CHEMICAL ENGINEERING TRANSACTIONS
VOL. 71, 2018
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.aidic.it/cet
Guest Editors: Xiantang Zhang, Songrong Qian, Jianmin Xu

ISBN 978-88-95608-68-6; **ISSN** 2283-9216
DOI: 10.3303/CET1871207

Application of Improved Support Vector Regression Method and Chemical Index in Classification and Evaluation of Water Resources Quality in China

Yu Ye

School of Landscape Architecture, Beijing Forestry University, Beijing 100083, China
yeyulab@163.com

With the rapid development of China's economy, the rapid increase in the number of chemical enterprises, which poses a serious threat to the security of surrounding water environment. Therefore, in the form of China's economy into the new normal, the connotation and extension of risk have changed, which makes the water environment risk become more and more complicated. Lake water resources play an important role in the sustainable development of national economy and society. It is an important way to implement the optimal control strategy for the sustainable development of lake water resources in china. However, there are serious chemical pollution problems in the exploitation of lake water resources in our country, such as pH, dissolved oxygen, permanganate and ammonia exceed the standard value, which cause a serious impact on people's production and life. Therefore, it is an important link to evaluate the water quality of Lake. The traditional model is based on the mass and energy of matter and the principle of momentum conservation. Although the physical concept is clear, it is limited to the amount of calculation and parameter identification. In practice, the model structure is simplified and assumed, which affects the accuracy and practicability of the model. For this reason, this paper establishes an input response relationship between water quality index and the corresponding pollution source, and then introduces the improved support vector regression method to classify and evaluate the water quality of Taihu Lake. The experimental results show that the method has the advantages of high precision, strong practicability, simple calculation and so on. It can give a reasonable water quality classification for the lake water samples, which is worthy of popularization and application.

1. Introduction

With the rapid development of China's economy, the rapid increase in the number of chemical enterprises, which poses a serious threat to the security of surrounding water environment. Therefore, in the form of China's economy into the new normal, the connotation and extension of risk have changed, which makes the water environment risk become more and more complicated. The accurate assessment of water quality is based on the detection of abnormal water quality, and the early-warning of water quality is an effective application of water quality assessment. For the content and the relationship between the three parts, domestic and foreign scholars have made a lot of research, and we only illustrate some of their important contributions. The research work on water quality anomaly detection began in the last century in 90s. Chen et al., (1997) set up a water quality early warning system based on the S-P model, and the water quality monitoring system is applied to the Guijiang. Dong et al., (2002) propose a comprehensive early warning technology, which is based on environmental information system and geographic information system. Based on the river water quality model and the GIS theory, the water quality early warning system of Hanjiang is established by Xiao (2005). By using the method of comprehensive judgment and analytic hierarchy process (AHP), the water environment early warning of Tianjin city is studied by Hu (2006). Hou (2013) proposes a method of water quality anomaly detection, which is based on radial basis function neural network and wavelet analysis. Taking the mainstream of Songhua River as an example, in view of the sudden water pollution accident, Liu (2011) uses the fault tree analysis method to determine the accidents possibility along the river, and uses it as a major indicator of risk identification. According to the typical organic pollutants in Songhua

Please cite this article as: Ye Y., 2018, Application of improved support vector regression method and chemical index in classification and evaluation of water resources quality in china, Chemical Engineering Transactions, 71, 1237-1242 DOI:10.3303/CET1871207

River, Wang (2013) and Li (2012) develop a set of WAF pollutant migration and transformation model based on WASP, which is suitable for the characteristics of Songhua River.

2. Intrusion detection system

The main research object of water pollution chemistry is pollution which is discharged into the water environment by human life and production activities, such as oxygen consumption of organic matter, nitrogen and phosphorus nutrients, heavy metals, pesticides, chemical carcinogens, radioactive substances and so on. However, in many nonpolluting substances such as inorganic salts, clay minerals, humus, iron and manganese aluminium hydrous oxide, as well as a variety of physical and biological factors such as light, radiation, meteorological, hydrological, ecological, geological and geographical conditions, they constitute the environmental background of the pollutants. They interact with pollutants, or indirectly affect pollutants. These are the contents of water pollution chemistry that must be considered at the same time. Therefore, the research object of water pollution chemistry should be an integrated system composed of pollutants and water environment. From the point of view of pollution chemistry, the meaning of chemical speciation can be summed up as valence state, chemical state, structural state and combination state. These different forms may have an important effect on the pollution effect of pollutants. Therefore, it is not clear to judge the environmental quality according to the total content of some pollutants.

This paper studies the water resources in Taihu Lake from the west of Wuxi, Changzhou and Huzhou area, and the water outlet is in the area of Suzhou. Because the water storage is very stable in that region, pollutants are considerable in Taihu. The Huzhou area will have a lot of water only in the flood season, the region usually has less water, and the content of pollutants is not high. See them on the map, it can be found that they are mainly distributed in Yixing and Wujin area. The region is a gathering place of chemical enterprises, and the phenomenon of illegal discharge of sewage is serious. Taihu Lake, in this paper, the importance of the use of data mining methods in water pollution control is discussed with the water quality assessment in Taihu.

Figure 1: The present situation of water pollution in Taihu Lake

3. The classification process of ls-SVM

Support vector machine (SVM) not only can solve the classification problem well, but also can be applied to regression problems. If the linear case is considered, the function of support vector regression is:

$$f(x) = wx + b \tag{1}$$

Assuming that all training data (x_i, y_i) can be fitted under the accuracy of X, then the minimum problem can be expressed as a convex optimization problem:

$$\min \frac{1}{2}\|w\|^2 \tag{2}$$

$$\begin{cases} y_i - wx_i - b \le X \\ wx_i + b - y_i \le X \end{cases} \tag{3}$$

Considering the regression errors, we introduce the relaxation factors which are a_i=0 and a_i^*=0, then the regression problem can be transformed into the constrained optimization problem:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(a_i + a_i^*) \tag{4}$$

$$\begin{cases} y_i - wx_i - b \le X + a_i \\ wx_i + b - y_i \le X + a_i^* \\ a_i \ge 0 \\ a_i^* \le 0 \end{cases} \tag{5}$$

To solve the above optimization problem of support vector machine, it can be transformed into a quadratic programming problem based on duality theory.
Then, the *Lagrange* equation is established:

$$l(w, a_i, a_i^*) = \frac{1}{2}w^T w + C\sum_{i=1}^{n}(a_i + a_i^*) - \sum_{i=1}^{n}(X + a_i + y_i - w^T x_i - b) \tag{6}$$

The nonlinear regression needs to map the data to the high dimensional feature space, and then carries on the linear regression in the high dimensional space. As long as the appropriate kernel function $K(x_i, x_j)$ is chosen in the original space, it can be equivalent to the inner product $H(x_i, x_j)$ of the nonlinear function of the high-dimensional space. Thus, the nonlinear transformation is realized, and the computational complexity is not increased.
Although there are many kinds of kernel function, we chose the radial basis function as kernel function:

$$K(x, x_i) = \exp(-\frac{\|x - x_i\|^2}{2^{e^2}}) \tag{7}$$

Among them, $\|x - x_i\| = \sum_{k=1}^{n}(x^k - x_i^k)^2$, and e is the width of the kernel.
The principle of least squares support vector machine as shown in Figure 2.

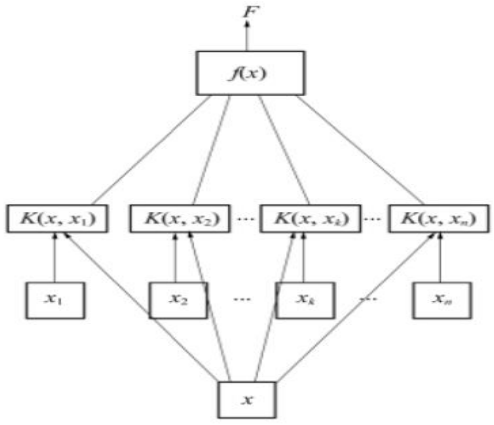

Figure 2: The principle of least squares support vector machine

4. Simulation Experiment and Result Analysis

4.1 The experimental data

In this paper, the national water quality standard (GB3838-2002) is used as the water quality standard. Based on the analysis of water quality monitoring data of 9 monitoring stations in Taihu in 2010-2011, we know that Taihu pollution is mainly organic pollution. Next, we have a brief introduction of 4 conventional water quality indicators which are biochemical oxygen demand (COD), dissolved oxygen (DO), permanganate index (COD mn), and ammonia nitrogen (NH3-N). They are shown in table 1.
PH: The index to characterize the acidity and alkalinity of water. When pH value is 7, the water quality is neutral, less than 7 is acidic, and more than 7 is alkaline. The pH value of natural surface water is generally between 6 and 9. The growth of algae in the water will be caused by the absorption of carbon dioxide by photosynthesis, and the pH value will be increased.

Dissolved oxygen (DO): Express the molecular oxygen dissolved in water. Dissolved oxygen in water is one of the important indexes to reflect water quality. The decomposition consumes dissolved oxygen in the water, which will blacken the water, and kill the fish, shrimp and other aquatic organisms. The growth of algae in the water is due to the oxygen produced by photosynthesis, which will cause the surface dissolved oxygen to rise above the saturation value.

Table 1: Limited value of water quality evaluation index in standard of GB3838-2002

Number	Chemical property	Class Ⅰ	Class Ⅱ	Class Ⅲ	Class Ⅳ	Class Ⅴ
1	PH	6-9				
2	Dissolved oxygen≥	7.5	6	5	3	2
3	Permanganate index≤	2	4	6	10	15
4	Ammonia nitrogen≤	0.15	0.5	1.0	1.5	2.0

Permanganate index (COD Mn): The amount of consumption by using Potassium Permanganate as the oxidant to dispose the surface water samples. Under these conditions, the Potassium Permanganate can be consumed by organic pollutants and the reductive inorganic substance in the water which are ferrite and sulfide. They are often used as a comprehensive index of surface water pollution by organic pollutants.
Ammonia nitrogen (NH3-N): ammonia nitrogen dissolved in the molecular ammonia (also known as free ammonia, NH3) and in the form of ammonium salt (NH4+) in the presence of water. It exists in the form of molecular ammonia with dissolved state and ammonium salt in the water. The ratio of both depends on the water pH value and water temperature, and we use the amount of N elements to express the content of ammonia nitrogen. The source of ammonia nitrogen in water is mainly domestic sewage and some industrial waste water, such as coking and synthetic ammonia. In addition, the surface runoff is the source of ammonia nitrogen, and it mainly refers to the use of fertilizers in agriculture.

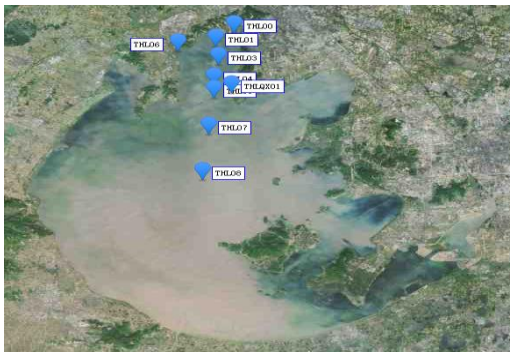

Figure 3: Sketch map of monitoring station in Taihu Lake

The monitoring items of water quality automatic monitoring station include water temperature, pH, dissolved oxygen, conductivity, turbidity, permanganate index, total organic carbon, and ammonia nitrogen. In addition, some of the monitoring stations have the function of testing for volatile organic compounds (VOCs), biological toxicity and chlorophyll A. The monitoring frequency of water quality automatic monitoring station is generally used every 4 hours. 6 monitoring results can be obtained each day, and then merges the data into monthly water quality report. In this paper, the monthly water quality reports of 24 months from 2010 to 2011 are selected as the sample data set, and the data from the first 6 months of the year of 2012 are used as the test data set. In Taihu, there are a total of 9 monitoring station, which numbers are set form THL00 to THL08. Due to space limitations, this article only gives the results data in THL00 monitoring station, as shown in table 2.

4.2 The experiment steps and the result analysis

After we have the sample data, we need to standardize and normalize the data, and then set the parameters of the SVM model:
Train_label: the label of the training set, that is, the grade of water quality.
Train: training set, that is, the data table of 6 kinds of water quality.
C min, C max: The range of change of penalty parameter C, that is, in the [2^cmin, 2^cmax] range to find the best parameter C. This paper takes c min=-6 and c max=6, that is, the range of default penalty parameter C is [2^(-6), 2^6].

Gmin, g max: The range of change of RBF nuclear parameter g, that is, in the [2^gmin, 2^gmax] range to find the best RBF nuclear parameters g. This paper takes g min=-6 and g max=6, that is, the range of RBF nuclear parameter g is [2^ (-6), 2^6].
C step, g step: The step size of C and g in parameter optimization process, that is, the value of C is 2^cmin, 2^ (c min+c step), , 2^cmax, and the value of g is 2^gmin, 2^ (g min+g step),... , 2^gmax.
The results of the training model are shown below:

Table 2: The sample data set of THL00 monitoring station

Number	Station	pH	COD Mn(mg/L)	DO (mg/L)	NH3-N (mg/L)	water temperature (°C)	conductivity (μS/cm)
1	THL00	7.39	6.05	6.46	10.54	4.2	560
2	THL00	7.93	9.49	7.36	6.525	6.7	586
3	THL00	8.04	7.35	8.65	5.715	10.5	520
4	THL00	8.99	13.6	12.5	1.482	19	635
5	THL00	8.23	7.45	6.74	6.016	23.5	570
6	THL00	9.21	11.62	13.44	1.337	26	405
7	THL00	8.09	8.4	6.73	3.676	30.2	475
8	THL00	7.87	7.04	4.9	3.554	29.2	420
9	THL00	7.6	5.1	2.48	3.372	24.2	418
10	THL00	7.41	5.06	1.47	4.809	19.6	425
11	THL00	7.51	5.33	1.9	6.06	14.3	460
12	THL00	7.77	6.82	4.97	5.16	8.4	500
13	THL00	7.7	8.73	2.5	2.854	8.4	610
14	THL00	7.76	5.18	9.49	3.692	8.4	425
15	THL00	7.79	9.89	1.58	10.01	12	1100
16	THL00	7.68	9.71	3.06	4.071	18.8	830
17	THL00	7.65	9.53	2.11	3.864	26	620
18	THL00	7.4	7.04	3.67	3.101	26.2	565
19	THL00	7.78	7.07	2.43	3.46	30	525
20	THL00	7.7	7.41	10.52	0.535	30	485
21	THL00	8.15	5.87	7.46	0.034	26.8	405
22	THL00	7.7	6.29	1.4	5.288	22.4	615
23	THL00	7.87	6.81	3.52	7.524	15	620
24	THL00	7.6	7.98	2.1	5.4	9.5	565

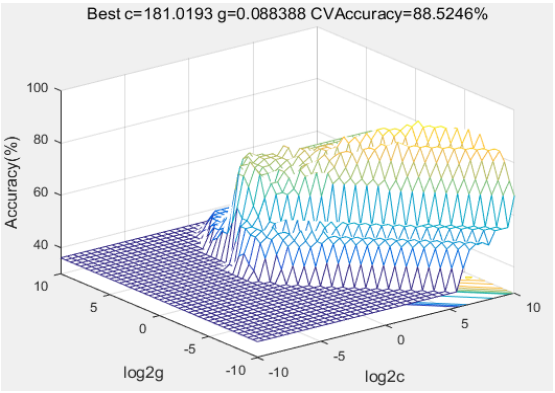

Figure 4. 3D Sketch Map of parameter selection results of SVM

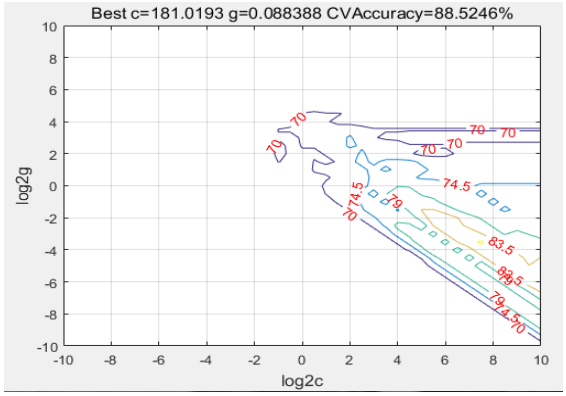

Figure 5: Contour map of parameter selection results of SVM

As can be seen from Figure 5, the optimal parameters of RBF radial basis kernel function are c=181.0193 and gamma=0.0088388, and the precision is 88.5246%. Therefore, we use this as the basic parameters of the model to construct the SVM classification evaluation model. Next, we use the first six months of 2012 as the data set to be detected, and then import it into the constructed SVM classification evaluation model, and the results of the calculation are shown in table 2.

Table 3: The sample data set of monitoring site (THL00)

Number	Actual water quality class	The detection in this paper	by clustering method	by BPNN method
25	I	I	I	II
26	II	II	III	II
27	III	III	IV	III
28	IV	IV	II	II
29	V	V	IV	V
30	V	V	III	III

As can be seen from table 2, through the comparative analysis of the three methods, the evaluation results of support vector machine evaluation method are more consistent with the actual situation of water quality, and the evaluation results are more objective. It gives an accurate assessment of the six samples, and the other two methods are significantly less accurate. Compared with the traditional evaluation methods, this shows that the new method is simple, fast, accurate and practical, and has obvious advantages in dealing with small samples. At the same time, the evaluation model is established by using this method only need to provide the actual observation data to the program. Then, the evaluation results can be obtained by computer analysis and calculation, which can help to realize the standardization of the model, even the automation. Therefore, we can see that the improved support vector machine has a very important value in the comprehensive evaluation of water quality with chemical evaluation standards.

5. Conclusions

With the continuous development of China's industrialization, many water pollution problems occur in this process. After the water is polluted by chemical substances, it has caused serious damage to the ecological environment. At the same time, the life and health of human beings have been greatly threatened. For this reason, this paper establishes an input response relationship between water quality index and the corresponding pollution source, and then introduces the improved support vector regression method to classify and evaluate the water quality of Taihu Lake. The experimental results show that the method has the advantages of high precision, strong practicability, simple calculation and so on. It can give a reasonable water quality classification for the lake water samples, which is worthy of popularization and application.

Acknowledgments

2016 Teaching Reform Research Project of Yulin University (JG1638). National Natural Science Foundation Youth Fund (Project Approval No. 31700633)

References

Chen H.J., Tang Y.J., 1997, Study on water quality early warning and forecast information system of Guangxi River, Shaanxi Hydro Electric Power, 2, 50-52.

Dong Z.Y., Wang J., 2002, Preliminary study on water quality early warning theory, Research on Soil and Water Conservation, 9(1), 36-38, DOI:10.3969/j.issn.1005-3409.2002.03.059

Hou D.B., Chen L., Zhao H.F., 2013, Water quality anomaly detection method based on RBF neural network and wavelet analysis, Sensors and Microsystems, 32(2), 138-141, DOI: 10.3969/j.issn.1000-9787.2013.02.042

Hu X.F., 2006, Study on ecological security assessment and early warning of water environment in Tianjin, Tianjin Normal University, Tianjin.

Li E., 2012, Research and application of cross boundary water pollution accident early warning system, Harbin Institute of Technology.

Liu Y.H., 2011, Study on environmental risk source identification of sudden water pollution accident based on safety theory, Harbin Institute of Technology.

Xiao C., Zhang Y.J., Peng H., 2005, Study and application of water quality early warning and forecasting system: A case study of Wuhan section of Hanjiang River, Environmental Science and Technology, 11(3), 1-6.

Wang C., 2013, Dynamic modeling and simulation of typical organic pollutants in Songhua River, Harbin Institute of Technology.