Prediction Analysis of Artificial Landscape Water Eutrophication

Jiaoyan Ai[1, a], Haiyang Xu[1, b], Yajuan Cai[1, c], Sizhi Wu[1, d] and Zengqiang Lei[1, e]

[1]Electrical Engineering College, Guangxi University, Nanning 530004, China.

[a]aijy@gxu.edu.cn, [b]554001527@qq.com, [c]374332945@qq.com;

[d]wusizhi1989@163.com, [e]leizengq@126.com

Keywords: Chl-a, Eutrophication, BP, Multiple Regression, Genetic Algorithms, SVM.

Abstract. The chlorophyll a (*Chl-a*)can express algae biomass, it is one indicator of the degree of eutrophication. In this paper, we take the Mirror Lake locate in Guangxi University as research object, and we fit the relationship between *Ch1-a* with every ecological factors in water by multivariate regression models,BP neural network, BP based onGenetic Algorithms and Support Vector Machines(SVM) to evaluate the eutrophication. Using BP neural network prediction model for the evaluation of eutrophication has good results which provide a theoretical basis for the landscape eutrophication's prevention and cure.

Introduction

Artificial landscape water,as the main water for leisure, entertainment and ecological environment improving, plays an important role in urban construction [1].Because of its relatively closed waters, relatively simple ecological system, and closely related to our life, it is prone to eutrophicationwhich leads to destruction of ecosystems, affecting the life quality of urban aesthetics and residents.

Research showsthat algae biomass can be represented by *Chl-a* contentthat is one of the most important indicators of eutrophication phenomenon[2]. The higher*Chl-a*concentration, the more biomass of the algae, it means thata serious eutrophication would happen. Mirror Lake,after a certain time observation and analysis, was taken as the research object in this paper.We determined 19 sampling points to monitor the temperature, illumination, PH, DO, COD, TN, TP, ammonia nitrogen and *Ch1-a* these 9 indicators,then we chose main instrument and test method to conduct a series of experimental work, at last we took a detailed analysis of the relation between the correlation of every indicators and the formation mechanism of eutrophication in the lake. The study shows that the temperature, light, TP, COD is alimiting factor for *Chl-a*which has an essential impact on the formation of eutrophication and DOhas a negative correlation to *Chl-a*, the others are not a strong impact in the lake and they are not very clear.

In order to quickly build and accurately predict *Chl-a* content model, we use multiple regression models,BP neural network,BP based onGenetic Algorithms and Support Vector Machine method to predict the content of *Chl-a*,to provide a scientific basisfor the prevention and treatment of landscape water through eutrophication evaluated by *Chl-a* content prediction.

Multiple regression model prediction

We put the temperature, light, PH, DO, COD, TN, TP, ammonia,*Chl-a*into the SPSS softwareby operating specifications,then make *Chl-a* as a dependent variableand the remaining factors as independent variables and filter eachindependentvariable by stepwiseregression. According to the independent variables influence on *Chl-a*,removing it from small to large until there are no variables in the equation can be removed. The best significant model is obtained:

$$Y=-54.558+0.694x_1+3.401x_2+143.283x_3 \quad (1)$$

Note: Y is *Chl-a* (mg / m^3), x_1 is COD (mg / L), x_2 is temperature ($℃$), $x_3$is TP (mg / L).

According to the variance of *Chl-a* multivariate regression model analysis, we can knowthat $P = 0.000<0.01$ (variance $P <0.05$, significantly), the model is pretty significantlyand credible [3]. Because the TP,the limiting factors of eutrophication and an important factor for *Chl-a*, can predict

more accurately to reflect the eutrophication status and be more in line with the ecosystem in actual situation. With this model, *Chl-a* content of 76 sample data will set to predict the relative error. The results are as follows in Fig.1.

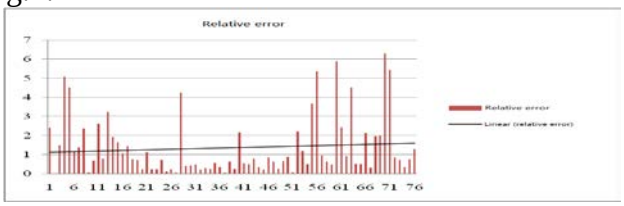

Fig.1 Relative error histogram

In the Fig.1, we can see that relative error is mostly between 0-1, but theothers are larger to the actual measurement, so the model can predict thetrend of *Chl-a*, but the effect is not precise enough.

Intelligent Prediction

BP neural networkmodel Prediction.BP neural network is to adjust the value of the network with the propagation learning algorithm. The 276 sets of data were interpolated by the 19 observation points in the Lake. The inputof the network,in the BP training and validation, were the temperature, light, PH, DO, COD, TP, TN and ammonia which were at the same time and place,we take the same *Chl-a*as the output terminalto predict *Chl-a*content.

The first 256 group of samples were as a network of training and validation set, the remaining 20 groups were as a test set. First,we input samples to normalize by using newff-function which is in MATLAB neural network toolbox to structure BP construction by repeated training 2,000 times in total to make the error reaches 10^{-4}, then we used the sim-function by simulation to output data for network anti-normalization after training the network. The results were compared with the target output to test the performance of the neural network. The predicted output is as follows in Fig.2.

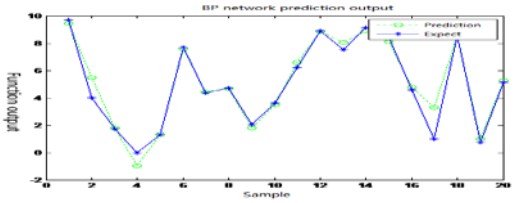

Fig.2 BP network prediction output

We rearranged by the 276 group sample, and then make the last 20 sets of data as a test set with BP network generalization ability. In order to get a output of the network, with the last 20 sets which have been trained for network simulation, we used RMSE to judge the merits of learning performance.The formula is:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N}(error_i)^2 / \mathrm{N}} = \sqrt{\sum_{i=1}^{N}(\mathrm{BP}output_i - output\_test_i)^2 / \mathrm{N}} \quad (2)$$

Note: where RMSE is RootMean Square Error, N is the Network test set samples, i is the i-th sample pairs, $error_i$is relative error of the i-th sample pairs, $\mathrm{BPoutput}_i$the i-th sample of the target output, $output\_test_i$ for the measured values of the i-th sample.

We get RMSE = 0.6844 by calculation,which means thatthe actual and the predicted values can be a good coincidence with 20 samples and the error is in the ideal range to indicate that this network already has a good generalization ability, it can predict the trends of *Chl-a* content in the lake for eutrophication rating evaluation.

Prediction of BP based onGenetic Algorithms. BP based onGenetic Algorithms uses Genetic Algorithms to optimize their initial weights and thresholds, so that the network can be optimized to predict the desired output function [4]. According to the theory of Genetic Algorithms and BP,eutrophication prediction of the lake has been implemented in MATLAB.Using the same 20 samples as a test for generalization ability of BP based onGenetic Algorithms and making RMSE evaluate the performance and learning ability, this output is shown in Fig.3.

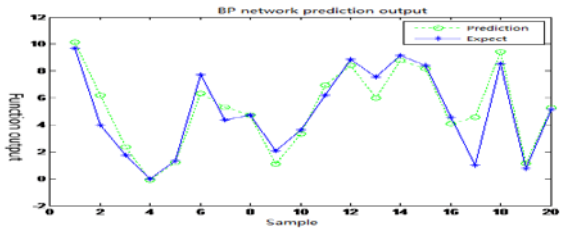

Fig.3output of BP based onGenetic Algorithms

Where RMSE = 1.1496, error is beyond the ideal range, so that the network can't be used to predict the trend of thelake *Chl-a* content and eutrophication of the lake rating evaluation.

Support Vector Machinemodel prediction. Support Vector Machine (SVM) method, as a new machine learning methods, mainly based on a limited sample data, in both the established model complexity and the machine's ability to find the best compromise, so as to achieve the best promotion performance [5]. When SVM is applied in regression analysis, the ε insensitive loss function is introduced into the SVM classification model, the researchers obtained Support Vector Regression (SVR).

In this paper,we used SVR to achieve the establishment and evaluation of the regression model in LIBSVM Toolbox and take the default parameters measured data to predict the same set of samples.We take first 256 asinput and the rest 20 groups of samples as a test to replace the kernel function for eutrophication predictionin MATLAB where respectively including three types of kernel functions: polynomial, RBF, Sigmoid.Through comparative analysis, using polynomial to reach the highest value, the output fitting result is shown in Fig.4:

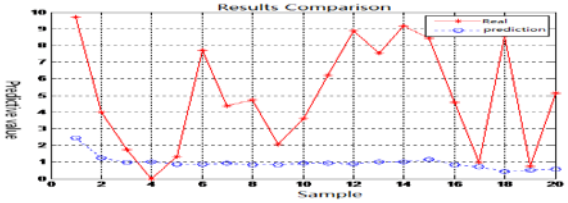

Fig.4 LIBSVM forecast renderings

Analysis, the error is too large, can't accurately predict *Chl-a* content, in order to improve a better accuracy and precision,we try tooptimize values of the main parameters of polynomial functions to affect learning performance optimization.

Using Cross-Validation (CV) optimization penalty function *C* and kernel parameter *g*, CV principle is in accordance with certain principles resulting sample data division, which means that there will be a part of sample as the training set and the rest as validation. Learning and training first, then the model performance verification, evaluating the performance of the classifier base on the accuracy obtained. Common CV methods are Hold-Out Method, Leave-One-Out Cross Validation (LOO-CV), K-fold Cross Validation (K-CV) [6-8]. Analysis of the characteristics of each method and CV can be realized, we use K-CV cross-validation method to optimize*C* and *g*, so in the absence of the test set label situation, finding the best parameter *C*, *g*, to avoid over-learning and less learning, and finally the training set to achieve the highest return accuracy. Using the forecast of cross validation optimization parameters in LIBSVM for measured data, The RMSE results obtained before and after are shown in Table 1:

Table 1predictions optimized parameters obtained

Kernel type	RMSE (Before)	RMSE (After)
Polynomial	5.0905	4.9977
RBF	5.2843	5.0570
Sigmoid	5.2138	5.2265

From Table1, theprediction accuracy of SVM parameter does not improve.Although the effect of selecting optimization parameters is slightly better,but fitting accuracy of the default parameteris in the same magnitude, indicating that using SVM for eutrophication prediction for the lake can't achieve satisfactory results.

Conclusions

Using Multiple regression model to predict *Chl-a* content of 76 sample data, we can know from the relative error bar chart(Fig.1) that most of the relative errors are between 0 and 1, but some prediction error is bigger than 1 and quite different from the actual measurements, so the model can predict trends of*Chl-a* content, but the overall effect is not precise enough.

The results of three intelligent prediction methodsare as shown in Table 2:

Table 2 three prediction method for predicting the results of comparison

Prediction	RMSE	R(All)
BP neural network	0.6844	0.9716
Genetic Algorithms optimization BP network	1.1496	0.9838
LIBSVM (optimization parameters)	4.9977	0.9801

According to Table2, from the comparison of the RMSE, the RMSE of BP predictionare minimum in these three methods, instead LIBSVM got maximum RMSE.In three methods, the total values of R is close to 1,indicating that all the network have goodperformance, but the R value optimized by BP based onGenetic Algorithms is closer to 1, so it has a better resultfrom the comparison of the simulation time, the Genetic Algorithms takes more time than LIBSVM and BP network.

Taking all these factors that the reasons for this prediction are:

Mirror Lake is semi-open and semi-enclosed artificial landscape water, it is divided into 19 sampling areas anddifferent areais affected by the severity of differentimpact factor.Most of research in natural lakes and reservoirs take one or two to get results, a lake including comprehensive 19 sampling points is more complex. Thus, ecological factors of network parameters of the lake is not sensitive and robustenough,making BP network, BP based onGenetic Algorithms and LIBSVMnetwork predictions are not very accurate.

BP network and BP based onGenetic Algorithms' RMSE are significantly better than LIBSVM, probably because SVM kernel functions inside the model are not suitablefor Mirror Lake eutrophication of water bodies. In the Genetic Algorithms optimization BP's R value is a slightly better situation, BP network simulation hasa shorter time and slightly smaller RMSE, so we will choose BP neural network prediction for *Chl-a*in an ideal tolerance scope,we will get a better and practicalresult whenwe evaluate the situation.

Traditional statistical multivariate regression model and BP neural network are compared by the relative error histogram in Fig.1, and we will know that the results of some relative error in multiple regression model is greater than 1, the prediction error is too large,instead, there is no obvious effect in BP network forecast.The reason for this phenomenon is likely due to the complexity of eutrophication reflected in the relationship between ecological factors of non-linear, linear regression do not predict this relationship well.

Therefore,in view of Mirror Lake,using BP neural network to predict *Chl-a*can be more effective to evaluate eutrophication, thereby to combat eutrophication of landscape water to provide a more accurate theoretical scientific basis.

Acknowledgement

This research was financially supported by Science and Technology Plan Chairman Foundation of Guangxi (1517-08). The corresponding author of this paper is Ai Jiaoyan, a professor at Guangxi University.

References

[1] Su Weizhong theoretical analysis and spatial organization of urban open space research [D] Kaifeng: Henan University: 37, 2002.

[2] Meng rain. Ecological effects of shallow water crab grass type lakes farming [D]. SoochowUniversity,2013.

[3] GuzhenRui, Sun Liping, bell away, and so on Tianjin Water Park in Lake chlorophyll a and environmental factors in the principal component linear regression analysis [J] Ecological Sciences, 2015, 34 (4): 125-130.

[4] TianXuguang, Song Tong, Liu new combination of structure and parameters of Genetic Algorithms to optimize BP neural network [J] Computer Applications and Software, 2004, 21 (6): 69-71.

[5] Zhang work on statistical learning theory and support vector machine [J] AutomaticaSinica, 2000, 26 (1): 32-42.

[6] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods[J]. Advances in large margin classifiers, 1999, 10(3): 61-74.

[7] Kearns M, Ron D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation[J]. Neural Computation, 1999, 11(6): 1427-1453.

[8] Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection[J].Icml, 1995, 14:1137--1143.