Output prediction of CMF based on improved hybrid genetic algorithm and support vector machine

Xu Danyu[1,a],Shi Yan[2,b],You Yangyang[2,c], Duan Yunxia[1,d], Hou Ying[2,e]

(1 Tianjin Academy of Environmental Science, Tianjin 300191, China ;

2 Tianjin United Environmental Engineering Design Co.,Ltd., Tianjin 300191, China.)

[a]danyu079@126.com, [b]147603835@qq.com, [c]654744669@qq.com, [d]284176690@qq.com, [e]510792531@qq.com

Keywords: continuous micro-filtration;support vector machine;accelerating genetic and simulated annealing algorithm;BP neural network;membrane flux

Abstract: Through improved select tactics and genetic operators, the accelerating genetic algorithm (AGA) and simulated annealing algorithm (SA) were combined to form a new algorithm called accelerating genetic and simulated annealing algorithm (AGSA). A modified method to develop the flow rate prediction model of the continuous micro-filtration (CMF) system was proposed based on improved hybrid genetic algorithm and support vector machine (SVM). A new self-adapting optimized algorithm was formed and applied to the SVM parameters. The hybrid genetic algorithm was utilized to perform variable selection, and SVM was employed to construct prediction models. The prediction models were verified by a flow rate experiment in a pilot-scale continuous micro-filtration system. Results showed that the proposed model can reveal the rule of flow rate variation in CMF. It produced a small error and exhibited strong correlation (R^2=0.91, MAE=0.0132, SSE=0.0055, RMSE=0.0155) between predicted and measured values. This result reveals that the model has strong predictability. According to the leave-one-out cross validation of training samples, the model also shows good robustness (R^2=0.89, MAE=0.0164, SSE=0.0073, RMSE=0.0178). The model developed by AGSA-SVM was compared with the model constructed by a BP neural network. The former exhibited optimal predictive capability and robustness in the comparison and is thus more suitable for the flow rate prediction of CMF.

Introduction

Continuous microfiltration (CMF) is mainly employed for product classification, concentration, separation, and purification; it has been widely utilized in the field of environmental protection, food, chemical, and others [1–3]. Its non-ideal flux changes because of membrane fouling. Thus, accurately predicting the membrane flux change rule, timely adjustment of operating conditions and operation parameters, and long-term stable operation of the membrane system are of practical significance. Given the diversity of the micro filter system and the complexity of the pollution mechanism, no universally applicable models exist to predict the flux change rule of the pollution membrane. Chinese and foreign scholars have established micro filtration flux models based on phase separation in the membrane hole shrinkage jams, surface adsorption, sedimentary characteristics, and multi-effect synergy to describe concentration polarization, membrane pore blocking, congestion, and sedimentary layer factors [4–7]. However, given the numerous parameters and complex process, relying on such models for optimization research at different operating conditions (temperature, pressure, etc.) is difficult. For this type of multi-factor,

multi-level, nonlinear complicated problem, the accuracy is not ideal when linear models, such as linear regression, time series method, and index smoothing method, are used to solve problems [8–10]. Piron [11] and Guangmin Sun [12] introduced the improved neural network method for the prediction of microfiltration flux; the method, to a certain extent, overcomes the shortcomings of traditional methods. Considering that the neural network has many defects to overcome, a general-scale membrane system experiences difficulty providing sufficient data for sample training in the short term. Hence, the application of this method to a practical range is limited. The support vector machine (SVM) method has improved generalization capability for future samples and a good prediction effect despite the lack of data; however, the system structure is not very clear [13, 14]. Therefore, this article focuses on the method to establish a forecast model of the CMF system. The hybrid genetic algorithm is improved and coupled with the SVM method to construct an adaptive optimization algorithm for SVM model parameters and improve the prediction accuracy of microfiltration flux on the surface of the membrane. The effect of changes in membrane flux prediction of sudden water pollution in short-term adjustment is analyzed to provide a scientific basis and technical methods to optimize operation conditions.

SVM

SVM is a machine learning method based on statistical learning theory [15]. In SVM, an appropriate inner product function is defined to achieve a transformation of the input space to a high-dimensional space. In this new space, the optimal linear hyperplane space is obtained. The specific principle is as follows.

For given sample data $\{x_k, y_k\} \subset R^n$, through nonlinear mapping $j(*)$, the training data are mapped into a high-dimensional feature space (Hilbert space). The nonlinear function in the input space estimation problem is translated into a linear function estimation problem in the high-dimensional feature space. The function has the form

$$f(x) = w^T j(x_k) + b, w \in R^{nh}, b \in R. \quad (1)$$

The dimension of the high-dimensional feature interval n is not fixed, and b is the offset. Based on the principle of structural risk minimization in statistics, the minimum of risk function $f(x)$ should be determined.

$$Rreg = \frac{1}{2} w^T w + CR_{emp}^{e} |f| \quad (2)$$

$w^T w = \|w\|^2$ in Equation (2) describes the complexity of function $f(*)$, e is the insensitive loss parameters, and C is the penalty factor. The introduction of non-negative slack variables x_i and x_i^* causes Equation (2) to minimize the rules risk functional equivalent optimization problem as follows:

$$\min_{w,b,x_i,x_i^*} J = \frac{1}{2} w^T w + C \sum_{i=1}^{l} \left(x_i + x_i^*\right)$$

$$s.t.\begin{cases} y_i - w^T j(x_i) - b \le e + x \\ y_i w^T j(x_i) + b - y_i \le e + x_i^* \ (i = 1,2,\cdots,l) \\ x_i, x_i^* \ge 0 . \end{cases} \quad (3)$$

The dual problem is

$$\min_{a,a_i^*} J = -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(a_i - a_i^*)(a_j - a_j^*)K(x_i, x_j) - e\sum_{i=1}^{1}(a_i + a_i^*) + \sum_{j=1}^{1} y_i(a_i - a_i^*) . \quad (4)$$

The kernel function $K(x_i, x_j)$ in Equation (4) is an arbitrary function satisfying the Mercer conditions. Parameters a and a^* are called Lagrange multipliers. To solve the problem involving parameters a and a^*, the Kuhn Tucker conditions can be utilized to obtain bias b. Then, one can obtain the output support vector machine for

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*)K(x_i, x) + b . \quad (5)$$

Given that this method is based on structural risk minimization instead of empirical risk minimization, the training is equivalent to solving quadratic programming problems. The phenomenon of learning does not appear when a small sample is trained. Thus, the model has strong generalization capability, and its solution is the global optimal solution. The problem of local extremum does not occur.

Improved Hybrid Genetic Algorithm

The genetic algorithm is a random search method that refers to natural biological selection and natural genetic mechanisms. The algorithm has strong adaptability, robustness, and global search capability. However, it has several shortcomings, such as premature convergence and local optimum [16].

The simulated annealing algorithm is a method to solve extreme programming problems by simulating the cooling process of the classical particle system in thermodynamics. It has a strong local search capability, and the search process can avoid local convergence. However, the process of finding the optimal solution requires a high initial temperature, slow cooling rate, and low end temperature; thus, the optimization process requires much time [17].

In view of the advantages and disadvantages of the two algorithms, the simulated annealing operator was embedded in the genetic operations [18] and improved selection strategy and genetic operators in this study. The characteristics of implicit parallelism genetic algorithm and simulated annealing algorithm were effectively combined for global optimization to establish the accelerating genetic and simulated annealing algorithm (AGSA). The specific process is as follows.

① The size of the population was determined. The initial temperature, crossover rate, mutation rate, and other parameters were set. A real code was used to encode each state, and the initial population was randomly generated.

② The annealing penalty factor fitness function was combined, the fitness of each individual was

calculated and decoded, and fitness evaluation was performed.
③ Genetic algorithms, such as crossover and genetic mutation manipulation, were used to optimize the initial population and produce new populations.
④ An optimal retention policy was introduced to train the new population through the simulated annealing algorithm.
⑤ After training the population, genetic selection manipulation, crossover, and mutation were applied to select and generate an excellent program group.
⑥ Determine whether to accelerate the iteration and if the termination conditions are satisfied. Return to step 1 until the conditions are met.

Model build

Based on SVM principles, the factors that have an impact on water production of the CMF system, such as temperature, pressure, and concentration, were selected as a sample input. Water production was selected as the sample output to constitute the SVM modeling sample data sets and establish the CMF water production forecasting model. Normalization method was selected to unify the sample data to [0, 1] and reduce the calculation error caused by different dimensions.

The SVM model commonly utilizes kernel functions, including polynomial kernel, RBF kernel function, sigmoid, and radial kernels. Hsu [19] reported that the radial basis function (RBF) or polynomial kernel function is highly suitable for nonlinear problems. Therefore, the RBF radial basis function was used as a kernel function in this model.

Parameter Optimization and Computing

Determination of the punishment factor, insensitive loss parameters, RBF radial basis functions, embedding dimension, and other parameters has a direct impact on the final result of prediction accuracy. Hence, these parameters have to be identified and optimized. Given the existence of the premature convergence defect of the standard genetic algorithm, in this study, it was coupled with improved AGSA with the SVM prediction model to achieve model parameter adaptive optimization while maintaining speed and efficiency in solving the model. Maintenance of both characteristics facilitates the achievement of the model solution and prevents local convergence or premature convergence [16] to achieve the desired results. The specific steps are shown in Figure 1.

Model validation

Test Equipment

A continuous microfiltration pilot system was investigated using test equipment with a daily capacity of 110 m^3, as shown in Figure 2. The main components of the system are as follows: microfiltration host, water supply system, backwash system, compressed air system, chemical cleaning system, PLC control system, and so on. The system, which contains four sets of independent control systems and membrane components, can be in parallel on the condition that four different water production membrane systems exist. Pilot tests were conducted using outside pressure type hollow fiber membrane modules at the reclaimed water workshop of the sewage treatment plant of TianJin RongCheng Iron & Steel Corp. The specific technical parameters are shown in Table 1.

Sample Acquisition, Classification, and Input

The CMF system was observed in continuous experiments under different conditions, and the samples were built separately. The changes in operation parameters (e.g., pressure) in the process flow were recorded by a single membrane module in the continuous microfiltration system. Historical load data, which were of the same type as the prediction data, were selected as training samples. Among them, the first 24 groups of data were selected as the training sample set, and the

next 24 groups of data were selected as the forecast validation sample set. The results are shown in Table 2. For comparison, while training the established model, the same experimental data were used to train the neural network (BP model).

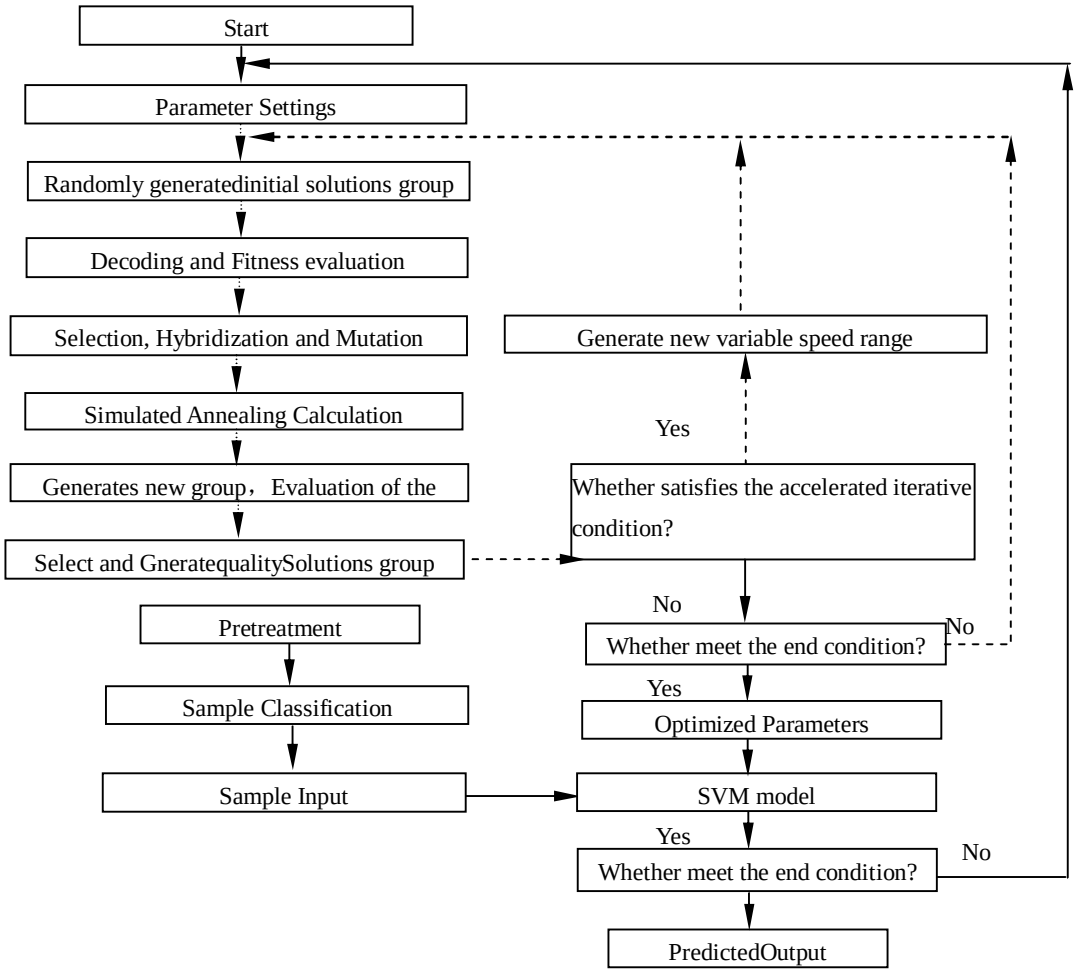

Fig.1 Flowchart of the AGSA-SVM

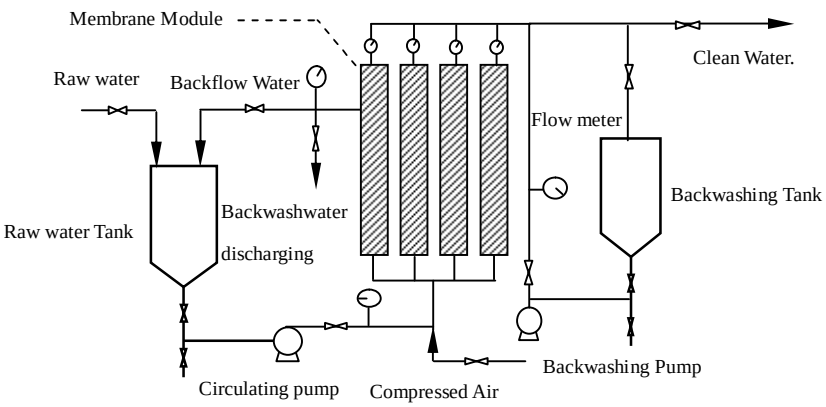

Fig.2 The Schematic of CMF

Table.1 Technical parameters of CMF

Filter form	Membrane materials	Aperture /（μm）	Filtration velocity /($m^3 \cdot h^{-1}$)	Maximum operating pressure /（MPa）	Membrane area /（m^2）	Operating temperature /（˚C）
External pressure type	Hollow fiber	0.2	1.10～1.20	≤0.15	42	5~45˚C

Result Analysis

Prediction Performance Evaluation

To test and compare the model, the accuracy and correlation of the model were analyzed by calculating R^2 (the correlation factor between predicted and measured values), mean absolute error (MAE), sum of squares due to error (SSE), and root mean square error (RMSE). The results are shown in Table 2.

Table 2 shows that the residual error between the values calculated by the BP model and the test samples, which was separately predicted using the trained BP and AGSA-SVM models, was in the range of 0.041–0.067. R^2, MAE, SSE, and RMSE were 0.71, 0.0265, 0.0216, and 0.71, respectively. The residual error between the values calculated by the AGSA-SVM model and test samples was in the range of 0.015–0.035. R^2, MAE, SSE, and RMSE were 0.91, 0.0132, 0.0055, and 0.91, respectively.

Table.2 Comparison of predicted values of testing sample

Sample number	Press /（MPa）	Temperature /（˚C）	pH	Measured values /（$m^3 \cdot d^{-1}$）	BP		AGSA-SVM	
					Predictive values / （$m^3 \cdot h^{-1}$）	Residual	Predictive values / （$m^3 \cdot h^{-1}$）	Residual
1	0.125	17	8.24	0.86	0.834	0.026	0.841	0.019
2	0.132	19	8.31	0.82	0.796	0.024	0.801	0.019
3	0.115	16	8.21	0.88	0.909	-0.029	0.89	-0.01
4	0.123	18	8.25	0.82	0.832	-0.012	0.812	0.008
5	0.126	14	8.32	0.81	0.825	-0.025	0.787	0.013
6	0.117	16	8.25	0.86	0.83	0.03	0.875	-0.015
7	0.128	15	8.24	0.94	0.883	0.057	0.927	0.013
8	0.119	18	8.26	0.74	0.711	0.029	0.734	0.006
9	0.124	13	8.29	0.86	0.871	-0.011	0.871	-0.011
10	0.118	18	8.32	0.92	0.897	0.023	0.93	-0.01
11	0.121	14	8.22	0.88	0.869	0.011	0.868	0.012
12	0.133	16	8.25	0.82	0.78	0.04	0.794	0.026
13	0.117	18	8.32	0.87	0.911	-0.041	0.876	-0.006
14	0.122	13	8.32	0.90	0.936	-0.036	0.876	0.024
15	0.121	16	8.28	0.83	0.795	0.035	0.832	-0.002
16	0.118	15	8.31	0.83	0.857	-0.027	0.826	0.004
17	0.12	16	8.25	0.82	0.753	0.067	0.785	0.035
18	0.119	18	8.27	0.78	0.768	0.012	0.789	-0.009
19	0.121	19	8.32	0.76	0.778	-0.018	0.738	0.022
20	0.123	16	8.31	0.80	0.779	0.021	0.814	-0.014
21	0.133	16	8.25	0.80	0.812	-0.012	0.812	-0.012

22	0.117	18	8.32	0.80	0.822	-0.022	0.811	-0.011
23	0.122	13	8.32	0.84	0.831	0.009	0.85	-0.01
24	0.128	15	8.24	0.86	0.842	0.018	0.855	0.005
		R^2				0.71		0.91
		MAE				0.0265		0.0132
		SSE				0.0216		0.0055
		RMSE				0.0306		0.0155

The results show that the benefits of the trained AGSA-SVM model in comparison with the BP model are as follows: the prediction values agree well with the measured values, the model has remarkable correlation and better forecast capability, the model can overcome the shortcomings of the BP neural network (easily falls into the local minimum value), reduced prediction error, and improved prediction precision and accuracy.

Model Cross Validation

The leave-one out procedural test was performed to cross validate and compare the robustness of the CMF membrane system prediction model [20]. The membrane flux of the training sample set was separately predicted by hybrid genetic SVM algorithm and BP neural network method daily. The forecast results of cross validation are shown in Table 3. The predicted and measured values of the correlation diagram are shown in Figure 3.

Table 3 shows that the residual error of cross validation by the BP model was between 0.055 and 0.039; R^2, MAE, SSE, and RMSE were 0.69, 0.029, 0.0235, and 0.69, respectively. The residual error range of the AGSA-SVM model was between 0.023 and 0.028; R^2, MAE, SSE, and RMSE were 0.89, 0.0164, 0.0073, and 0.0178, respectively. Figure 3(a) shows the BP model correlation diagram of the predicted and measured values. The visible part of the deviation is large, and the correlation between predicted and measured values is low. The error is obvious. The verification results show that based on the neural network BP algorithm of CMF water production, the robustness of the forecast model is poor. Figure 3(b) shows the AGSA-SVM model correlation diagram of the predicted and measured values. The error is small. The prediction model based on AGSA-SVM has good robustness. In conclusion, comparison of two different prediction models, namely, neural network BP algorithm and AGSA-SVM, shows that AGSA-SVM is better than the BP algorithm in terms of model robustness and is more suitable for the prediction of and research on water production of the CMF system.

Conclusion

(1) Through improved select tactics and genetic operators, AG and SA algorithms were combined to form AGSA. A modified method to develop the flow rate prediction model of the CMF system was proposed based on improved hybrid genetic algorithm and SVM. A new self-adapting optimized algorithm was formed and applied to the SVM parameters. The hybrid genetic algorithm was utilized to perform variable selection, and SVM was utilized to construct prediction models.

Table.3 Comparison of predicted valuesbycross-validation

Sample number	Temperature/（˚C）	Press /（MPa）	pH	Measured values /（ $m^3 \cdot d^{-1}$）	BP		AGSA-SVM	
					Predictive values /（ $m^3 \cdot h^{-1}$）	Residual	Predictive values /（$m^3 \cdot h^{-1}$）	Residual
1	17	0.125	8.32	0.86	0.838	0.022	0.848	0.012
2	19	0.132	8.31	0.8	0.763	0.037	0.819	-0.019
3	16	0.115	8.24	0.8	0.855	-0.055	0.773	0.027
4	18	0.123	8.32	0.76	0.732	0.028	0.772	-0.012
5	14	0.126	8.32	0.78	0.802	-0.022	0.773	0.007
6	16	0.117	8.31	0.82	0.786	0.034	0.803	0.017
7	15	0.128	8.26	0.88	0.867	0.013	0.864	0.016
8	18	0.119	8.31	0.91	0.925	-0.015	0.889	0.021
9	13	0.124	8.28	0.88	0.915	-0.035	0.893	-0.013
10	18	0.118	8.31	0.84	0.826	0.014	0.831	0.009
11	14	0.121	8.22	0.86	0.843	0.017	0.852	0.008
12	16	0.133	8.25	0.88	0.927	-0.047	0.9	-0.02
13	18	0.117	8.31	0.86	0.831	0.029	0.88	-0.02
14	13	0.122	8.21	0.78	0.754	0.026	0.769	0.011
15	16	0.121	8.22	0.82	0.793	0.027	0.792	0.028
16	15	0.118	8.25	0.84	0.851	-0.011	0.858	-0.018
17	16	0.12	8.31	0.78	0.811	-0.031	0.755	0.025
18	18	0.119	8.21	0.78	0.826	-0.046	0.796	-0.016
19	19	0.121	8.31	0.84	0.828	0.012	0.827	0.013
20	16	0.123	8.21	0.9	0.872	0.028	0.885	0.015
21	15	0.121	8.26	0.9	0.864	0.036	0.909	-0.009
22	14	0.119	8.29	0.9	0.931	-0.031	0.881	0.019
23	14	0.119	8.24	0.8	0.842	-0.042	0.823	-0.023
24	13	0.115	8.25	0.76	0.721	0.039	0.744	0.016
		R^2			0.69		0.89	
		MAE			0.0290		0.0164	
		SSE			0.0235		0.0073	
		RMSE			0.032		0.0178	

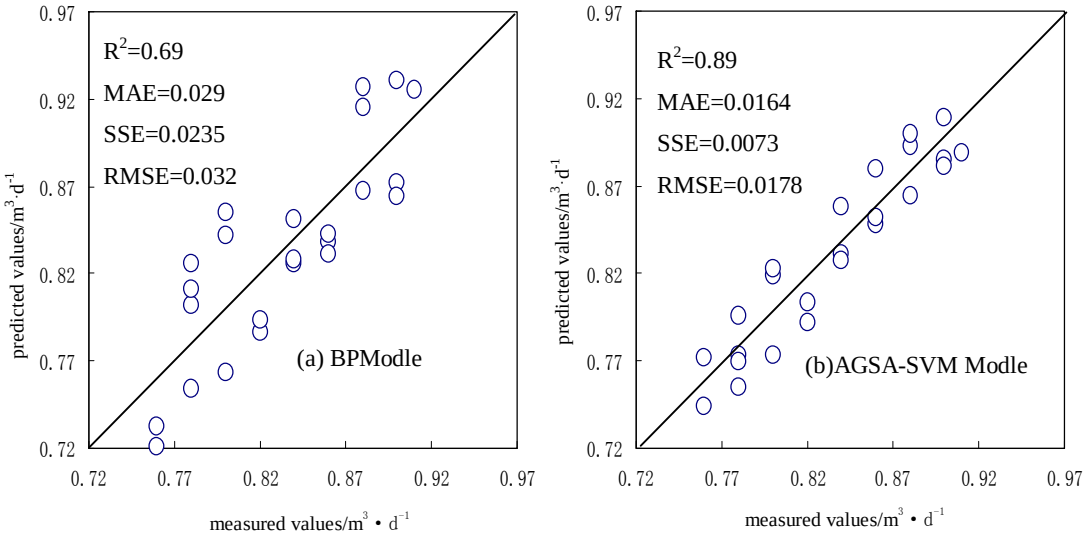

Fig.3 Plots of predicted values vs. measured values of flux by GA-SVM and BP respectivedlyin cross validation

(2) The prediction model was verified through a flow rate experiment in a pilot-scale continuous micro-filtration system. The results showed that this model can reveal the rule of flow rate variation in CMF. It had a small error and strong correlation (R^2=0.91, MAE=0.0132, SSE=0.0055, RMSE=0.0155) between predicted and measured values. The model has strong predictability. According to the leave-one-out cross validation of training samples, the model also has good robustness (R^2=0.89, MAE=0.0164, SSE=0.0073, RMSE=0.0178).

(3) The model developed by AGSA-SVM was compared with the model constructed by the BP neural network. The former showed optimal predictive capability and robustness, indicating that it is more suitable than the latter for the flow rate prediction of CMF.

Acknowledgements

Funding for this work was provided by the National Natural Science Foundation of China (Grant Nos.51208358, 51178311), and Tianjin science and technology support project(13ZCZDSF00700).

Reference

[1]Yang F,WangX,Lu X L,etal.Study on optimization of coagulation- microfiltration pilotplant. Membrane Science and Technology,2009,29(1):73-78.

[2]Su X J, Huang L J. Study on MF-RO mcombined membranes process in concentrating grosvenori infusion. Membrane Science and Technology, 2009,29(1):66-68.

[3]Wang K,Zhou C. Membrane Separation Technology (second edition) [M]. Beijing: Chemical Industry Press,2006:33-34.

[4]Sablani S S,Goosen M F,Al-Belushi R. Concentration polarization in ultra-filtration and reverse osmosis: a critical review. Desalination, 2001,141:269-289.

[5]Wei Yuan, Aleksandra K, Andrew L. Analysis of humic acid fouling during micro-filtration using a pore blockage-cake filtration model, J. Membr.Sci,2002,198:51-62.

[6]Krystyna K.Modeling of membrane filtration of natural water for potable purposes. Desalination,2002,143:123-139.

[7]Wang Z, Wu W J, Zhang X M,et al. Research progress of flux predict ion models of micro-f iltration membrane. CIESC Journal,2005,56 (6):972-978.

[8]Wang L,Zhang H W,Niu Z G. Application of supportvector machines in short-term prediction of urban water consumption.Journal of Tianjin University,2005,38(11):1021-1025
[9]El-KeibA,Ma X, Ma H. Advancement of statistical based modeling techniques for short-term load forecasting. Electric Power Systems Research,1995,35(1):51-58.
[10]Al-Kandari A M,Soliman S A, El-Hawary M E. Fuzzy short-term electric load forecasting. Electric Power Systems,2004,26(2):111-122.
[11]PrionE,LatrilleE,ReneF.Applicationof artificial neural networks for crossflow microfiltration modeling:black-box and semi-physical approaches,Comput. Chem.Eng,1997,21(2):
1021-1030.
[12]Sun G M, Zhang C H, Wang Z. Flux prediction of micro-f iltration devices based on genetic neural network. CIESC Journal,2009,60(9):2237-2242.
[13]Zhang L, Chen X H, Liu B J. SVM model of water demand prediction based on AGA. Journal of Chinahydrology,2008,28(1):38-42.
[14]Liong S Y, Sivapragasm C. Flood stage forecasting with SVM, Journal of the American Water Resources Association,2002,38(1):173-186.
[15]Vapnik V N. The Nature of statistical learning Theory[M].New York:Spring-Verlag,1999
[16]Wang D W,Wang J W,Wang H F,etal.Intelligent Optimization Methods.Beijing:Higher Education Press,2008.
[17]Xing W X, Xie J X. Modern Optimization Methods [M].Beijing:Tsinghua University Press,2005.
[18]Duan H F , Yu G P. Improved hybrid genetic algorithms for optimal scheduling model of urban water-supply system. Journal of Tongji University (Natural Science)2006,34(3):377-381.
[19]Hsu C W, Lin C J. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks, 2002(2), 13(2):415-425.
[20]Qi J, Niu J F, Wang L L. Research on QSPR for n-octanol-water partition coefficients of organic compounds based on genetic algorithms-support vector machine and genetic algorithms-radial basis function neural networks, Envirionmental Science,2008,29(1):212-218.