2018 17th IEEE International Conference on Machine Learning and Applications

Design of River Water Quality Assessment and Prediction Algorithm

Sheng Cao
Department of Microelectronics Science and Technology
Harbin Institute of Technology
Harbin, China
sky_sheng1995@hotmail.com

Shucheng Wang
Department of Information Engineering
Harbin Institute of Technology
Harbin, China
wangshucheng0424@foxmail.com

Yan Zhang
Department of Information Engineering
Heilongjiang University
Harbin, China
zhangyanheida@126.com

***Abstract*—Due to the rapid population growth and economic development, water environmental protection pressures has been increasing recently. This paper focuses on the pollution of water quality, building a water quality assessment model to analyze the water quality level, and makes an objective further prediction of the trend of its factors. In this paper, the mutation factor of genetic algorithm is introduced into the PSO algorithm. The Least Squares Support Vector Machine (LS-SVM) based on adaptive Particle Swarm Optimization (PSO) algorithm used to optimize the hyper-parameter builds one water quality classification assessment model. The fuzzy information granulation method is combined with the Least Square Support Regression (LS-SVR) to set up a water quality time series model, which can predict the trend of changes in water quality data in three days. With the help of the theoretical analysis and experimental data, this assessment model and the prediction algorithm are faster in training speed and higher in accuracy, compared with the traditional BP neural network.**

Keywords*—*water quality assessment, water quality prediction, least squares support vector machine

I. INTRODUCTION

As is known to us, water is a vital guarantee for human survival. It is of utmost importance to automatically evaluate and predict water quality [1]. An assessment of water quality, as well as a prediction theory and analysis methods are integrated in paper, building a model and prediction model that correspond with water quality factors. Not only does it provide more scientific and reasonable assessment of water pollution levels, but can be more objective and accurate in the reflection of the future trend of water quality [2].

Support Vector Classification Machine (SVC) is superior to neural network algorithm in solving practical problems such as high-dimensional nonlinearity and small sample size problems [3]. Least squares support vector sorter (LS-SVC) is an improved SVC algorithm which inherits many advantages of SVC and can ease the workload and reduce the worktime. This paper uses LS-SVC to set up a water quality assessment model to evaluate the water quality. The adaptive Particle Swarm Optimization (PSO) algorithm and the cross validation principle are used to optimize the parameters. Finally, we evaluate the water quality data by the trained model and analyzed the water quality.

Support Vector Regression (SVR) known as a machine learning algorithm performs an optimal compromise solution between the model's learning ability and the VC dimension to obtain the best generalization ability [4]. This paper sets up a partial minimum LS-SVR water quality prediction model based on the historical data of water quality factors, taking the historical data as the dependent variable. The water quality factor is used as an independent variable to predict the further trend of water quality. In order to do that, this paper combines the fuzzy information granulation method with LS-SVR to set up a time series water quality model, which can predict the trend in the next three days.

The experimental results show that the LS-SVC model based on the adaptive Particle Swarm Optimization algorithm can predict the water quality more accurately than the BP neural network, with a considerable faster training speed. The LS-SVR model combined with the fuzzy information granulation method have a more accuracy to predict the changing trend of water quality.

II. RELATED WORK

Paper [5] presented a holistic and low cost approach to the water quality monitoring problem for drinking water distribution systems and for consumer sites. It is based on the development of low cost sensor nodes for real time and in-pipe monitoring and assessment of water quality on the fly. Paper [6] presented a portable sensor implemented as an electronic embedded system featuring disposable measurement cells, which is suitable of measuring bacterial concentration in water samples. A new algorithm, named the multigrouped particle swarm optimization (MGPSO), for the multimodal function optimization based on the particle swarm optimization (PSO) is proposed in paper [7]. MGPSO kept basic concepts of the PSO, and, thus, shows a more straightforward convergence compared to conventional hybrid type approaches. In order to reduce the additive errors and improve the accuracy of measurement, a suitable temperature compensation algorithm is proposed in paper [8] and the Least squares support vector regression (LS-SVR) is adopted to adjust temperature errors for real measured data.

III. WATER QUALITY ASSESSMENT MODEL

The Least Squares Support Vector Machine (LS-SVC) is improved on the basis of SVC [9]. It replaces the slack variable with the square of the training error and replaces the inequality constraint with the equality constraint to avoid understanding the quadratic programming problem. This can speed up the training time and reduce the workload.

This paper introduces a mutation-based particle swarm algorithm combined with the LS-SVC algorithm to establish a water quality assessment model. Not only does it have a more accurate classification of water quality, but also greatly enhances the speed of hyper-parametric search and model training.

978-1-5386-6805-4/18/$31.00 ©2018 IEEE
DOI 10.1109/ICMLA.2018.00146

A. Least Squares Support Vector Machine (LS-SVC)

For two types of classification problems, the training sample set is $S = \{(x_i, y_i), x_i \in R^n, y_i \in \{-1, +1\}\}_{i=1}^{l}$, in which x_i represents the input variable, y_i represents the corresponding category, and l represents the number of training samples.

Then the objective function of the Least Squares Sorter (LS-SVC) is

$$\min \frac{1}{2} \| w \|^2 + \frac{\gamma}{2} \sum_{i=1}^{l} e_i^2 \tag{1}$$

The constraints are:

$$y_i (w^T \phi(x\ ) + b) = 1 - e_i, i = 1, 2. \cdots, l \tag{2}$$

Using the Lagrange factor method can be expressed as:

$$f(x) = \sum_{i=1}^{l} \alpha_i^* y_i K(x, x_i^*) + b \tag{3}$$

The kernel function of least support vector machine (LS-SVM) used in this paper is Radial Basis kernel Function (RBF). The performance of the LS-SVM largely depends on the penalty factor γ and the parameter σ of the radial basis kernel function. A small change in these two parameters may cause the difference in the accuracy of the classification and prediction. So it is important to quickly find the optimal value of these two parameters.

B. PSO Algorithm with Mutation Factor

As a heuristic intelligent algorithm, Particle Swarm Optimization (PSO) can solve some optimization problems with non-convergence problems and obtain the optimal solution of them [10].However, because of its small number of parameters and the relatively simple search mechanism, the result tends to fall into a local optimum, which is also the main reason for limiting the application of the particle swarm algorithm in wider areas.

Absorbing the merits of genetic algorithm, the mutation operation introduced in the PSO algorithm means reinitializing the variables such as position and velocity with a certain probability. The search principle of PSO algorithm is a process of changing from diversity to identity. In the final state, particle population concentrates to the neighborhood of the current global best point for intensive search, losing its particle population diversity. Then, the introduction of mutations in the basic PSO can broaden the search space which is shrinking in the PSO iterations, enabling the particles to jump to the currently searched optimal position to search in a larger space and guaranteeing population diversity, which increases the possibility of an optimal value. The specific scheme is apply the velocity and position variation to the particle with a certain probability in the iteration. The variation formula is as follows:

$$if \quad rand > 0.6 \tag{4}$$

$$v_i^{k+1} = v_{\min} + \xi\left(\frac{v_i^k - v_{\min}}{v_{\max} \text{-} v_{\min}}\right) \tag{5}$$

$$\begin{aligned} x_i^{k+1} &= Pbest_i \\ &+ c_m (x_{\max} \text{ - } x_{\min})(rand - 0.6) \end{aligned} \tag{6}$$

In the iterative process, the PSO algorithm updates the positions and velocities of partial particles in the form of probability. In this method, the Logistic chaotic sequence variogram function is used to update the velocity, and the position is updated with the global optimal slight variogram function. ξ is a n-dimensional random vectors subject to (0,1) Gaussian distribution, c_m is the coefficient of variation. In this method, the search space can be continuously updated during the iterative process to prevent the algorithm from prematurely converging into a local optimum and finally obtain an optimal fitness matrix.

C. LS-SVC Based on Adaptive Particle Swarm Optimization

For the LS-SVM model with Gaussian Radial basis Function (RBF) as kernel function, the parameters to be optimized are RBF parameter and penalty factor γ [11]. According to the PSO algorithm, the algorithm can quickly obtain the optimal solution when performing LS-SVM parameter optimization. The value range of the parameters σ and γ to be optimized is taken as the solution space of the particle, and the adaptive PSO algorithm is used to optimize the two parameters.

The steps of using adaptive mutation particle swarm algorithm for LS-SVM optimization are as follows:

1) Initialize algorithm parameters: population number s initial position and velocity of particles, the maximum of inertia weights and the minimum of it are $\omega_{\min}$, $\omega_{\max}$, learning factors c_1, c_2, and so on.

2) Use the initialization coordinate of each particle as the initial value of kernel function parameter σ and the penalty factor γ . Use training data to train the LS-SVM model to obtain the initial fitness function f_i of each particle and all the average of the particle fitness value f_{avg} . Then, the coordinates of the fitness value minimum particle (whose fitness value is $f_{\min}$) are taken as global extremum points, and the current position of each particle is taken as its local extremum point.

3) Calculate the inertia weight w of the particle, and update the particle velocity and position.

4) Compare the current fitness value of the particle with its optimal fitness value. If it is better than the local extremum point, replace the local optimum coordinate with the current particle coordinate. If the fitness value of the particle is better than the global optimal fitness value, the replacement is also performed.

5) Determine the particle mutation according to (4). If the generated random number is bigger than 0.6, the mutation in particle position and velocity is performed according to (5) and (6).

6) Judge weather the accuracy requirement or the maximum number of iterations is satisfied. If it is satisfied, the iteration is ended and the results can be output, otherwise, the process returns to step 3).

D. Build Water Quality assessment Model

The main steps of building a water quality assessment model that combines LS-SVC and an adaptive PSO are as follows:

1) Standardize the training and testing samples randomly selected by the National Surface Water Monitor Center.

2) Train the multi-class LS-SVC model with the training samples, and optimize the parameters of the model through the 5-fold cross validation principle and adaptive particle swarm algorithm until a set of parameters with the highest classification accuracy is found.

3) Use the parameters (σ, γ) found in step 2) to test the classification accuracy of training samples and test samples.

4) Treat the water quality data and training samples of the Yangtze River collected in this project in the same standardized way to eliminate the effects of different dimensions. Finally, the trained LS-SVC model was used to automatically evaluate the water quality data of the Yangtze River.

IV. WATER QUALITY PREDICTION MODEL

In the application of water quality prediction, we hope to accurately predict the specific values of water quality factors, and to predict the future development trend of water quality factors. With the increase of collected data and time series, it is important to simplify the timing without losing the important information contained in the original time series data. Therefore, this paper combines the LS-SVR prediction method and the fuzzy information granulation method to predict the development trend of water quality factors which is important in the Yangtze River.

A. Least Squares Support Regression (LS-SVR)

For the training set of the regression problem, in the least squares support machine regression algorithm (LS-SVR), the corresponding optimization problem is:

$$\min_{w,b,e} \frac{1}{2} \| w \|^2 + \frac{\gamma}{2} \sum_{i=1}^{l} e_i^2 \tag{7}$$

The constraints are

$$y_i = w^T \phi(x_i) + b + e_i (i = 1, 2, \cdots, l) \tag{8}$$

The Lagrange factor method can be expressed as

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) + b \tag{9}$$

The least squares support regression (LS-SVR) kernel used in this paper is the radial basis kernel function (RBF). Similar to LS-SVC described above, the water quality prediction model built in this paper is also optimized using adaptive Particle Swarm Optimization to find the optimal penalty factor γ and the radial basis kernel function parameter σ .

B. Fuzzy Information Granulation Method

The Fuzzy Information Granulation (FIG) of time series needs to divide information into independent windows first. If the window size is selected properly, not only the original information of the data can be maintained, but also the time series can be effectively simplified, which helps us to discover the main knowledge in the time series. Next, blur each window to generate fuzzy information particles. That is, generate a fuzzy set on the original data set which is used to replace important information in the original data.

For a given data set X , if there is a grained window of size K in the set, and there are K consecutive elements in the window, then we call this non-overlapping window W_G . Therefore, the goal of fuzzy information granulation is to generate a fuzzy set P in the data set X , which can reasonably represent the granulation window as a fuzzy concept of information particles. Therefore, the following two optimization problems have been formed:

1) Minimizing of support of particle fuzzy sets. This can be achieved by minimizing the support set, that is

$$\min \sup p(b-a) \tag{10}$$

Among them, b and a denote the upper and lower bounds of the support set of the particle fuzzy set P ;

2) Maximizing the sum of particle membership fuzzy degree P membership. The greater the sum of degrees of the fuzzy information particle P , the more information that the particle contain, the more they can represent the original data set. That is

$$\max \sum_{k=1}^{K} f(x_k) \tag{11}$$

Among them, $f(x_k)$ represents the membership function of P . In order to satisfy the two optimization problems, they can be combined by a performance indicator function:

$$Q = \frac{\sum_{k=1}^{K} f(x_k)}{\sup p(b-a)} \tag{12}$$

Generally, there are many membership functions that can constitute a fuzzy set P , mainly parabolic, triangular, Gaussian and so on. This paper selects the triangular membership function as shown in (13) to represent the set P .

$$f(x,a,m,b)=\max(\min(\frac{x-a}{m-a},1-\frac{x-m}{b-m}),0) \quad (13)$$

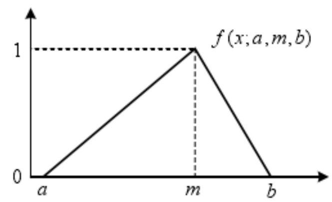


Fig. 1. Triangular membership function

Fig.1 shows the triangular membership function. Through this function image, we can obtain an important membership function simplification, that is, the support set $\sup p(b-a)=b-a$.

For a single window W_G, we assume that the W_G consists of a non-decreasing set $\{x_1,\cdots,x_M,\cdots,x_K\}$. Then, the set $\{x_1,x_2,\cdots,x_M\}$ can be used to represent the left range of the fuzzy set of membership functions, and $\{x_M,x_{M+1},\cdots,x_K\}$ is used to represent the right range. As shown in Fig.1, the triangle membership function can be divided into two parts. For the left range, the membership function expression is:

$$f(x,a)=\frac{x-a}{m-a} \quad (14)$$

It's performance index function $Q(a)$ is expressed as:

$$Q(a)=\frac{\sum_{k=1}^{M}(x_k-a)}{(m-a)^2}=\frac{\sum_{k=1}^{M}x_k-Ma}{(m-a)^2} \quad (15)$$

Similarly, its performance index function is expressed as:

$$Q(b)=\frac{\sum_{k=M+1}^{K}x_k-b(K-M)}{(m-b)^2} \quad (16)$$

The fuzzy parameters a,b,m constitute the important parameters of the fuzzy information particle, wherein the parameter a represents the minimum value of the original data change corresponding to each window, which is called the feature lower bound value; and parameter b represents the maximum value of the original data change, which is called the feature upper bound value. The parameter m represents the mean level of the data change, which is called the characteristic modal value. Therefore, we can use the feature interval a,b,m after the processing of FIG to describe the information of the original data.

C. LS-SVR Water Quality Prediction Model Combining Fuzzy Information Granulation

The main steps to build a water quality interval prediction model combining LS-SVR and fuzzy information granulation are as follows:

1) Set the size of the window to 3, based on the number of water quality samples in the project, and then perform fuzzy information granulation on the original water quality data according to (13), obtaining three groups fuzzy granulation sequence after granulation:

$$A=\{a_1,\cdots,a_W\}$$
$$B=\{b_1,\cdots,b_W\} \quad (17)$$
$$M=\{m_1,\cdots,m_W\}$$

2) Separately normalize the three groups of information granulated sequences by:

$$p_i=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \quad (18)$$

Among them, P_i is the standardized data of the original water quality factor x_i, $x_{\max}$ is the maximum value in the sample, and $x_{\min}$ is the minimum value in the sample.

3) Use the normalized three groups of information-granulated sequences to establish the LS-SVR model with adaptive PSO algorithm.

4) Use the optimized LS-SVR model to predict each sequence, and predict the next window a,b,m to achieve the predictions of intervals in which water quality factor will change in the next 3 days.

V. EXPERIMENTS ANALYSIS

A. Evaluation of Water Quality Assessment Model Based on Adaptive PSO

In order to verify the accuracy of multi-class LS-SVC model for water quality assessment and prediction, this paper randomly selected 200 sample data from the National Surface Surveillance Center to complete the training of the multi-class LS-SVC model. Using the adaptive mutation PSO algorithm and the best parameters searched by the cross-validation principle, the multi-class LS-SVM model was trained with 200 training samples input. Select randomly 100 test samples to complete the water quality assessment automatically. The sample set is shown in TABLE I. Compared with the actual water quality level, the correct rate of multi-class LS-SVM assessment and prediction can be obtained.

TABLE I. TRAINING SAMPLES SET

No.	*PH*	*DO*	*COD*	*NH-3*	*Category*
1	7.6	8.3	0.08	1.6	I
…	…	…	…	…	…
200	7.09	8.3	11	1.14	V

This paper uses the adaptive mutation PSO algorithm and cross-validation principle to optimize the multi-class LS-SVC water quality assessment model. Fig.2 is the curve of the best fitness and the average fitness value of the adaptive mutation PSO algorithm for parameter optimization. The best parameters found were γ =91.501, σ =19.912, and the best accuracy rate was 91%. Fig.3 show the results of the prediction of training samples and test samples used the model. The accuracy of the model in predicting the training sample is 97%, and the accuracy of the prediction of the test sample is 94%.

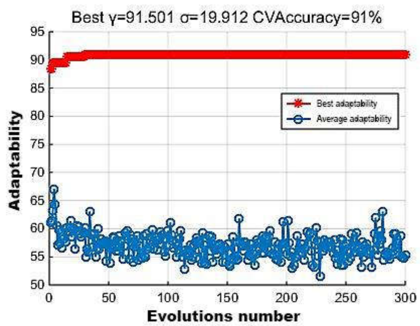
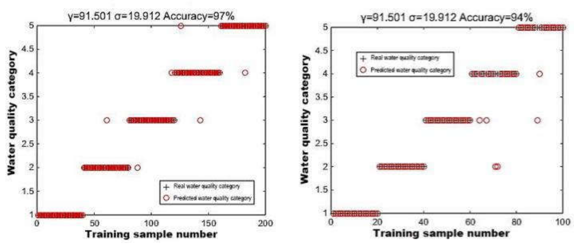


Fig. 2. PSO fitness function curve

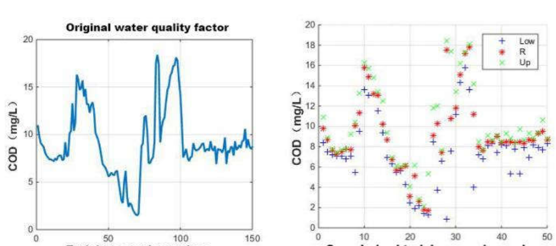


a) training samples b) test samples

Fig. 3. Water quality assessment results of training set

B. *Evaluation of Water Quality Prediction Model Based on Fuzzy Information Granulation*

According to COD water quality data of 160 groups (in terms of "days") actually monitored by a sewage treatment plant in Yangtze River from December 10, 2017 to May 20, 2018, water quality interval prediction modelling studies were conducted. The first 150 groups, 153 groups and 156 groups of data were processed for fuzzy information grain processing, and the LS-SVR regression model was used to predict the water quality interval change in the next window of each group of data. That is to predict water quality intervals in the next three days. Fig.4 a) shows the variation of the raw data of 150 groups of COD water quality factors.



a) raw data curve b) fuzzy information granulation views

Fig. 4. COD water quality factor raw and granulation view

The water quality data in Fig.4 a) is granulated for each of the three groups as a window. After the granulation, the curve of the highest value UP (b) of each window and the average value R(m) and the lowest value LOW(a) is visualized, shown in Fig.4 b).

The lowest value LOW (a) is used as an example to illustrate the process of establishing the LS-SVR prediction model. First, the minimum value LOW(a) sequence is standardized, and then the adaptive mutation PSO algorithm and the cross validation principle are used to perform parameter optimization to get the preferred parameters γ =325.1347, σ =123.1001,which can be used to predict the LOW(a) data series and the original grained data is compared with the predicted data. The result of the comparison of the original grained data and the predicted data is shown in Fig.5. The prediction result of the LOW (a) in the next window using this model is 8.1161.

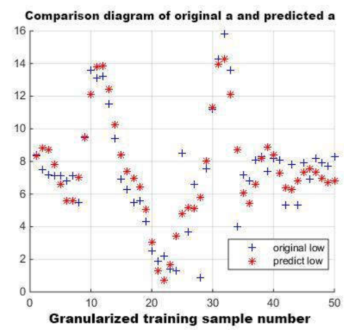


Fig. 5. LOW prediction graph after granulation of fuzzy information

Using the same prediction method as described above for the lowest value LOW(a), the predicted values of the average value R(m) and the highest value UP(b) after granulation are shown in Fig.6 and Fig.7, respectively. For the next window, the prediction results for R(m) and UP (b) are 9.3630 and 8.8980, respectively.

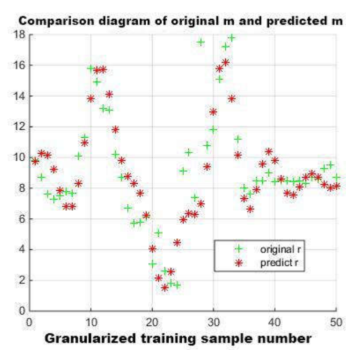


Fig. 6. R prediction graph after granulation of fuzzy information

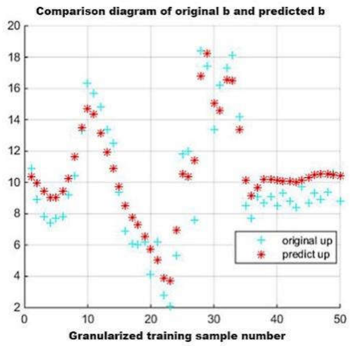


Fig. 7. UP prediction graph after granulation of fuzzy information

In this paper, the first 153 sets and 156 sets of data of 160 sets of monitored COD water quality factors were respectively used for the next window interval prediction. The specific forecast results are shown in TABLE II.

TABLE II. COD WATER QUALITY REAL VALUE (RV) AND PREDICTION VALUE (PV)

	UP		*LOW*		*R*	
	RV	PV	RV	PV	RV	PV
150 groups	9.12	9,3630	8.24	8.1161	8.79	8.8980
153 groups	8.51	8.6129	8.37	8.1965	8.43	8.5030
156 groups	9.32	9.2708	8.51	8.3012	8.53	8.6180

Taking advantage of the fuzzy information granulation method and the LS-SVR water quality prediction modelling method described above to predict the interval of the next three days by using the first 150 groups, the first 153 groups and the first 156 groups of the NH3-N water quality data. The comparison of the prediction and actual value results were shown in TABLE VIII.

TABLE VVVI. NH3-N WATER QUALITY REAL VALUE (RV) AND PREDICTION VALUE (PV)

	UP		*LOW*		*R*	
	RV	PV	RV	PV	RV	PV
150 groups	0.85	8.7728	0.72	0.7431	0.80	0.8351
153 groups	1.01	1.0350	0.68	0.7116	0.84	0.8512
156 groups	0.72	0.7392	0.51	0.5324	0.59	0.6132

From the TABLE V, it can be seen that the prediction of the interval variation of the water quality factors COD and NH3-N is consistent with the actual water quality factor data. It can be concluded that the method of water quality factor interval predicting combined with the fuzzy information granulation and the LS-SVR method dose have a good effect.

VI. CONCLUSION

This paper focuses on the pollution of river water quality, builds a multi-class LS-SVC water quality assessment model to achieve water quality assessment, and sets up an important water quality prediction model based on historical data and fuzzy information granulation to get the predictions for the specific values and future interval changes of important water quality factors, then to predict the water quality levels in the future.

This paper has completed the establishment of multi-class LS-SVC water quality assessment model, which combines the adaptive mutation PSO algorithm to complete and accelerate the optimization of parameters. Through the experimental simulation results, we can see that the improved PSO algorithm can improve the results' accuracy by 1.5% and the convergence speed is significantly faster when compared with the basic PSO algorithm.

This paper also builds an important water quality factor prediction model based on fuzzy information granulation for the river water quality. And the water quality interval variation in the next three days can be predicted. Through the simulation results, we can see that the prediction intervals and the water quality factors of all water quality factors are consistent with their actual ranges. The prediction of future water quality levels and their interval changes can be achieved by analyzing the prediction of the important water quality and its change intervals, while the results are consistent with the actual levels.

In summary, the multi-class LS-SVC water quality assessment model established in this paper can objectively evaluate the water quality of each site, and the water quality prediction model can realize the prediction of specific data and future interval changes of important water quality factors, achieving the forecasting the future water quality conditions.

REFERENCES

[1] Magnus Larson, Lennart Jönsson. Remotely sensed assessment of water quality levels in the Pearl River Estuary, China[J]. Marine Pollution Bulletin, 2007, 54(8)：1267-1272.

[2] Taheriyoun M, Karamouz M, Baghvand A. Development of an entropy-based fuzzy eutrophication index for reservoir water quality evaluation[J]. Iranian Journal of Environmental Health Science & Engineering, 2010, 7(1)：1-14.

[3] Xi-zhao Wang, Shu-xia L. Improved Fuzzy Multicategory Support Vector Machines Classifier. International Conference on Machine Learning and Cybernetics, 2006, 3585 – 3589.

[4] Lingmin He, Xiaobing Yang, Huijuan Lu. A Comparison of Support Vector Machines Ensemble for Classification. International Conference on Machine Learning and Cybernetics, 2007, 6, 3613-3617.

[5] Theofanis P, Lambrou, Christos C. A Low-Cost Sensor Network for Real-Time Monitoring and Contamination Detection in Drinking Water Distribution Systems. IEEE Sensors Journal, 2014, 14(8): 2765 - 2772.

[6] Marco Grossi, Roberto Lazzarini, Massimo Lanzoni, et al. A Portable Sensor With Disposable Electrodes for Water Bacterial Quality Assessment. IEEE Sensors Journal, 2013, 13(5): 1775-1782.

[7] Jang-Ho Seo, Chang-Hwan Im, Chang-Geun Heo, et al. Multimodal function optimization based on particle swarm optimization. IEEE Transactions on Magnetics, 2006, 42(4): 1095-1098.

[8] Ping He, Guofeng Pan, Lin Li, et al. Temperature compensation for pressure sensor based on LS-SVR. Proceedings of the 29th Chinese Control Conference, 2010, 4813-4816.

[9] Zi-Jie Chen, Bo Liu, Xu-Peng He. A SVC Iterative Learning Algorithm Based on Sample Selection for Large Samples. International Conference on Machine Learning and Cybernetics, 2007, 6, 3308-3313.

[10] Shuzhou Wang, Bo Meng. PSO Algorithm for Support Vector Machine. Third International Symposium on Electronic Commerce and Security, 2010, 377-380.

[11] M. Hoseinzade, K. Mohamedpour, S.M.H. Andargoli, et al. OFDM Channel estimation based on Gaussian Radial basis function interpolation. 11th International Conference on Advanced Communication Technology, 2009, 1, 9-13.