Research on water pollution prediction of township enterprises based on support vector regression machine

Yue Wang[1], Song Xue[1,2], Junming Ding[3]

[1]School of Business, Hohai University, Nanjing 211100, China
[2]Project Management Institute, Hohai University, Nanjing 211100, China
[3]Construction Management Department, Shandong Water Transfer Engineering Co., LTD., Jinan 250101

Abstract. The construction and development of township enterprises plays a key role in promoting the development of rural economy. With the implementation of the rural revitalization strategy, township enterprises develop rapidly, but there are problems in the development process that have a negative impact on the quality of local rural water environment. Rural water environment is related to the health of farmers, the healthy development of agriculture and the sustainable development of rural areas, so it is necessary to predict the water pollution of township enterprises. The application of support vector regression forecasting model to the prediction of water pollution of township enterprises can better predict the water pollution of township enterprises with the characteristics of complexity, nonlinear and small sample. This intelligent forecasting method will help to scientifically prevent the development of township enterprises from having negative impact on the quality of local water environment.

1 The introduction

Rural water environment is an important ecological security guarantee for rural economic and social development. The quality of water environment is directly related to agricultural production, farmers' life, and the development of rural regional economy and society. It is of great significance to realize coordinated regional development and water environment protection for the sustainable development of China's economy and society. The state attaches great importance to ecological and environmental protection in the development of towns and townships. The 2018 Central Document Opinions of the CPC Central Committee and the State Council on The Implementation of The Rural Revitalization Strategy indicates that by 2035, the rural ecological environment will be fundamentally improved and beautiful and livable villages will be basically realized. By 2050, rural areas will be fully revitalized, with strong agriculture, beautiful rural areas and rich farmers. We will strictly prohibit the transfer of pollution from industry and urban areas to agriculture and rural areas, strengthen rural capacity building for environmental supervision, and ensure that rural areas at the county and township levels assume primary responsibility for environmental protection. Several Opinions of the CPC Central Committee and the State Council on Giving Priority to Agricultural and Rural Development and Doing a good job in the work related to Agriculture, Rural areas and Farmers in 2019, the No. 1 Document of the CPC Central Committee clearly states that green development of agriculture and rural areas should be promoted in a coordinated way in the field of pollution control and environmental protection in rural areas. We will implement the systems of river chiefs and lake chiefs, improve the water environment in rural areas, and strictly manage the shorelines of rivers and lakes and other water ecological spaces in rural areas.

In recent years, due to the rapid economic and social development, urban and agricultural water use has increased substantially, and the discharge of waste water and sewage has also increased simultaneously. A large number of industrial sewage and urban sewage exceeding the standard were discharged into the countryside and used for irrigation. Non-point source pollution from rural agricultural production, pollution from township enterprises, domestic sewage and waste pollution accelerated the decline of rural water environment quality.

The rapid development of township enterprises to made great contributions to the development of national economy, most extensive production of township enterprises, however, to the local environment caused by pollution is getting worse, the main pollutants in the industrial pollution emissions rising, the proportion of the pollution of township enterprises has become the main pollution sources in rural areas. The characteristics of township enterprises, such as high consumption, high emission, low efficiency and decentralized distribution, increase the difficulty of governance.

The water pollution prediction of township enterprises can effectively reduce the rural water environment pollution caused by the development of township enterprises, Sehgal[1]Et al. used linear fitting to predict the trend of water pollution, but in fact the equation fitting

based on linear regression could not meet the demand for accurate prediction of water environmental pollution. Guo Qingchun[2]Et al. used BP artificial neural network model to predict water pollution index of Taihu Lake, and found that with the increase of neural network training times, the effect of error parameters presented uncertainty, and it was easy to fall into local extremum.

Li Yinghui[3]predicted and analyzed the water pollution of the Three Gorges Reservoir based on the grey system model, and the operation process was relatively complex. At present, there are many methods for water pollution prediction at home and abroad with different focuses. Based on the support vector regression machine model, this paper will explore its effect on water pollution prediction of township enterprises, so as to provide a basis for the treatment and improvement of rural water environment and the realization of sustainable development of rural economy.

2 Construction and comparative analysis of water pollution prediction model of township enterprises

Water pollution prediction is a basic work for water environment management planning and water pollution prevention and control. At present, there are some forecasting methods for water pollution at home and abroad, and different industries and monitoring departments choose different forecasting methods. According to the different mechanisms of various water pollution prediction methods, they can be roughly divided into statistical prediction method, intelligent prediction method and mechanism model prediction method [4] .

(1) Statistical prediction methods mainly include regression analysis prediction method, exponential smoothing prediction method and gray system theory prediction method. Regression analysis is widely used, but it requires a large amount of data basis, and is only suitable for medium - and long-term prediction, with large deviations.

(2) The prediction rule of mechanism model is that it requires a deep understanding of the characteristics of real objects and careful analysis of the inherent laws before the model can be established and assumptions can be made to achieve the purpose of water pollution prediction. Wang Siwen et al. [5] used the WASP model to predict the reduction amount of sewage and water environmental capacity of Ashe river and Hulan River in the tributary of Songhua River in 2014. Chen Yue et al. [6] selected CO,

QUAL2K model was used to predict meixi section of Xitiaoxi trunk stream.

(3) Intelligent prediction methods are widely used in artificial neural network (ANN) and support vector machine (SVM), and BP artificial neural network is one of the most widely used ANN models at present. However, BP model is only applicable to the learning of deterministic relations and cannot deal with the contradictory sample wood and samples containing non-real measurable factors[7]. Artificial neural network (ANN) prediction method needs a lot of data to ensure good prediction effect. Compared with the traditional learning method of neural network, support vector machine (SVM) based on structural risk minimization principle, solving a quadratic optimization problem, and get the global optimal solution effectively solves the model selection and learning problems, nonlinear and dimension disasters and the local minimum problem, in solving the small sample, nonlinear and high dimensional pattern recognition problem in exhibit many unique advantages [8] .

(4) Support vector machine (SVM) prediction method is a prediction method based on statistical theory, which requires little data, so it is more suitable for the prediction of water pollution in township enterprises. Support vector machine (SVM) is a new type of machine learning method developed based on statistical learning theory. It is the concrete realization of structural risk minimization criterion [9] . Support vector machine (SVM) can better deal with the prediction of water environment and water quality indexes with the characteristics of complexity, nonlinearity, high dimension, local minimum point, small sample, etc., and it has extensive generalization ability, which has become one of the research hotspots of water quality prediction [10] . Support vector machines have the functions of self-organization, self-learning and associative memory, which overcomes the problems of "overlearning", local optimization and large sample size in general machine learning theories (such as artificial neural networks) [11] . Recently, SVM has become more and more popular methods in data mining fields such as classification, regression and singularity detection. There have been a lot of research reports on SVM internationally, and SVM has been successfully applied in many aspects [12]

3 Construction and analysis of prediction model of support vector regression machine

3.1 Data selection and sources

The accuracy and authority of data are important factors for the reference value of test results. To ensure the accuracy and reliability of the data in this paper, the paper is obtained from China Environmental Statistics Yearbook, China Statistical Yearbook of Township Enterprises, Guangxi Statistical Yearbook, Guangxi Environmental Quality Bulletin, Guangxi Water Environment Quality Bulletin and Guangxi government department website. The statistical data from 2009 to 2018 were selected for analysis in this paper. In order to reduce the modeling error, the data is normalized.

3.2 Construction of regression prediction model

Support vector regression (SVR) usually first trains the model according to the existing data to predict the time series. Set a time series $x=\{xi|xi \in R, I =1..... L\}$, it is hoped to predict the value of the last M moments through the value of the first N moments of the sequence. N data of the sequence can be used as the sliding window and

mapped to M values, which represent the predicted value of the next M moments of the window. The prediction model of the data is listed in table 1 below, which divides the data into K data segments with certain overlap length of N+M, and each data can be regarded as a sample, thus K=L-(N+M) +1 sample. The first N values of each sample can be taken as the input of the network, and the last M values as the output of the network. Through learning, the network can realize the mapping from the input space R^N to the output space R^M, so as to achieve the purpose of time series prediction. N is the number of input variables and M is the number of output variables.

Table1. Data prediction model.

N inputs	M output
X_1, X_2And... , X_N	X_{N+1}, X_{N+2}And... , X_{N+M}
X_2, X_3And... , X_{N+1}	X_{N+2}, X_{N+3}And... , X_{N+M+1}
...	...
X_K, X_{K+1}And... , X_{N+K-1}	X_{N+K}, X_{N+K+1}And... , $X_{N+M+K-1}$

Taking economic development scale index as an example, there are 10 pollution samples. Set the value of sliding window N as 3 and the value of output variable M as 1, from which it can be calculated that there are a total of samples K=10-(3+1)+1=7 in the sample set. The first five groups are taken as training sample sets, and the last two groups of data are test sample sets. The specific input and output vectors are shown in table 2.

Table2. Input and output of training samples and test samples.

The sample name	The input	The output
The training sample	The 2009-2011	2012
	The 2010-2012	2013
	The 2011-2013	2014
	The 2012-2014	2015
	The 2013-2015	2016
Test samples	The 2014-2016	2017
	The 2015-2017	2018

This paper uses LIBSVM to construct support vector machine regression model.. LIBSVM is the use of MATLAB and C language as a platform for a specific support vector machine algorithm software package, which can be changed by each parameter value to achieve kernel function and its correlation coefficient selection optimization, and regression prediction. The selection of kernel function and training parameters is very important to the performance of support vector machines. Therefore, choosing the appropriate kernel function type and parameters plays an important role in improving the prediction accuracy.

Common kernel functions include polynomial kernel function, radial basis kernel function and sigmoid kernel function. Among them, the polynomial kernel function has a slow operation speed, while the sigmoid kernel function will lead to divergence in some cases. If the radial basis function is used, a better balance can be obtained between the calculation time and the prediction effect. In radial basis function networks, the output of the basis function from the input layer to the hidden layer is a nonlinear mapping, while the output is linear. This can be seen as first transforming the original non-linear separable feature space to another space (usually high-dimensional space), making the original problem linearly separable in the new space through reasonable selection of this transformation, and then using it to solve the problem, which conforms to the idea of support vector machine. In this paper, the radial basis function kernel function is selected for the prediction.

In the process of using support vector regression machine to predict, the parameters to be selected mainly include penalty coefficient C, parameters in radial kernel function γ and insensitivity coefficient ε. $K(x, x_i) = \exp\{-\gamma \left|x - x_i\right|^2\}, \gamma > 0$ .In the case of noise in training data, penalty coefficient C can control the complexity of fitting curve, kernel function parameters γ can affect the smoothness of regression curve, and insensitivity coefficient ε can control the generalization ability of the model. γ ε In the actual prediction, the default parameter value provided by the system may not be optimal, so it is necessary to find out the optimal parameter.

Liu Jingxu[13] found that when fixed at different values, the CV error changes with the other two parameters. On the one hand, it shows that a better combination of C γ and C can be found by fixed values ε. On the other hand, it also provides some support for the empirical method of selecting parameters, that is, the best combination of C, γ can always be found for the selected parameters ε, and this combination is robust to the parameters. Based on this, this paper uses a heuristic search algorithm to search for parameters and numbers. The idea of the algorithm is: first of all, for the fixed ε, we use the step search method to find a better group of C, γ and group. Then, we use a certain group C, γ ,in this paper, we use the algorithm GA (genetic algorithm) to search for the best parameters, C and γ. By grid search, we can find the highest class accuracy in the sense of CV, that is, the global local optimal solution, but sometimes if we want to find the best parameter C and γ in a larger range It will be very time-consuming. The heuristic algorithm can be used to find the global optimal solution without having to go through all the parameter points in the grid.

4 Analysis of model prediction results

By using the LIBSVM toolbox to calculate on the MATLAB platform, the final optimal value of the parameters is: when COD is taken as the dependent variable, C=1.7, γ =3.2, P =0.1. Finally, the regression index is obtained through calculation. The Mean squared error of the training set is 0.0107663. Squared correlation coefficient Squared correlation coefficient = 0.925718.Mean squared error of test set = 0.180513;The Squared correlation coefficient of the test set was 0.984908.The measured data and the predicted values

made by the support vector regression machine are plotted in figure 1.

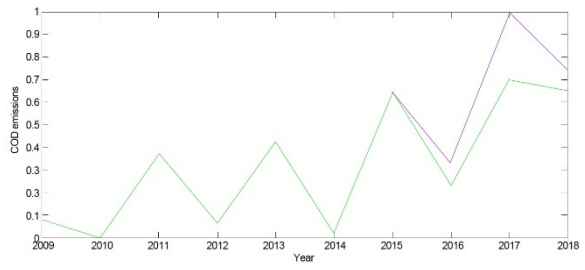

Fig1. Comparison of COD measured and predicted curves.

In the figure above, the seven data from 2012 to 2018 are public data, the top line is the curve drawn by measured values, and the bottom line is the prediction curve by support vector regression model. From the intuitive reflection of regression model indexes and curves, it can be seen that the regression prediction model has strong generalization ability and high prediction accuracy rate.

In the same way, the parameters of the regression model of wastewater discharge were selected, and the final optimal value was: C=2.5, γ =1, p=0.1. After calculation, the regression index was obtained. The Mean squared error of the training set was 0.00866302. Squared correlation coefficient Squared correlation coefficient = 0.84485,Mean squared error of test set = 0.0982102;The Squared correlation coefficient of the test set was 0.976821.The measured data and the predicted values made by the support vector regression machine are plotted in figure 2.

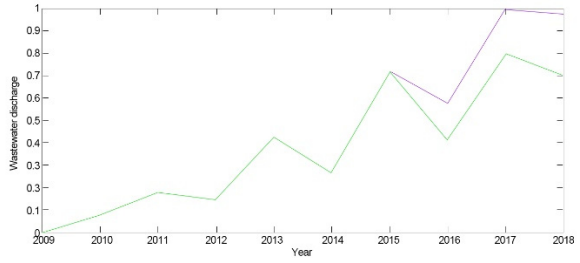

Fig2. Comparison of measured and predicted curves of wastewater discharge

It can be seen from the relevant indicators of the regression model that the prediction model performs well, and the trend of the graph shows that the accuracy of the regression prediction model of support vector machine is relatively high.

The changes of industrial wastewater and industrial COD emissions in Guangxi are characterized by stages, with decreasing and increasing alternately. The average annual growth rate of COD is 3.37% higher than that of wastewater discharge, and the overall change is small, but the fluctuation range is large. Because clean drinking water is directly related to people's life and health, people pay more attention on water pollution than other pollutants. Compared with groundwater, people are more concerned about the pollution of surface water system. The discharge of industrial wastewater directly affects the surface water quality, so although the discharge is decreasing and increasing alternately, the fluctuation range is not large.

5 conclusion

Water pollution prediction is an important content of water environment research and an important work to ensure the quality of water environment. Rural water environment is also related to the implementation effect of Rural Revitalization Strategy. This paper studies the problem of water pollution prediction in township enterprises based on support vector regression machine. Through the construction of support vector regression model for water pollution prediction and analysis in some areas of Guangxi, this paper explores the application effect of support vectorregression machine, a relatively new technology in the field of machine learning. According to the experimental results, in terms of the overall prediction effect, support vector regression machine has advantages over other prediction methods in solving small sample problems and nonlinear problems, and its prediction results can predict the water pollution of township enterprises better. The support vector regression machine has good generalization ability and more concise mathematical form, which is more convenient to use. In the prediction of water pollution in township enterprises, it has certain theoretical significance, promotion value and application value. But support regression vector machine is still in the development and improvement stage. There is a lot of room for development.

References

1. Sehgal V, Tiwari M K, Chatterjee C.Wavelet bootstrap multiple line-ar regression based hybrid modeling for daily river discharge forecas-Ting [J]. Water Resources Management, (2014,28 (10):2 793-2 811).
2. Guo Qingchun, He Zhenfang, LI Li, Li Haining. Application of BP artificial neural network Model in water pollution index prediction of Taihu Lake [J].Journal of southern agriculture, (2011,42(10):1303-1306).
3. Li Yinghui. Research on water pollution Prediction and Treatment Countermeasures of the Three Gorges Reservoir Area based on grey system Model [J].Journal of chongqing university of technology (social sciences), (2020,34(03):46-55).
4. Yu Ting Zhan, Li Yong, Bai Yun, Li Chuan.Research Status of Water pollution Prediction Methods [J]. Green Science and Technology, (2017(14):53-55+58).
5. Wang Siwen, Qi Shaoqun, Yu Dandan, Zhang Yuwei, Wan Luhe.Research on Water Environment Quality Prediction and Evaluation based on WASP Model -- A Case study of Halbinjiang Section of Songhua

River [J].Journal of natural disasters, （2015,24(01):39-45）.

6. Chen Yue, XI Beidou, HE Liansheng, Wang Jinggang. Application of QUAL2K Model in water Quality Simulation of Meixi Section of Xitiao-Xi Trunk Stream [J].Journal of Environmental Engineering, （2008(07):1000-1003）.
7. Jiang Baiquan. Application of artificial neural network in water environment quality evaluation and prediction [D].Capital Normal University,2007.
8. Liu Kun, Liu Xianzhao,Sun Jin, Meng Cuiling.Comprehensive Evaluation of Water Environment Quality based on Support Vector Machine [J]. China Environmental Monitoring, （2007(03):81-84）.
9. Zhang Xiuju, An Huan, Zhao Wenrong, Zhang Qinling.Examples of Water Quality prediction based on Support vector Machine [J]. China Rural Water Conservancy and Hydropower, （2015(01):85-89）.
10. Yu Ting Zhan, Li Yong, Bai Yun, Li Chuan.Research Status of Water pollution Prediction Methods [J]. Green Science and Technology, (2017(14):53-55+58).
11. VapmkV.The essence of statistical learning theory. Zhang Xue-gong translation. Beijing: Tsinghua University Press, 20
12. Zhang Sen. Research on Water Quality Evaluation and Prediction based on Partial Least squares support Vector Machine [D].Chongqing University, （2014）.
13. Liu Jing-xu. Model Selection and Application of Support vector Regression [D]. National University of Defense Science and technology, （2006）.