Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 2 (2010) 970–979

Procedia Environmental Sciences

International Society for Environmental Information Sciences 2010 Annual Conference (ISEIS)

Forecasting and Evaluating Water Quality of Chao Lake based on an Improved Decision Tree Method

Hao Liao[a], Wen Sun[a, b, *]

[a]*School of Software Engineering, University of Science and Technology of China, Hefei, Anhui,230027, China*
[b]*Department of Computer Science and Technology, Hefei Normal University, Hefei,Anhui, 230001, China*

Abstract

As the water quality of Chao Lake becomes diverse and eutrophied, predicting and evaluating water quality become more and more important. But prediction and evaluation are complex problems. In this paper, we come up with an improved decision tree learning method making water quality prediction easier and forecast more accurate. The classification standards are based on the evaluation mechanisms provided by the Hong Kong Environment Protection Department. We released an online web forecast system to apply to the classification and prediction of Chao Lake. Experimental results show that the improved method is better than artificial neural network or genetic algorithm with higher recognition rate and forecast accuracy and strong practical value.

Keywords: Improved decision tree learning model; Classification; Water quality

1. Introduction

Traditional methods of predicting and evaluating water quality, such as Mathematical statistics, the gray system theory, neural network modeling and water quality simulation model method are based on chaos theory. These four categories of methods have several shortcomings: the existence of mathematical statistics calculations, equal treatment to the old data and new data, and difficult prediction due to multivariate overlapped.

It is a very important matter to predict water quality which can enhance economic efficiency as a result. However, water quality prediction becomes a very complicated issue due to the complexity and diversity. In turn, the data will also affect water quality and authenticity of the limited accuracy of the forecasts. Meanwhile, building the models upon an accurate prediction in the limited monitoring data is difficult. Some factors have mutual influence on other factors, all of them having effects on prediction consequently, while some factors are independent, so it is difficult to use any formula to describe this relationship. The traditional water quality data which is not processed is very difficult to handle. In this article, we apply an improved model of decision tree learning (Improved decision tree learning model. IDTL), considering the impact of water between the elements and some crude factors, into water quality prediction.

* Corresponding author. Tel.: +86-551-3500590.
E-mail address: sunwen@mail.ustc.edu.cn

1878-0296 © 2010 Published by Elsevier Open access under CC BY-NC-ND license.
doi:10.1016/j.proenv.2010.10.109

The data we study is based on monitoring the water quality of Chao Lake, which is the largest lake in Anhui Province, one of the largest freshwater lakes in China. With the continuous economic development in recent year, water quality of Chao Lake is deteriorating due to sewage, fertilizer and pesticide widely used around the wetland area to compensate the lost caused by soil erosion and decreasing lake sedimentation rate leading to an obvious lake eutrophication.

In recent years, more and more achievements of artificial intelligence are applied into environmental protection. A popular method named neural network model has a drawback that is usually difficult to clearly understand the data model of the association between the variables.

The purpose of this paper is to propose a modified decision tree method taking better and more accurate information into account to predict and evaluate the water quality. This model is mainly a combination of some feathers of neural network and main method of decision tree to clarify of the data of water quality and forecasting.

The rest of this paper is organized as follows: section two introduces the background and strategies of this research; it introduces neural network and decision tree method and discusses how an Improved Decision Tree Learning model (IDTL) is used in the prediction. Section three focuses on application problem domain and study area. Section four presents experimental results and performance evaluation of the model. The final section summarizes the results of the research and concludes some directions for future work.

2. Background and Model

2.1. Artificial neural network in water quality prediction

Artificial neural network (ANN) was first introduced in 1943 (McCulloch and Pitts,1943),research into applications of ANNs has blossomed since the introduction of the back propagation training algorithm for feed forward ANNs in 1986 (Rumelhart et al.,1986a). Using conventional methods can be time-consuming and expensive in terms of labor and computational resources since the inter-relationship among parameters are complicated. On the other hand, compared with the conventional methods, the ANN approach has been shown to generate more accurate and repeatable results. (Mohaghegh et al.,1995). Some recent efforts at applying ANN methods in water prediction include the following. (Maier et al,2000)conducted a study reviewing 43 research papers that employed neural networks in the prediction and forecasting of water resources variables. They observed that neural network models always work well and their use in the study of water is on the increase due to their ability to handle large amounts of non-linear, on-parametric data. The use of data-driven techniques for modeling the quality of both freshwater (Chen and Mynett, 2003) and seawater (Lee et al., 2000, 2003) has met with success in the past decade. Reckhow (1999) studied Bayesian probability network models for guiding decision making regarding water quality in the Neuse River in North Carolina (Chau. 2006) has reviewed the development and current progress of the integration of artificial intelligence (AI) into water quality modeling.

2.2. Decision tree learning for water prediction

Decision tree is able to generate a graph or a decision model which is going to result in a better generation tool. These include the ability to forecast on the possibilities, resource costs, and effectiveness. In the field of data mining, decision tree is a predictive model, a tree structure similar to the flowchart, which can help people to obtain a target value through the classification and analysis.(JiaWei and Macheline, 2001).

Decision tree has an advantage in water quality classification. For example, the potential water quality among times can be easily identified by the set of terminal nodes in the tree that has better water quality data classification, and then the user can focus on the specific data described by those nodes. Comparing to other methods, a decision tree can be constructed relatively easily and quickly.

Some researches use decision tree learning for water prediction are discussed as follows: Qiuwen Chen et al.(2004) applied decision trees method into models construction to qualitatively predict Phaeocystis globosa blooms in the Dutch coastal waters. Jinsuo Lu et al.(2009) used decision tree model to predict the level of chlorophyll in next day from Online Monitoring Station. Hendrik Blockeel et al.(1999) introduced two applications of decision tree learning in simultaneous prediction of multiple physic-chemical properties and past physic-chemical properties of

the river water from biological properties. The research mainly emphasizes on decision tree learning in the existing papers used to predict water parameters without considering the interrelationship of attributes which is an adequately expressive representative if the data sets are supposed to show attributes independence. But in water quality, the data sets are not independent. Thus, we adopt a method combining consideration of interrelationship and prediction model. And we can significantly improve the accuracy on water quality prediction in our study area – Chao Lake.

2.3. Improved Decision Tree Model

Forecasting Model Based on Improved Decision Tree (IDTL) combines artificial neural networks and decision tree algorithm. The main advantage of this approach lies in the clustering of data when processing the data with stress on inter-dependence between parameters and an aim at reducing rough disadvantages at the same time. Therefore, more accurate predictions can be made by using this method. The results of this method are mainly aim to handle multiple variables of dataset, the following Figure 1 gives the forecasting process diagram. Forecasting model, based on improved decision tree algorithm, puts the original data set into the neural network training, and this step mainly consists of the input good decision tree algorithm; and what the final decision tree rules generate is the result of this process. Arrow shows the direction of data flow from input to output, and arrows are marked with the numeric types of data. Our goal is to precast the data of N+1 month based on that of the known N month by collecting the data of successive months before the N+1 month.

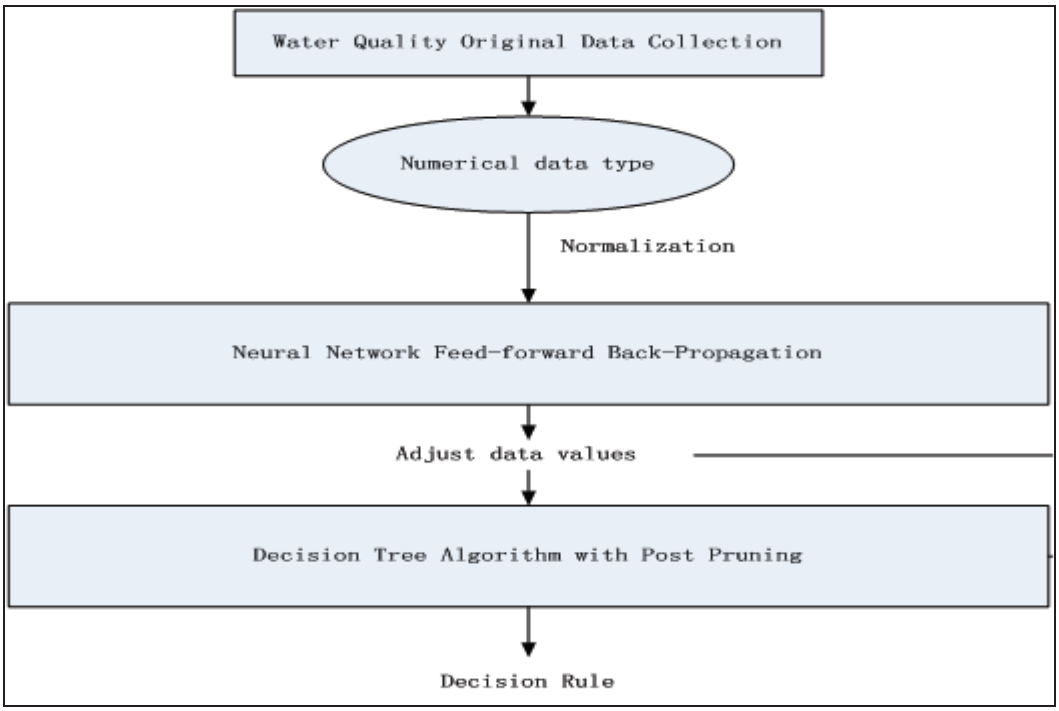

Fig. 1. Design of the IDTL Model

3. Application problem domain

3.1. Introduction of water quality prediction

The objective is to use the IDTL model for identifying the relationships among the core variables responsible for water quality prediction. The performance of the IDTL model is assessed by using data sets obtained from Chao Lake in Anhui Province, China.

3.2. Study Area

The Chao Lake is one of the biggest fresh lakes of China. It is also a very important fresh water reservoir to several provinces around the lake. The quality of water resources in this area are particularly vulnerable because of nutrients and pollution discharged from factories. The water quality prediction project established by Chinese Academy of Science helps people and government to develop best management practices to minimize adverse effects.

According to the average water level in years, Chao Lake area is calculated to 760km2, an average depth of 3 m, maximum depth of 6.78 m [SHAN Ping]. The following figure shows Chao Lake Basin (local) and distribution of monitoring point. Weekly or biweekly samples of DO, BOD5, PH, temperature, Ammonia-Nitrogen and other selected water quality variables were collected at 12 stations showing in the figure 2.

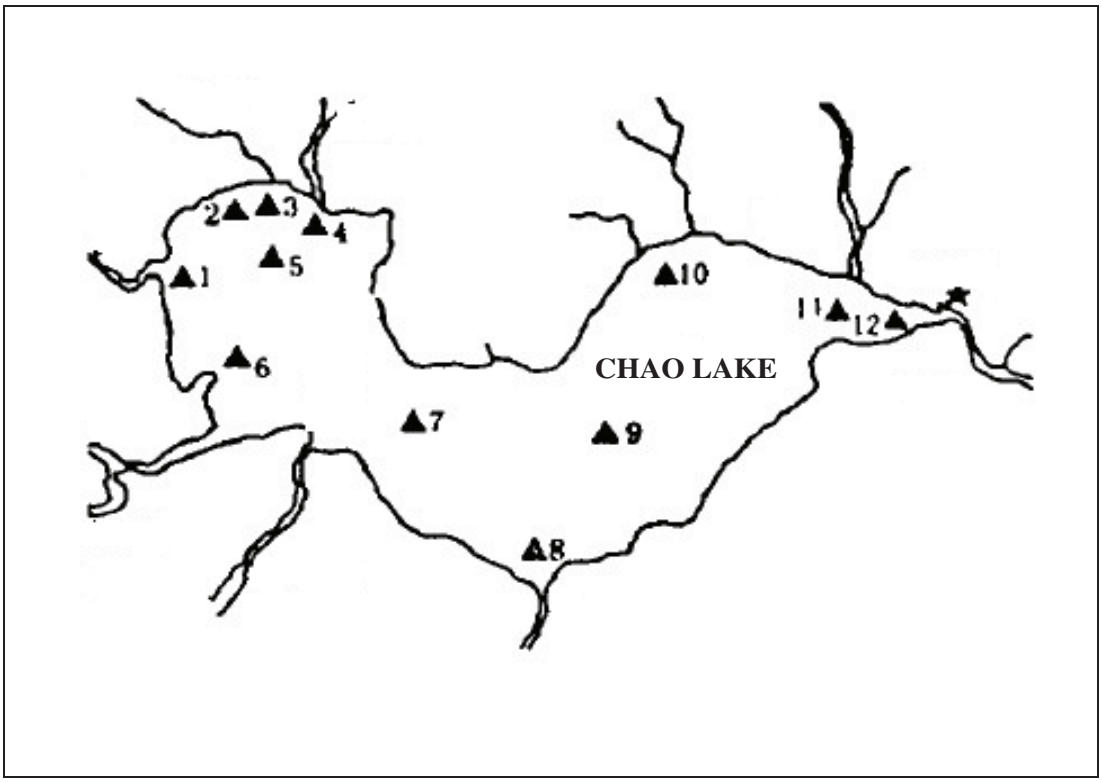

Fig. 2. Chao Lake Basin(local) and distribution of monitoring points

4. Result and Discussion

4.1. Evaluation of water quality

From all kinds of the monitoring indicators, referring to the evaluation criteria of the water quality by the Hong Kong Environmental Protection Department, we select three most influential parameters to evaluate water quality in details: dissolved oxygen (DO), five-day biochemical oxygen demand(BOD5) and ammonia-nitrogen levels. Having a score named water quality index (WQI) under each parameter in the table 1, the smaller number WQI shows, the better water quality. Then WQI is classified into five levels which are water quality index I, II, III, IV IIV corresponding to excellent water quality, good, general, bad, very bad, which are listed from low to high:

Class I: [3.0 –4.5], Class II: [4.6—7.5], Class III: [7.6 – 10.5], Class IV: [10.6 – 13.5], Class V: [13.6 – 15.0]

For example, a stream with oxygen saturation equal to or greater than 91%, the five-day biochemical oxygen demand below 3 Mg/L, and the ammonia-nitrogen below 0.5 Mg/L, is 3 for its total WQI, so the stream is classified to Class I. In other situation, a river may get a poor score of 15 points, so WQI is Class IIV. The water quality of Chao Lake monitoring stations for each index was presented on a monthly and annual figure. Monthly figures are based on the weekly or biweekly collection of water samples for test score containing all the three parameters and the annual water quality index is the annual average of the monthly water quality index.

Table 1. Calculation of Water Quality Index (WQI)

Score	DO(Saturation/%)	BOD5(Mg/L)	Ammonia-Nitrogen(Mg/L)
1	91 -110	< 3	< 0.5
2	71 – 90/111-120	3.1-6.0	0.5-1.0
3	51–70/121 - 130	6.1 – 9.0	1.1 – 2.0
4	31- 50	9.1 – 15.0	2.1 – 5.0
5	<30 or >130	>15.0	>5.0

Table 2 shows the classification of sample data collected in the estuary of Nafei River. As seen from the graph below, most of the data fall into class III and II. This is because in reality, water qualities level of most river is general and only at higher level for few times.

Table 2. Sample data of Nafei River of Chao Lakeq

Date	DO	BOD5	Am-Ni	Score	Level
2008-1-7	6.58	4.18	1.270	8	Class III
2008-2-14	9.83	2.50	0.590	4	Class I
2008-3-4	10.46	3.40	3.700	7	Class II
2008-4-2	10.05	5.70	1.850	6	Class II
2008-5-2	5.50	6.64	1.200	9	Class III
2008-6-5	6.94	6.50	1.070	8	Class III
2008-7-3	7.37	5.60	0.710	6	Class II
2008-8-4	7.20	4.23	3.340	8	Class III
2008-9-2	7.35	6.00	0.170	5	Class II
2008-10-6	7.81	5.55	1.110	7	Class II
2008-11-3	9.13	4.10	0.634	5	Class II
2008-12-2	11.34	6.40	0.985	7	Class II

4.2. Classification and prediction results

From the result trained by the neural network part in figure 3, we obtained the link weights between the input layer and the hidden layer. Then, the original water quality data sets are trained by using equations shown below into a new dataset. IDTL model was applied to the new data set to generate decision rules.

$$O1 = 3.6DO + 3.57BOD5 + (-2.51)\,Am\text{-}Ni + (-0.41) \quad (1)$$

$$O2 = 1.3DO + 1.25BOD5 + (-2.09)\,Am\text{-}Ni + (-0.74) \quad (2)$$

$$O3 = 1.8DO + 1.21BOD5 + (-2.15)\,Am\text{-}Ni + (0.87) \quad (3)$$

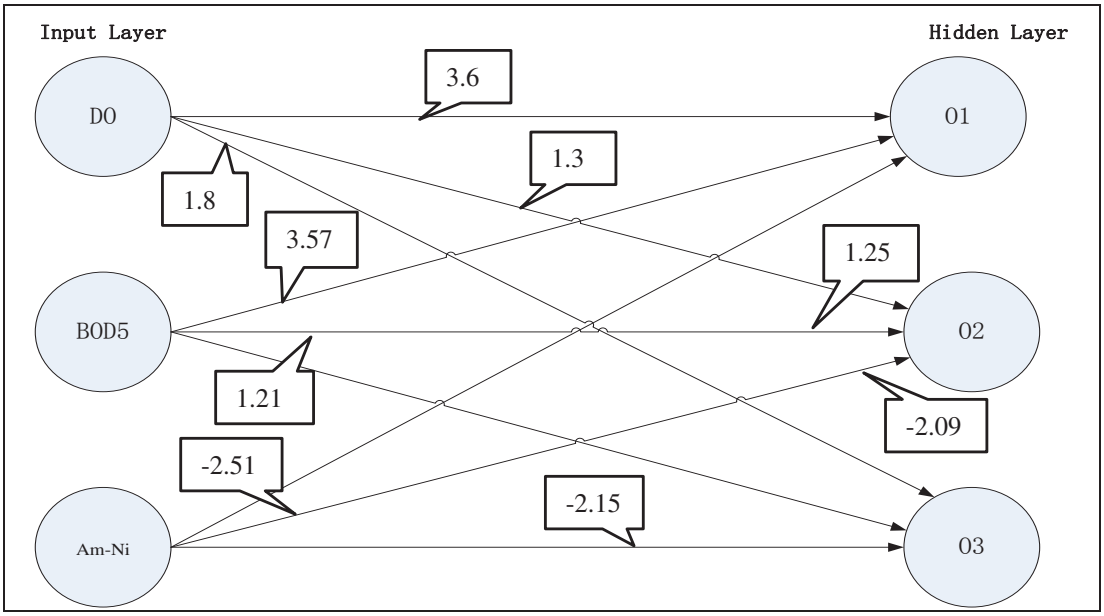

Fig. 3. Neural training result of the original numeric dataset

As shown in Table 3, it can be seen that when compared the results generated by the C4.5 decision tree learning algorithm with the IDTL model, the IDTL model reduces the tree size and number of rules by half with only 3% decrease in classification accuracy. A decrease in the number of rules is an improvement because domain experts such as environmental engineers can more easily validate them. Hence, the IDTL model applied in this study provides better explanation capability since it generates a comparative smaller rule set with an acceptable level of classification accuracy. In other words, with the sacrifice of a little classification accuracy, the IDTL model is able to provide some explicit heuristics for classification that enhance predicting water quality level from water monitor.

Table 3. Comparison of IDTL Model and C4.5 results

Measures	C4.5 with pruning	IDT with pruning
Tree size(nodes)	84	41
Number of rules(leaves)	32	16
Test set size	228	228
Correct classification rate(%)	80.21	78.75
Misclassification rate(%)	19.79	11.25
Mean absolute error	0.12	0.18
Mean squared error	.24	14.10

Some sample rules generated with the new data set by using the IDTL model are shown in Figure 4, and Figure 5 shows the sample tree for the IDTL models.

```
IF O1 <= 40.1068 THEN
    IF O2 < = 17.278 THEN
        IF O3 < = 15.6944 THEN
            WATER QUALITY = Class III[7.6-10.5]
```

Fig.4.Sample rule generated by IDTL model

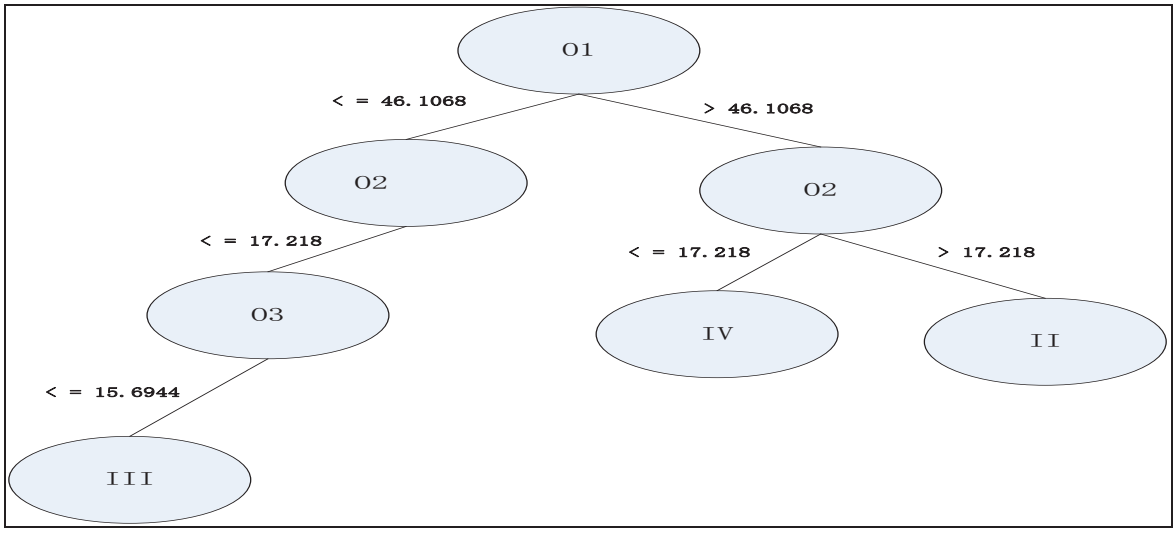

Fig. 5. Sample tree generated by IDTL Model

To build this model and predict the water quality of Chao Lake, See5 software was used to calculate and give the result of prediction. We are directly predicting the score of water quality by using the three key training parameters—O1, O2 and O3. Meanwhile, there is a relationship with the score and water quality level. So we can get the water quality level directly.

In summary, the prediction correct rate of IDTL model is 85% as the experiment shows by using See5 soft package which is better than decision tree model having the prediction correct rate at 70%. We are also delivering an online water quality prediction system as the Figure 6 presents below：

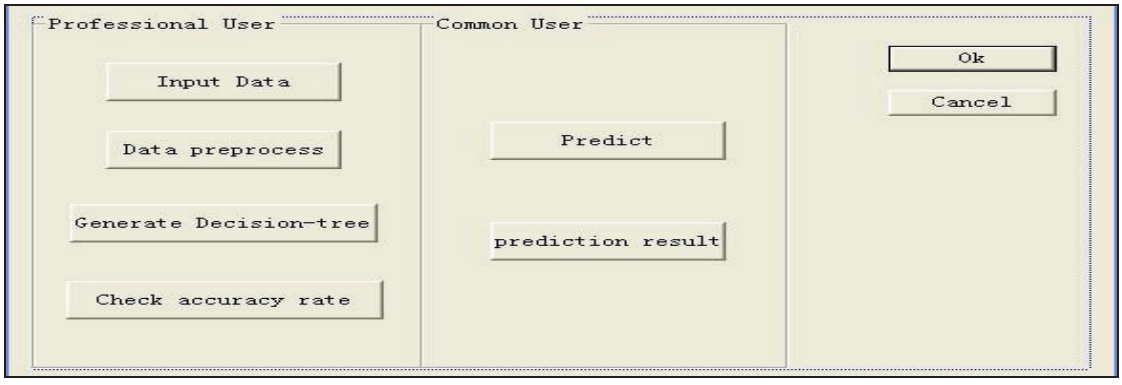

Fig. 6. Demo water quality prediction system

5. Conclusion and Future Work

Our research is to evaluate and predict the water quality of Chao Lake by using IDTL model. Study indicated that, compared with ordinary decision tree method like C4.5, the IDTL model has many advantages.

In addition, we also found that IDTL model has several advantages compared to ANN model: (1) For policy maker, IDTL model is easy to understand generated tree structure. (2) IDTL model training process is faster than ANN model, and IDTL is always convergent. (3) Knowledge of IDTL model can help us choose parameters and assess the dependencies between related attributes.

We noted that although the overall prediction performance of IDTL model is good, the predication error rate is also high. Causes of the high error rate including: (1) Limited availability of data sets; (2)Indistinct difference between the data sets.

In the future, there are several parts can be further studied. The issues including: (1) To define a formal process of integrating attributes grouping into the construction of a multivariate decision tree for categorical data modeling. Multivariate decision tree model improving the link weights can be used to prune errors. (2) To integrate other water quality parameters into IDTL model is a good way to improve prediction accuracy, which is necessary to collaborate with water evaluation experts who can help classify water parameters into different information groups and develop models for each group.

Acknowledgement.

We would like to thank the anonymous reviewers for their valuable comments. Financial support from Chinese Academic Science Special Grant for Postgraduate Research, Innovation and Practice, and School of Software Engineering of USTC are gratefully acknowledged. Water quality data of Chao Lake support from Environment Protection Bureau of Anhui Province and Evaluation standard reference from Environment Protection Department of Hong Kong are highly appreciated. We would like to thank Prof. Ruijian Zhang from Purdue University for his contribution to this work.

References

[1] Agrawal R, Shafer J C. Parallel mining of association rules. In: Proceedings of IEEE Transactions on Knowledge and Data Engineering. USA: IEEE Educational Activities Department;1996, 8(6): 962- 969.

[2] Blockeel H, Dzeroski S, Grbovic J. Simultaneous prediction of multiple chemical parameters of river water quality with TILDE. In: Proceedings of Third European Conference on Principles of Data Mining and Knowledge Discovery. Berlin: Springer;1999, 15-18

[3] Breiman L, Friedman J. H., Olshen R. A., and Stone C. J. Classification and regression trees. California: CJ Stone - Wadsworth Inc; 1984.

[4] Chen Q, Mynett A E. Predicting phaeocystis globosa bloom in dutch coastal waters by decision trees and nonlinear pecewise regression. *Ecological Modeling* 2004; **176**: 277-290.

[5] JiaWei Han, Micheline Kamber. *Data mining: concepts and techniques.* 2rd ed. CA: Morgan Kaufmann Publishers; 2001.

[6] J R Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies* 1987; **27(3)**: 221 – 234.

[7] J R Quinlan. *C5: Programs for machine learning*. CA: Morgan Kaufmann Publisers; 1993.

[8] K Lee, S D Choi, G H Park., R Wanninkhof, T-H Peng, R M Key, C L Sabine, R A Feely, J L Bullister, F J Millero, Alex Kozyr. Updated anthropogenic CO2 in the Atlantic Ocean. *Global Biogeochemical Cycles* 2003; **17(4)**: 1116-1122.

[9] Kwok-wing, Chau. A review on integration of artificial intelligence into water quality modeling. *Marine Pollution Bulletin* 2006; **7**: 726-733.

[10] Lee J H W, Huang Y, Dickmen M, Jayawardena A W. Neural network modeling of coastal algal blooms. *Ecological Modeling* 2003; **159**: 179–201.

[11] Lee J H W, Wong K T M, Huang Y, Jayawardena A W. A real time early warning and modeling system for red tides in Hong Kong. In: Proceedings of the Eighth International Symposium on Stochastic Hydraulics. Beijing: Balkema;2000, p.659–669.

[12] McCulloch W, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 1943; **7**:115 – 133.

[13] Mehta M, Agrawal R, Rissanen J. SLIQ: A fast scalable classifier for data mining. In: Proceedings of 5th International Extending Database Technology Conference. Berlin:Springer;1996, p.18-32

[14] Maier H R, Dandy G C. Neural networks for the prediction and forecasting water resources variables: a review of modeling issues and applications. *Environmental Modeling & Software* 2000; **15**: 101-124.

[15] Reckhow K H. Water quality prediction and probability network models. *Canadian Journal of Fisheries and Aquatic Sciences* 1999; **56**: 1150-1158.

[16] Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation. *Parallel Distributed Processing* 1986; **1**:318-362.

[17] SHAN Ping, Yin Fu-cai. Backward assessment and instrument approach for the Chao Lake pollution prevention and control. *Journal of Anhui Normal University (Natural Science)* 2003;**3**: 289-293.

[18] Shahab Mohaghegh, Ameri S, Hefner M H. A methodological approach for reservoir heterogeneity characterization using artificial neural networks. In: Proceedings of 1994 SPE Annual Technical Conference and Exhibition. Louisiana: Society of Petroleum Engineers; 1994, p. 337-346.

[19] W Lam, F Bacchus. Learning bayesian belief networks: an approach based on the MDL principle. *Computational Intelligence* 1994; **10**:269-293.