

Ensemble learning is a new machine learning paradigm, which is applied to solve the same problem by using multiple (usually homogeneous) learning machines, and it can significantly improve the generalization ability of learning system. The basic idea is to use multiple models or solutions to solve the same problem. Due to the high precision and error distribution of different input space of the individual learner, it has achieved satisfactory results [7].

Let $\overline{h}(x)$ donote the ensemble output of the input x, it is described by the equation:

$$\overline{h}(x) = G[\alpha_m, h_m(x, y)](m = 1, 2, \cdots\cdots, M)$$

Where $h_m(x, y)$ is the m-th individual learner for the input and output pair (x, y), α_m is the weight for each individual learner $h_m(x, y)$ in the integration, and G is the method of individual combination, the most simple is the linear method. It is called simple integration if the weights are the same.

The concept of the Bagging (Bootstrap aggregating) algorithm was proposed by Breiman Leo in a technical report "Bagging Predictors" in 1994 [8].The main idea of the Bagging technique is to set up weak learning algorithms for training sets$(x1,y1),(x2,y2)\cdots(xn,yn)$ firstly. Then training *m (m<n)* sample each time from the training set, the sample will be returned to the training set when the training completed. So initial training examples can appear repeatedly or not in other training process.

After training, a forecast function sequence$h_1,h_2,\cdots h_t$can be obtained, and the final prediction result $\overline{h}$ is calculated by the weighted average method. The specific steps of Bagging algorithm as follows [9]:

(1) Given original data set $S=\{(x1,y1),(x2,y2)\cdots(xn,yn)\}$;

(2) Initialization of data sets;

(3) For t=1, ···, T Do;

(4) For each cycle t: take out M samples from the original data set S to form a new training set $D=\{(x1,y1),(x2,y2)\cdots(xm,ym)\}$;

(5) Obtain the learning model ht in the new training set *D* by using the basic learning algorithm;

(6) Save the learner model ht of t round, and integrate individual learning outcomes $h_1,h_2,\cdots h_t$ into a total learning result by using weighted average method.$\alpha_t(t=1,2,\cdots;T)$ is contribution weight for each individual, and it can be taken the same value.

Breiman [10] also pointed out that, to make Bagging effective, the basic learning algorithm must be unstable, that is, Bagging is sensitive to the training data. The learning algorithm of the basic classifier is more sensitive to the training data, the better the Bagging effect, so Bagging learning algorithm is quite effective for the network. In addition, due to the characteristics of the Bagging algorithm, it is very suitable for parallel training of multiple basic classifiers, which is a major advantage.

3. Algal bloom Prediction in Jinshu water source, Taihu Gonghu basin

3.1 data sources

LCD touch-screen is used as the operating platform, which is currently the most simple, convenient and natural way for the human-computer interaction. The serial communication technology is used to connect the GPS and YSI with the touch screen for real-time data transmission. SD card interface is used to achieve a large number of storages and rapid extraction of data. Communication between USB to RS-232 USB universal serial port and touch screen by using FTDI chip, to achieve the function of wireless remote transmission.

Sampling from 2009 to 2012, once per 4 hours in Jinshu water source, Taihu Gonghu bay were carried out. Determination and analysis of environmental factors including: hydrogen ion concentration index (PH), oxygen consumption (OC), temperature (T), turbidity (TB), ammonia nitrogen (NH4), total nitrogen (TN), total phosphorus (TP), dissolved oxygen (DO); biological factor analysis including: chlorophyll a(Chl_a), the density of algae (AD). All the data from 2009 to 2012 were processed to obtain the correlation between environmental factors and biological factors by using SPSS software, as shown in Table 1.

Table 1 correlation analysis of each factor of Jinshu water source, Taihu Gonghu basin, in 2009-2012

	pH	OC	T	TB	NH4	TN	TP	DO	Chl_a	AD
pH	1	0.125	0.627	0.315	0.362	0.555	0.276	0.054	0.240*	0.038
OC		1	0.035	0.112	0.023	0.094	0.065	0.124	0.153	0.186
T			1	0.321	0.486	0.709	0.392	0.480	0.176*	0.200
TB				1	0.256	0.462	0.279	0.045	0.281*	0.199
NH4					1	0.533	0.226	0.199	0.167	0.012
TN						1	0.341	0.364	0.459*	0.083
TP							1	0.102	0.167*	0.032
DO								1	0.039	0.294
Chl_a									1	0.335
AD										1

With the above 10 environmental factors and biological factors as the independent variables, the correlation coefficient was calculated, and select the independent variables on the 0.01 level (bilateral). Finally the significant correlate factors- pH, T, TN, TP were selected.

As TB, Chl_a and AD are the similar parameters of phytoplankton in the water, Chl_a is an important parameter for the eutrophication of water body recognized by experts and scholars, so removing TB and AD, and keeping Chl_a factor. A large number of experiments result show that the respiration of algae and decomposition of dead algae consume large amounts of oxygen when the excessive growth of algae reproduction, which will cause the DO in water dramatic changing, even may cause the water in severe hypoxia and the algal blooming, so DO is taken as one of the main factors affecting the algal bloom. Taking into account OC, NH4 are difficult to measure in the actual environment, and the correlation with the chlorophyll is not high, so remove these factors. In summary, 6 factors of the above 10 factors are selected (excluding OC, TB, NH4, AD) as reference factors of the algal bloom prediction model.

3.2 Algae bloom forecasting method based on BP network integration model

YSI6600 multi parameter water quality monitor imports of US is used as water quality sensor, which is suitable for multi-point sampling measurement, long-term field monitoring and profile analysis of different water bodies, and can monitor up to 17 parameters simultaneously. It also has the characteristics of small size and strong function. The water quality sensor was connected to the intelligent instrument through the serial interface of the RS-232 communication mode. The water quality sensor can detect pH, DO, TN, TP, SD, Chl_a, T,conductivity, Secchi disk depth(SD) and other indicators, and provide basic data onto the whole water quality analysis of the whole water system.

The artificial neural network is a new type of information processing system which can imitate and extends the function of human beings. It is a kind of adaptive nonlinear dynamic system and connected by a large number of simple processing units. Robert Hecht-Nielson had proved that the mapping with a 3 layer BP network can complete any n dimensional to m dimension. So the prediction model of algal bloom in a single BP network is divided into three layers, and the Bagging algorithm is integrated into the establishment of the BP network algal bloom prediction model.

Input layer: 18 neurons,the first three moments value of the pH, T, TN, TP, DO, Chl_a ;

Hidden layer: 3 neurons, the transfer function is tangent function;

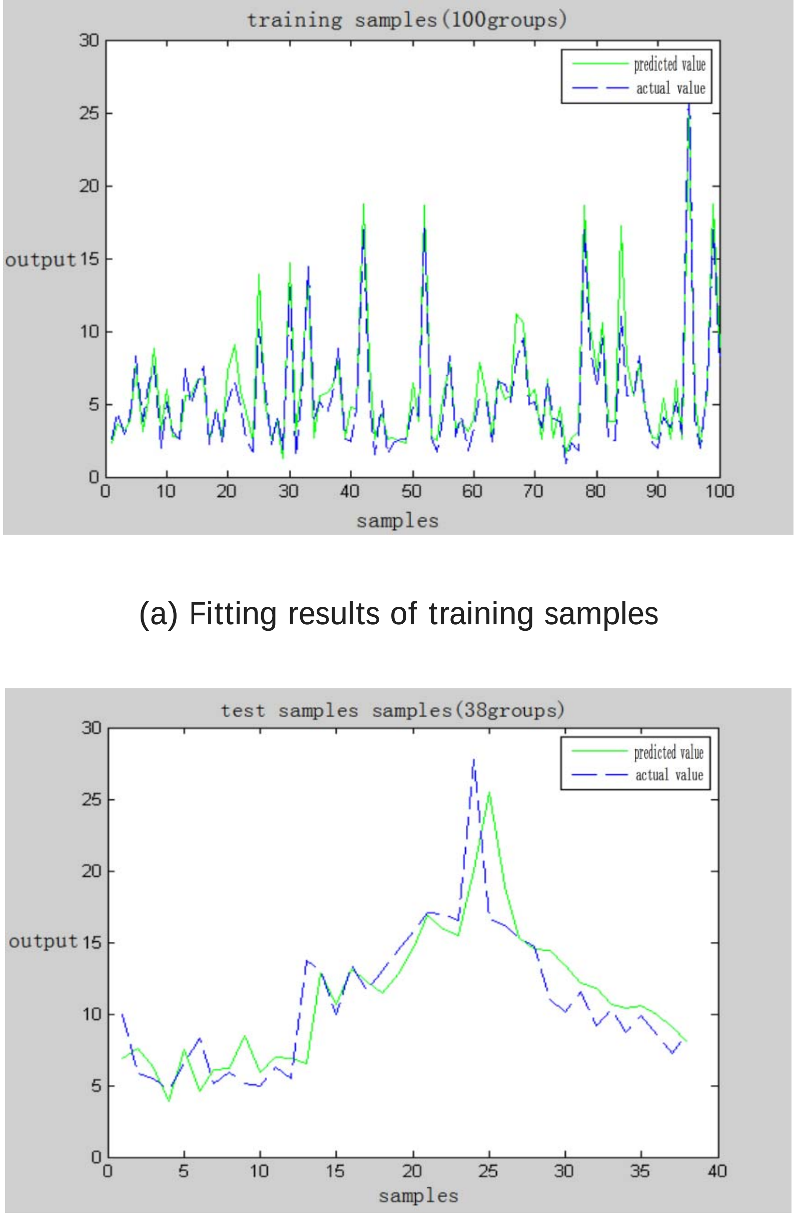
Output layer: 1 neuron, the chlorophyll value of the next moment, the transfer function is purelin.

4. Experimental and simulation results

There are 138 groups of sample data in Jinshu water source, Taihu Gonghu basin were selected from May to October, 2011, and the data of each site at different times of the same day were processed to obtain the average value. Then marking the data in time sequence. Between the first group and the 100th group were used as the training samples, the 101st group and the 138th group as the test samples. The gradient descent algorithm was used in the BP network, the training times and training error were setting 10000 times and 0.001 accruracy respectively.

4.1 Algae bloom forecasting method based on simple BP network integration model

10 BP network models were obtained after the same training samples training. Figure 1 shows the simulation results of the simple BP network integration model, the blue line and green solid line represent the measured value and the predicted value respectively. Table 2 shows the relative error of both the 10 Single BP network models and the simple BP network integration model by simple average method.



(a) Fitting results of training samples

(b) Fitting results of test samples

Fig.1 Algae bloom forecasting method based on simple BP network integration model

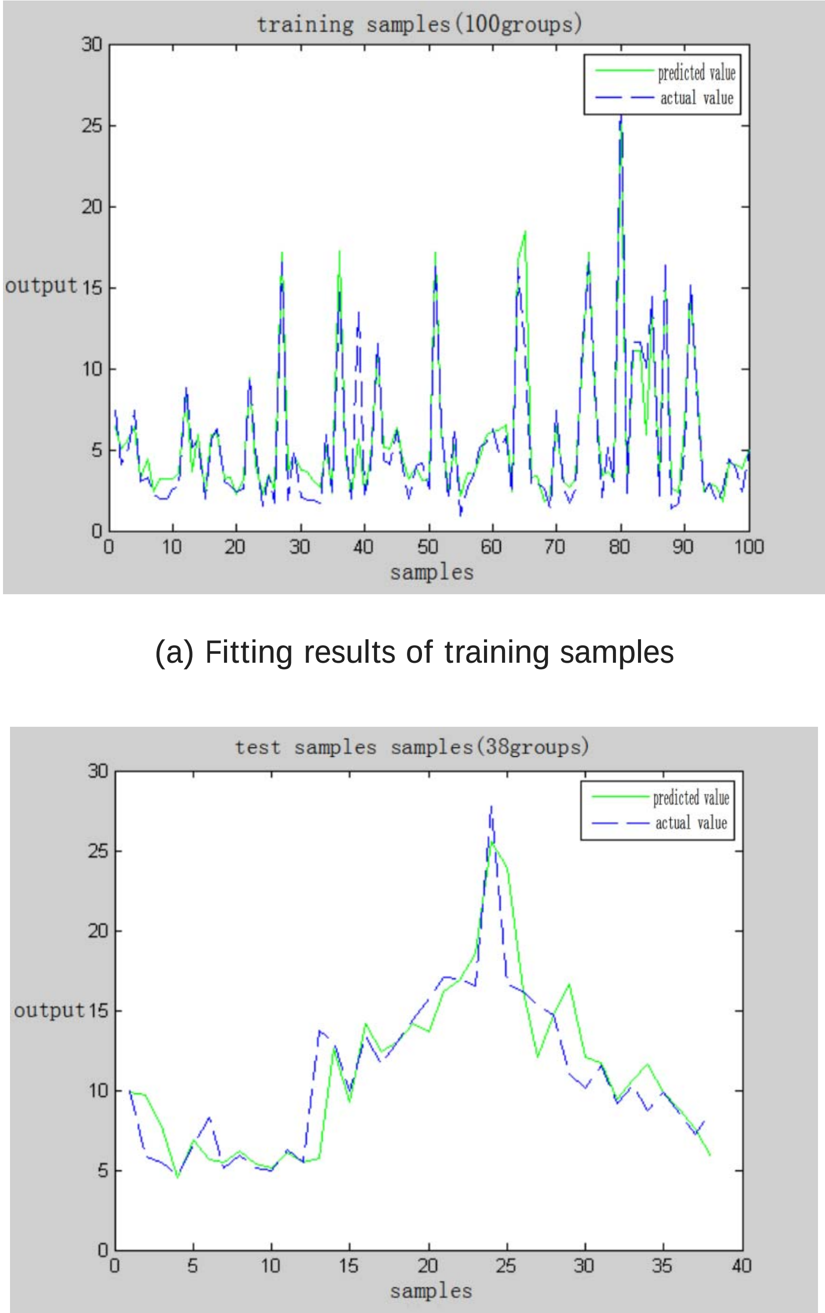
Table 2 Comparison of Single BP network with simple BP network integration model

Serial number	Single BP network model										Simple ensemble model
	1	2	3	4	5	6	7	8	9	10	
Relative error (%)	41.6	30.8	37.1	29.6	30.7	33.7	26.1	44.7	45.0	26.0	24.9

The table 2 shows the maximum relative error of a single BP network model is 44.7%, the minimum relative error is 26%, so the average relative error is 34.5%. While the relative error of the simple BP network integration model is 24.9%. Compared to the single BP network model, the model error dropped by 1.1%, and the average error of the 10 simulation model dropped by 9.4%. The Figure 2 can be seen that the predicted value of Chl_a s consistent with the trend of the actual value, which shows that the simple BP network integration model can basically reflect the growth trend of Chl_a.

4.2 Algae bloom forecasting method based on multiple BP network integration model by Bagging algorithm

Using the same training sample, 10 BP network models can be achieved after training, figure 2 is the simulation results by using the 10 BP network integration model by Bagging algorithm, the blue line represents the measured value, and the green solid line represents the predicted value. Table 3 shows the relative error of both the 10 Single BP network models and the 10 BP network integration model.



(a) Fitting results of training samples

(b) Fitting results of test samples

Fig.2 Algae bloom forecasting method based on multiple BP network integration model by Bagging algorithm

Table 3 Comparison of single BP network with multiple BP network integration model by Bagging algorithm

Serial number	Single BP network model										ensemble model
	1	2	3	4	5	6	7	8	9	10	
Relative error (%)	37.1	35.7	32.9	27.5	34.4	31.8	22.2	29.3	22.0	28.4	19.8

The table 3 indicates that the maximum relative error of a single BP network water bloom prediction model is 37.1%, the minimum relative error is 22%, so the average relative error is 29.1%. And the relative error of multiple BP network integration model by Bagging algorithm is 19.8%. The accuracy of the optimal value is 2.2% higher than that of a single BP network model, which is 9.7% higher than the average of the 10 simulation model. Compared with the simple BP network integration model without using the method of Bagging algorithm, the model is improved by 5.1%. According to the Figure 3, the value prediction of Chl_a cannot correspond with the actual value of each point, but their trend is consistent, which shows that multiple BP network integration model by Bagging algorithm can reflect the growth trend of Chl_a, and the prediction precision is in reasonable range.

5. Conclusion

By comparing these algal bloom prediction models based on the single BP network, simple BP network integration and multiple BP network integration model by Bagging algorithm respectively, the results show that the multiple BP network integration model based on Bagging algorithm is much better than the other two kinds of algal bloom prediction model. It shows that the Bagging algorithm can fully excavate the information contained in the sample, and get the main influence factors, which reflect the change rules of algal bloom. So multiple BP network integration model by Bagging algorithm can effectively improve the performance of the network, and obtain a higher accuracy of prediction results.

Acknowledgements

This work was financially supported by Innovation ability promotion project of Beijing municipal colleges and universities (PXM2014_014213_000033), Major Project of Beijing Municipal Education Commission science and technology development plans

(KZ201510011011), and General Project of Beijing Municipal Education Commission science and technology development plans (SQKM201610011009). Those supports are gratefully acknowledged.

References

[1] Kong Fanxiang, Ma Ronghua, Gao Junfeng. The theory and practice of prevention forecast and warning on cyanobacteria bloom in Lake Taihu. Lake Science, 2009, 21(3), p314-328.

[2] Zhang Kexin, Lu Kaihong, Zhu Jinyong. Predicting Model of Algal Blooms Based on BP Neural Network. 2012, 28(3), p53-57.

[3] Zhu Shiping, Liu Zaiwen, Wang Xiaoyi. Gray -BP neural network water bloom prediction methods. Intelligent Automation Conference: Nan Jing China, 2009, p1067-1072.

[4] Li Dagang, Wang Xiaoyi, Liu Zaiwen. Research on water-bloom prediction based on process neural network. Computer and Applied Chemistry, 2011, 28(2), p173-176.

[5] Liu Zaiwen, Cui Lifeng, Wang Xiaoyi, Lu Siying. The Method of Soft Sensing for Water Bloom in River and Lakes Based on RBF Neural Network. Chinese Control Conference, Hunan, China, 2007, p108-111.

[6] Shi Yan. Research on neural network ensemble model for logistics center site selection. Computer Engineering and Applications, 2009, 45(16), p211-214.

[7] Wang Lili. Research on Ensemble Learing Algorithm. Guang xi: Guang xi University. 2006.

[8] Jiao Licheng. Neural Network System Theory. Xidain University Press, China 1992, p34-41.

[9] Ma Ranran. Research of Ensemble Learning. Shan dong: Shan Dong University of Science and Technology. 2010.

[10] Breiman, L. Bagging Predictors. Machine Learning, 1996, 24(2), p123-140.