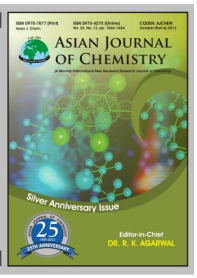
Asian Journal of Chemistry; Vol. 25, No. 14 (2013), 7853-7856

ASIAN JOURNAL OF CHEMISTRY

http://dx.doi.org/10.14233/ajchem.2013.14654



BOD-DO Concentration Forecasting Model Based on Hybrid TS-SVM

JIAYANG WANG*, YUN LIU, ZUOYONG LI and XUEQIAO ZHANG

College of Resources and Environment, Chengdu University of Information Technology, Chengdu 610225, P.R. China

*Corresponding author: E-mail: goodwjj@foxmail.com

(*Received*: 8 October 2012; *Accepted*: 20 July 2013) AJC-13825

A hybrid TS-SVM model is provided, in which taboo search (TS) was used to optimize the key parameters of support vector machines (SVM) to make enhancement on the forecasting effect of support vector machines on BOD and DO. The hybrid TS-SVM model was used in forecasting the concentration of the indexes BOD and DO. Then by forecasting the concentration of the indexes BOD and DO in two cases. The results show that this method not only provides a simple and practical method for water quality forecasting, but also has a better forecasting effect, especially for small sample forecasting problem.

Key Words: BOD concentration, DO concentration, Support vector machines, Taboo search, Forecasting model.

INTRODUCTION

Analysis and forecasting the chemical composition of water quality indicators is the base of water environmental assessment, planning and management. Because there were so many factors related to transformation and migration of the contaminants in water, it is a complex process to simulate and forecast. And it is hardly to obtain the complete statistical data of pollution sources investigation, so the prediction accuracy of water quality is low precision. Now, there were several methods on water forecasting such as pollutant sources investigation method, Mann-Kendall trend analysis model, linear regression method, sliding average method, *etc.*[1-3]. Studies shown that key indicator is oxygen balance to influence the water quality, which indicates the condition of dissolve oxygen. DO and BOD are two important indicators.

The article adopted the hybrid TS-SVM (Taboo Search-Support Vector Machines) model to forecast the concentration of BOD and DO, the method can be used when forecasting other indicators of water quality. Support vector machines (SVM) was proposed by Boser *et al.*[4] and Vapnik[5] to improve the accuracy of classifiers in machine learning and pattern recognition. Support vector machines is a powerful methodology for solving problems in nonlinear classification, function estimation and density estimation which has also led to many other recent developments in kernel based methods in general[5]. The advantage of SVM is that it can achieve high accuracy with relatively small training sets. However, as a key part of statistics theory, the application of SVM was far from the expect effect, main reason is how to choose the parameters of SVM[6,7]. In order to express the performance of SVM fully, the parameters must be chosen appropriately.

Studies shown that error punishment factor C and nucleus function parameter (σ) were not only played a large role in the precision of SVM, but also had a great influence on the performance of machine[6,7]. So, support vector machines and Taboo search are combined to establish an improved hybrid TS-SVM model for BOD and DO concentration forecasting, in which parameters of SVM were optimized by Tobaco search. Here, two representative studies shown that the hybrid TS-SVM model has the feasibility to forecast the concentration of BOD and DO, better results were received in condition of the limited samples. The advantages of this model have been shown in comparing the results achieved by other models.

EXPERIMENTAL

Hybrid TS-SVM model: Support vector machines has the advantages of fast learning, global optimization and strong promotion compares with the traditional neural network methods which is based on empirical risk minimization. It can not only obtain the complex mapping relationship between the dependent variable and independent variables, but also get the evaluation and prediction results significantly better than other methods of pattern recognition and regression[6,7]. However, as a core element in the statistical learning theory, support vector machine's application is far from the desired theoretical effect. Parameter selection problem is a key factor to their application. In order to give full play to the performance of SVM, kernel function parameters must be chosen appropriately. Parameter σ (the kernel function) and C (the error discipline) are the key

factors to the performance of SVM. Selecting the appropriate kernel function parameter σ and the error penalty factor C is essential for the performance of learning machine[6,7]. Therefore, a hybrid TS-SVM model was proposed based on the parameter optimization of SVM by Taboo search[8], the generic flow of TS-SVM is given as follows:

Step 1: Initialization: length of the taboo search list, range of some main parameters, the spanning way of the neighbourhood, the number of cycles, *etc*.

Step 2: Calculates the fitness value: Divide a small portion (generally the 10 % of all training samples) of the sample data used for training as the test sample set. The remaining training samples were used in SVM to fit the requirements. Predicting values were calculated by the test samples after obtaining the parameter values of the SVM model.

Step 3: Updates the taboo search list and the optimal solutions.

Step 4: Check the termination condition, if satisfied, end the optimization; otherwise t = t + 1, go to step 2. Ending conditions are the maximum evolution algebra in optimal achieve or the error is less than a given range;

Step 5: The optimal value of the TS algorithm is the optimal parameter vector (C, σ) in the SVM. Learn all the training samples to get simulation accuracy by the optimized SVM.

Basic principle of SVM: The details of SVM are discussed in previous references[4-7]. Set the training sample $\{x_k, y_k\}^1$, where, $x_k \in R^n$ is n-dimensional input sample, $y_k \in R$ is the output sample. The approximation problem of the function is to find a corresponding function f for samples other than x after training through the sample. Support vector machines estimates the unknown function by the following equation[4,5]:

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

where, $\phi(x)$ presents the mapped high dimensional feature space, b is the deviation value. The following formulae can be obtained by minimizing the risk function to coefficient w and b:

$$\min\left(\frac{1}{2}\| w \|^2 + C\sum_{i=1}^{l} | y - f(x) |_\varepsilon\right) \tag{2}$$

where, the first item $\frac{1}{2}\| w \|^2$ is called the model complexity item, the second item is the empirical error term determined by insensitive. Support vector machines method transforms the above problems to the following antithesis question through introducing dot product nuclear function and using the Wolfe antithesis skills. Expressed as:

$$\max_{\alpha,\alpha^*} J = -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j)$$

$$-\varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*)$$

$$s.t \begin{cases} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases} \tag{3}$$

where, C(C > 0) is the penalty constant. Finally, the corresponding regression function (1) can be directly expressed as follows:

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{4}$$

$k(x_i, x)$ is the nuclear function, which satisfy the Mercer condition. The study selects the Gauss nuclear function:

$$k(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{5}$$

where, σ is the nucleus function parameter.

RESULTS AND DISCUSSION

BOD concentration forecasting of the trunk stream of Dongjiang River in Huizhou City: BOD concentration of the trunk stream of Dongjiang River in Huizhou City was monitored in August 2002. The monitoring data adopted by reported method[9] are shown in Table-1. BOD concentration were forecasted by TS-SVM. BOD concentration forecasting models of Huizhou Bridge, Lantang, Ping Mountain, respectively are shown as follows:

Step 1: Due to the limitation of training sample, the front eight samples were used as the training sample in the hybrid TS-SVM model, the ninth sample was used as test sample. Square root of test result with the actual value is flexibility value to quit the program. The tenth sample was used as the forecasting sample.

TABLE-1
MONITOR SAMPLE AND FORECASTING RESULTS OF BOD CONCENTRATION (mg/L)

Time	Lantang		Ping Mountain		Huizhou Bridge	
	Monitoring data	Prediction value	Monitoring data	Prediction value	Monitoring data	Prediction value
1	1.20	1.2538	0.40	0.4696	0.40	0.4151
2	2.20	2.1724	0.80	0.6362	0.80	0.8049
3	3.20	3.1871	0.80	0.9849	1.40	1.3819
4	4.20	4.1733	1.70	1.6016	1.70	1.7340
5	5.00	5.1951	2.30	2.3252	2.40	2.3624
6	6.20	5.8459	2.90	2.8977	3.40	3.4584
7	5.80	6.1095	3.40	3.3570	4.60	4.5183
8	6.80	6.6733	3.90	3.9904	5.55	5.6297
9	7.50	7.4896	4.90	4.8365	7.10	7.0452
10	7.60	7.6219	5.50	5.4875	7.55	7.5338

Step 2: The time value samples were used as the input sample in forecasting part of TS-SVM, while the BOD concentration samples were used as the output part. When the optimal parameters of SVM optimized by TS were obtained, the hybrid TS-SVM was formed.

Step 3: Forecast the future concentration of BOD by the optimized TS-SVM model.

Parameter setting of TS-SVM is shown as: taboo list length: 100, the max iteration times: 1000, error punishment factor $C \in [0.001, 200]$, nucleus function parameter $\sigma \in [0.001, 2]$. The optimal parameter value are: Lantang: $C = 121.5686$, $\sigma = 0.9974$; Huizhou Bridge: $C = 198.8038$, $\sigma = 0.6782$; Ping Mountain: $C = 198.2813$, $\sigma = 1.4583$. The prediction values of BOD are shown in Table-1.

Dynamics simulation model was common when forecasting the concentration of BOD. So, the analysis of variance between prediction value by dynamics simulation model[9] and by TS-SVM model is shown in Table-2. From the table, it can be concluded that the fitting error and inspection error of prediction value by TS-SVM model is less than that by dynamics simulation model.

The prediction values by TS-SVM model and the monitor sample of BOD are shown in Fig. 1. The simulation results by dynamics simulation model are also shown in Fig. 1.

Another case study on BOD-DO concentration forecasting model: The normal water level of a reservoir is 1140 m and the total storage capacity is 4.589 billion m^3. The capacity of the reservoirs, regulative flow, inflow and the experience values of the corresponding BOD and DO concentration of wet season are shown in Table-3[10]. In the table, the experience values of BOD and DO concentration which can reflect the true value were calculated by empirical formula. The article adopted the TS-SVM model to forecast the BOD and DO concentrations by the factors total inflow, outbound flow and capacity change.

The sample data during May to December were used as the training data in the TS-SVM model, the square root of BOD-DO concentration and the simulation value of TS-SVM was used as training fitness. Parameter vector (C, σ) was optimized by TS. Then, the sample data during January to April are used as the forecasting data in SVM, then the prediction concentration of BOD-DO are obtained.

TABLE-2
ANALYSIS OF VARIANCE

Sampling stations	Dynamics simulation model		TS-SVM	
	Residual sum of squares	Regression sum of squares	Residual sum of squares	Regression sum of squares
Huizhou bridge	344.09	440.20	0.0229	60.0924
Lantang	0.61	42.81	0.2804	43.8487
Ping mountain	4.01	74.79	0.0904	27.9761

TABLE-3
PREDICTION CONCENTRATION OF BOD AND DO (UNIT: mg/L)

Month	Total inflow m^3/s	Outbound flow m^3/s	Capacity of the reservoirs, $m^3/(s.\ months)$	Surface layer				Bottom of the reservoir			
				BOD Experiential data	BOD Prediction data	DO Experiential data	DO Prediction data	BOD Experiential data	BOD Prediction data	DO Experiential data	DO Prediction data
5	99.4	136.2	1043.3	0.418	0.3529	9.133	9.1336	0.359	0.3603	8.201	8.1981
6	341.0	72.1	1312.1	1.087	1.0721	8.272	8.2785	1.564	1.5575	5.177	5.1914
7	906.0	471.1	1747.0	1.176	1.1727	7.764	7.7739	1.473	1.4672	6.871	6.8757
8	614.0	614.0	1747.0	0.685	0.687	7.877	7.8862	0.826	0.8253	7.402	7.4036
9	243.0	243.0	1747.0	0.435	0.4158	8.323	8.3291	0.551	0.5518	7.916	7.9147
10	198.0	198.0	1747.0	0.351	0.3988	8.962	8.9637	0.417	0.4182	8.308	8.3045
11	99.4	178.2	1668.1	0.215	0.209	9.780	9.7762	0.233	0.2358	8.659	8.6535
12	58.8	131.0	15959	0.165	0.1652	10.775	10.7644	0.168	0.1724	8.039	8.037
1	50.6	134.6	1511.8	0.179	0.209	11.160	11.1507	0.177	0.19	8.013	8.0118
2	52.9	139.7	1424.9	0.193	0.2085	11.131	11.1278	0.193	0.1734	8.035	8.0342
3	32.0	146.9	1310.1	0.141	0.1539	10.646	10.639	0.120	0.1292	8.319	8.3155
4	36.5	158.3	1247.8	0.121	0.1423	9.744	9.7861	0.136	0.1348	7.661	7.8402

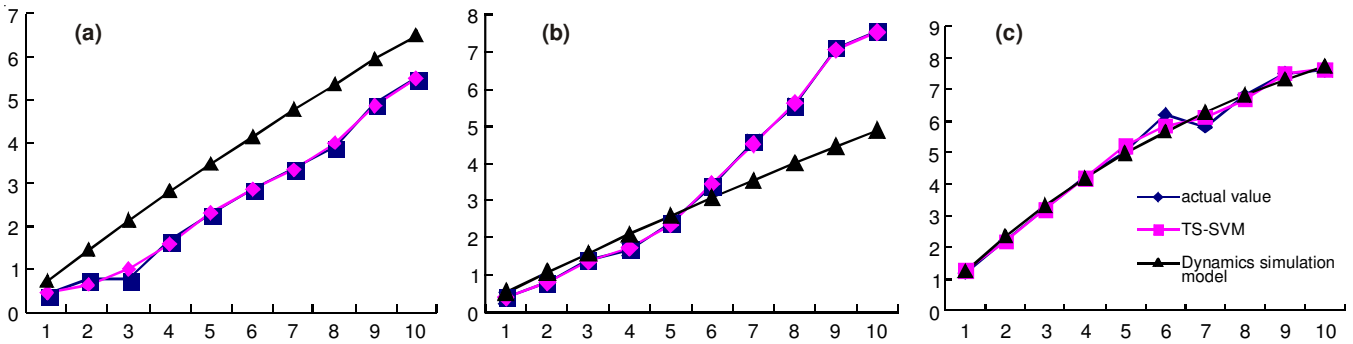


Fig. 1. Prediction value of BOD concentration in Huizhou Bridge (a), Ping Mountain (b), Lantang (c), respectively

The optimal parameter value are: BOD from bottom of the reservoir: C = 126.3479, σ = 0.1285; DO from bottom of the reservoir: C = 173.9194, σ = 0.0037; BOD from reservoir surface: C = 193.9464, σ = 9.9809; DO from reservoir surface: C = 145.5814, σ = 0.0071. The prediction values by hybrid TS-SVM and the calculative values by empirical formula are shown in Fig. 2. It can conclude that the prediction values are close to that calculated by empirical formula in the figure.

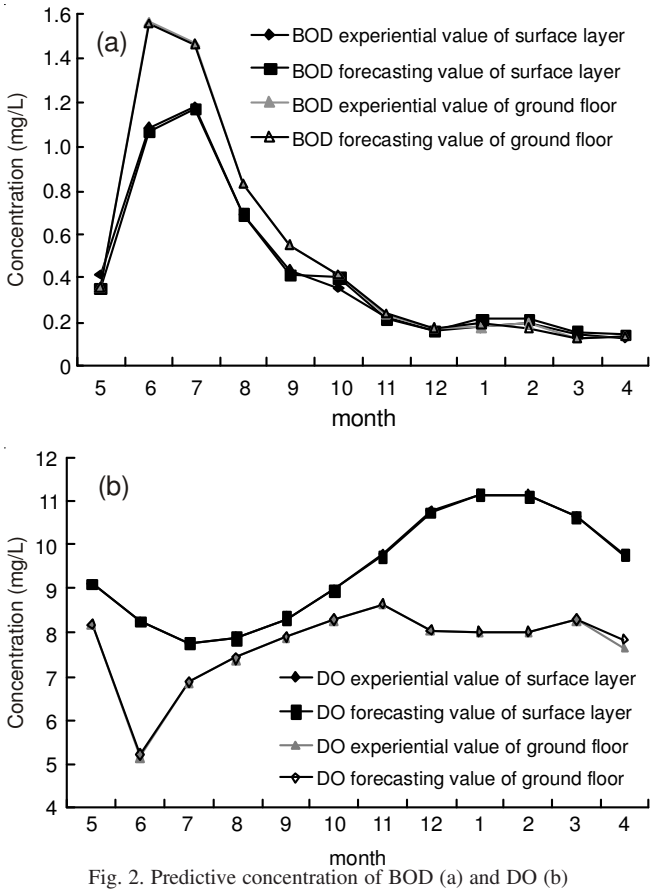


Fig. 2. Predictive concentration of BOD (a) and DO (b)

Conclusion

• SVM and TS were combined to establish an hybrid TS-SVM model, in which TS was used in tuning parameters of SVM, in order to overcome the blindness of the model parameters. The hybrid TS-SVM model has a more clearly theoretical guidance compared with other methods which present widely used through cross-validation method.

• A simple and practical method TS-SVM was proposed for concentration forecasting of BOD and DO, results shown that it can gain an ideal forecasting effect by the way of combination, especially for small sample forecasting problem.

• The article only adopted Gauss nuclear fuction as the kernel function, but in fact different kernel function selections have direct effect to SVM model. Therefore in-depth studies on how to choose other type of kernel function will be needed in future.

ACKNOWLEDGEMENTS

This work was financially supported by the National Natural Scientific Foundation of China (No. 51209024), Soft Science project of Sichuan Department of Science and Technology (No. 2013ZR0080), the Science Research Foundation and the Science Talents Foundation of Chendu University of Information Technology (No. J201214, KYTZ201303).

REFERENCES

1. S.E. Jorgensen, *Lake Reservoirs: Res. Manage*., **3**, 139 (1998).
2. G. Bendoricchio and G. De Boni, *Ecol. Model.*, **184**, 69 (2005).
3. R. Revelli and L. Ridolfi, *Adv. Water Resour*., **27**, 943 (2004).
4. B.E. Boser, I.M. Guyon and V. Vapnik, A Training Algorithm for Optimal Margin Classifiers, The 5th Annual ACM Workshop on Computational Learning Theory (1992).
5. V. Vapnik, The Nature of Statistical Learning Theory, NY, Springer (1999).
6. V. Cherkassky and Y.Q. Ma, *Neural Networks*, **17**, 113 (2004).
7. S.-W. Lin, K.-C. Ying, S.-C. Chen and Z.-J. Lee, *Expert Syst. Appl*., **35**, 1817 (2008).
8. V.V. Kovacevic-Vujcic and M.M. Cangalovic, *Comp. Math. Appl.*, **37**, 125 (1999).
9. X.P. Mo, *People's Pearl River*, **63**, 68 (2005).
10. F.W. Zhang, D.X. Chen and E. Wang, *J. Liaocheng Univ. (Nat. Sci.)*, **18**, 27 (2005).