Water demand forecast model of Least Squares Support Vector Machine based on Particle Swarm Optimization

Kun Yan[1a], Min-Zhi Yang[2]

[1]Zhejiang Institute of Hydraulics & Estuary, No.50, Fengqi East Road, Hangzhou 310020, China.
[2]College of Hydrology and Water Resources, Hohai University, No.1, Xikang Road, Nanjing 210098, China.

Abstract: In order to solve the problem of precision of water demand forecast model, a coupled water demand forecast model of particle swarm optimization (PSO) algorithm and least squares support vector machine (LS-SVM) are proposed in this paper. A PSO-LSSVM model based on parameter optimization was constructed in a coastal area of Binhai, Jiangsu Province, and the total water demand in 2009 and 2010 were simulated and forecasted with the absolute value of the relative errors less than 2.1%. The results showed that the model had good simulation effect and strong generalization performance, and can be widely used to solve the problem of small- sample, nonlinear and high dimensional water demand forecast.

1 Introduction

Water demand forecast is an important part of water supply system optimal operation. Accurate water demand forecast can allocate limited water resources reasonably and effectively, it can not only avoid the waste of resources caused by wrong water allocation, but also ease the tension of water resources to a large extent. Because of the late start of water demand forecast in our country [1], the series length of water demand data is short and the reliability of data is low [2], and there are many factors influencing water demand forecast, such as quota method, time series method, trend analysis method and other traditional forecast methods. They not only cost a large amount of work, but also difficult to guarantee the accuracy of the forecast.

With the rapid development of information technology, many intelligent technologies based on data mining have gradually appeared, artificial neural network, support vector machine and so on have been favored by a wide range of researchers. However, because artificial neural network is a method following the principle of empirical risk minimization, the generalization performance of its model is much worse than that of support vector machine which follows the principle of structural risk minimization when dealing with small sample problems [3]. Therefore, support vector machine and its improved models are widely used to solve the problems of small sample, nonlinear and high dimensional water demand forecast.

Therefore, in order to further improve the forecast accuracy of support vector machine in the field of water demand forecast, and simplify the operation process of machine learning, a water demand forecast model based on least squares support vector machine [4] and particle swarm optimization algorithm was established in this paper. We took an area in Binhai, Jiangsu province as our study area. The model was verified by using the historical water use data and related influencing factors from 2000 to 2010 in order to provide a reference for the establishment of high precision water demand forecast model.

2 Water demand forecast model of LSSVM based on PSO

2.1 Particle Swarm Optimization

Particle Swarm Optimization (PSO) [5] is a stochastic and parallel optimization algorithm proposed by Kennedy and Eberhart in 1995 which has strong global optimization ability [6-9].

The mathematical description of PSO is as follows: suppose that in the k-th iteration of a m-dimensional space, a particle population consists of n particles: $X^k=\left\{X_1^k, X_2^k, \cdots, X_i^k, \cdots, X_n^k\right\}$. Among them, the position vector of the i-th particle can be expressed as $\boldsymbol{X}_i^k=(x_{i,1}^k, x_{i,2}^k, \cdots, x_{i,d}^k, \cdots x_{i,m}^k)$, i=1,2,……, n. $x_{i,d}^k$ represents the position of particle i in the k-th generation, d-dimensional search space. The fitness value can be obtained by substituting $\boldsymbol{X}_i^k$ into the objective function which needs to be solved.

The current particle's individual optimal value *pbest* can be expressed as $\mathbf{P}_i^k=(P_{i,1}^k, P_{i,2}^k, \cdots, P_{i,\mathrm{d}}^k \cdots P_{i,\mathrm{m}}^k)$, i=1, 2, ……, n, and there must be one best particle in the whole population, that is the group optimal value *gbest*, $\mathbf{P}_g^k=(P_{\mathrm{g},1}^k, P_{\mathrm{g},2}^k, \cdots, P_{\mathrm{g,d}}^k \cdots P_{\mathrm{g,m}}^k)$, $\mathrm{g}\in\left\{1,2,\cdots\cdots \mathrm{n}\right\}$. The velocity vector of particle i moving in the search space can be expressed as

[a] Corresponding author: 424156367@qq.com

$V_i^k=(v_{i,1}^k, v_{i,2}^k, \cdots, v_{i,d}^k, \cdots v_{i,m}^k)$, i=1, 2, ……, n. In the iterative process, the algorithm uses the following formula to update the velocity and position of the particle:

$$v_{i,d}{}^{k+1}=\omega\cdot v_{i,d}{}^{k}+c_1\cdot r_1\cdot(P_{i,d}^k-x_{i,d}^k)+c_2\cdot r_2\cdot(P_{g,d}^k-x_{i,d}{}^k) \quad (1)$$

$$x_{i,d}{}^{k+1}=x_{i,d}{}^{k}+v_{i,d}{}^{k+1} \quad (2)$$

Where k is the current iteration number. $v_{i,d}{}^{k+1}$ and $v_{i,d}{}^{k}$ represents the velocity component of the d-th dimension of particle i in the k+1-th and the k-th generation respectively. ω is the inertial coefficient, and its value is between 0 and 1. c_1 and c_2 are learning factors, usually their values are 2. r_1 and r_2 are random numbers with uniform distributions between 0 and 1. $P_{i,d}^k$ is the position of the individual extreme point of particle i in the d-dimensional search space. $P_{g,d}^k$ is the position of the global extreme point of the whole particle population in the d-dimensional search space. $x_{i,d}{}^{k}$ and $x_{i,d}{}^{k+1}$ are the positions of particle i in the d-dimensional search space before and after iterative updating respectively.

In order to limit the variation of velocity vector in each optimization process, the following constraints should be satisfied.

$$|v_{i,d}^{k+1}|\le V_{\max} \quad (3)$$

Where $V_{\max}$ is the maximum allowable limit for contemporary velocity variation.

2.2 Least Squares Support Vector Machine

Based on the traditional SVM [10, 11], the least squares support vector machine transforms inequality constraints into equality constraints, and uses square terms as the optimization index to make the calculation more convenient. n samples are mapped to high dimensional feature space by nonlinear mapping, and the optimal decision function is constructed as follows.

$$y(x)=\boldsymbol{\omega}\cdot\varphi(x)+\mathrm{b} \quad (4)$$

The loss function of LS-SVM uses the least square linear system, and its optimization problem can be transformed into:

$$\min\ \frac{1}{2}\|\boldsymbol{\omega}\|^2+\frac{1}{2}C\sum_{i=1}^{n}\xi_i^2 \quad (5)$$

$$s.t\ \ y_i(\boldsymbol{\omega}\cdot\varphi(x_i)+b)+\xi_i=1,\ \ i=1,2,\cdots\cdots n \quad (6)$$

Unlike traditional support vector machines, LS-SVM transforms the non-negative relaxation factor ξ_i into a binary norm of error. Similarly, by using Lagrange method to solve the above optimization problem, we can change it into the following quadratic programming problem.

$$\min\ L(\boldsymbol{\omega},b,\boldsymbol{\lambda})_{LS-SVM}=\frac{1}{2}\|\boldsymbol{\omega}\|^2+\frac{1}{2}C\cdot\sum_{i=1}^{n}\xi_i{}^2-\sum_{i=1}^{n}\lambda_i\cdot[y_i(\boldsymbol{\omega}\varphi(\boldsymbol{x}_i)+b)+\xi_i-1] \quad (7)$$

Where λ_i is the Lagrange operator corresponding to the i-th sample.

According to the conditions of optimization and the definition of kernel function $K(x,x')=\varphi(x)^T\varphi(x')$, the optimization problem can be further transformed into solving the following linear equations.

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & K(x_1,x_1)+1/C & \cdots & K(x_1,x_n) \\ 1 & \vdots & \ddots & \vdots \\ 1 & K(x_n,x_1) & \cdots & K(x_n,x_n)+1/C \end{pmatrix}\cdot\begin{pmatrix} b \\ \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}=\begin{pmatrix} 0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \quad (8)$$

The LS-SVM model is obtained by solving λ and b by the least square method.

$$f(x)=\sum_{i=1}^{n}\lambda_i K(x,x')+b \quad (9)$$

2.3 PSO-LSSVM

Parameter optimization

For LS-SVM, the accuracy and generalization ability of the model mainly depend on the penalty coefficient C, the kernel function and its parameters. However, there are no effective methods to select the parameters reasonably [12]. Most studies show that it is better to select RBF as the kernel function when using LS-SVM for regression estimation [13-15]. Therefore, PSO algorithm is used to optimize the penalty coefficient C and σ of RBF.

The simulation accuracy and generalization performance of the LS-SVM model are not only related to C and σ themselves, but also closely related to the relationship between them. Therefore, each parameter can not be optimized separately, and the optimization of the parameter pairs should be considered at the same time under the premise of the given fitness function.

In this paper, the *k-fold* cross validation technique [16] is used to optimize the two parameters. The data sets are randomly divided into k equal parts. Each part has the same number of non-intersecting data. For each group (C, σ), the model is trained by using (k-1) parts of them, the remaining one is used to validate. It simulates k times, and calculates the average error of k times. The average error is the cross-validation error. According to different (C, σ) combinations, different cross-validation errors are obtained. The final fitness function is the minimum value of data set cross validation error *MAPE*, see formula 10.

$$\min \ MAPE = \frac{1}{k}\sum_{i=1}^{k}\left|y_i - \overline{y}_i\right| / y_i \quad (10)$$

Where k is the number of subsets of the dataset, usually 5 or 10, and y_i and $\overline{y}_i$ are the actual value vector and the analog value vector of the i-th subset, respectively.

Model construction

The essence of using PSO-LSSVM model to forecast water demand is to solve such a regression problem.

$$y = f(x_1, x_2, \cdots x_i \cdots x_n) \quad (11)$$

Where y indicates the water demand to be forecasted, x_i is the i-th factor affecting the water demand.

Because there are many factors affecting water demand forecast, the complicated nonlinear and high dimensional problems between water demand and its influence factors can be solved by establishing a water demand forecast model by PSO-LSSVM. The model transforms this problem into a least squares support vector machine regression problem of n input variables (influencing factors) and one output (water demand).

The algorithm flow of PSO-LSSVM water demand forecast model is shown in Figure 1.

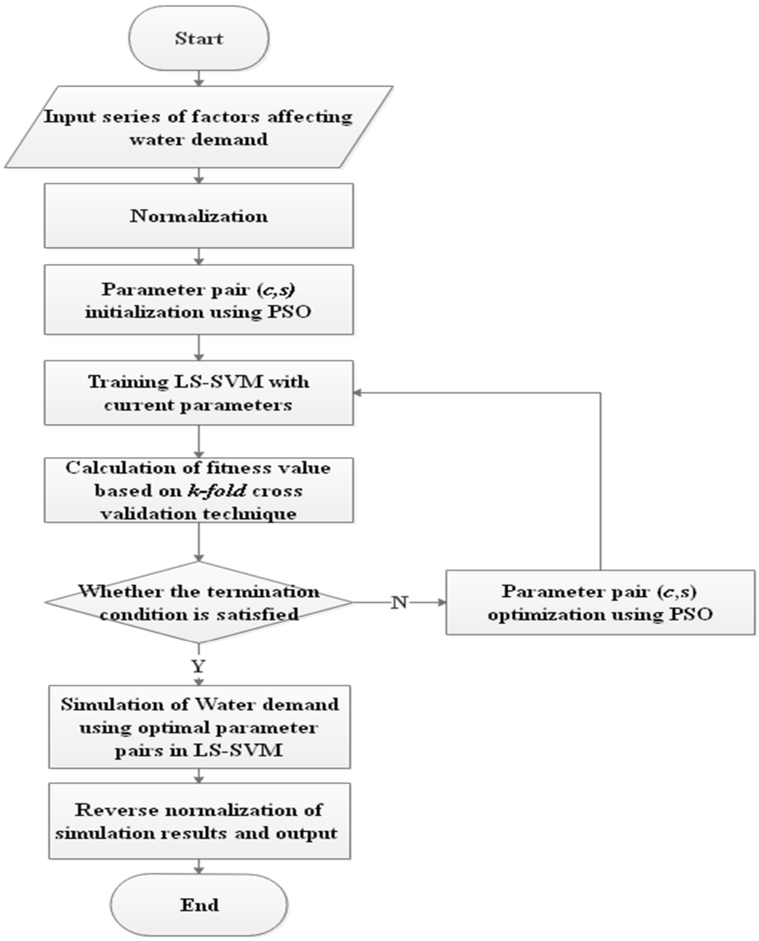

Figure 1. Flow chart of PSO-LSSVM model

3 Case study

In this paper, the PSO-LSSVM model was applied to the simulation and forecast of water demand in Binhai of Jiangsu Province. The water users were divided into four categories, namely life, industry, agriculture and ecology. The factors affecting the water demand of each type of users were different. Therefore, by using the method of sub-forecast, the influencing factors of water demand of different kinds of users were clarified, and the PSO-LSSVM water demand forecast models were established separately. The annual water consumption of all kinds of users in the study area from 2000 to 2010 were collected through the Water Resources Bulletin of Binhai in Jiangsu province, which were divided into two periods: model training period from 2000 to 2008 and verification period from 2009 to 2010. Through statistical yearbooks, comprehensive planning of water resources and other data, we collected the important influencing factors corresponding to different kinds of water users.

3.1 Domestic water demand forecast

The main factors affecting domestic water demand are the total population, per capita urban water use quota, per capita wage, leakage rate of urban water supply pipe network and popularization rate of water-saving apparatus from 2000 to 2010.

The above five factors are used as the input of PSO-LSSVM model to simulate the domestic water demand in the study area. The relative errors of the training period and the verification period of the model are shown in Table 1.

Table 1. Relative errors of simulated domestic water demand

Year	Actual value (million m^3)	Simulation value (million m^3)	Error (%)
2000	2057	2057	0.00
2001	2074	2073	-0.05
2002	2078	2082	0.19
2003	2084	2085	0.05
2004	2115	2112	-0.14
2005	2117	2117	0.00
2006	2117	2122	0.24
2007	2133	2132	-0.05
2008	2135	2138	0.14
2009	**2142**	**2149**	**0.37**
2010	**2158**	**2148**	**-0.46**

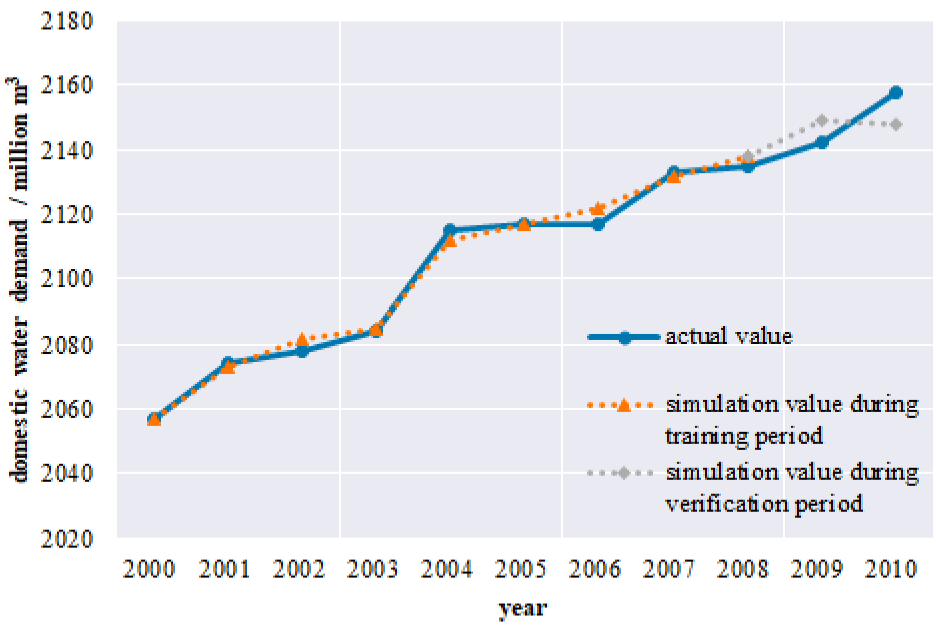

Figure 2. Fitting results of domestic water demand

From Figure 2, we can see that the simulated value of domestic water demand of the model fits well with the actual value. Combined with the results showed in Table 1, it can be seen that the absolute values of relative errors in the whole simulation period are less than 0.47% and the simulation accuracy is good. Among them, the maximum absolute value of relative error during training period and verification period are 0.24% and 0.46%, respectively.

3.2 Industrial water demand forecast

The factors affecting industrial water demand include gross industrial output value, industrial added value, reuse rate of industrial water, average consumption rate of industrial water and water use quota of per unit industrial added value from 2000 to 2010.

The above five factors are used as the input of PSO-LSSVM model to simulate the industrial water demand in the study area. The relative errors of the training period and the verification period of the model are shown in Table 2.

Table 2. Relative errors of simulated industrial water demand

Year	Actual value (million m^3)	Simulation value (million m^3)	Error (%)
2000	1121	1122	0.09
2001	1457	1454	-0.21
2002	1653	1689	2.18
2003	4076	4066	-0.25
2004	5327	5387	1.13
2005	7551	7422	-1.71
2006	7725	7733	0.10
2007	7874	7872	-0.03
2008	8021	8021	0.00
2009	**7845**	**8154**	**3.94**
2010	**7832**	**8094**	**3.35**

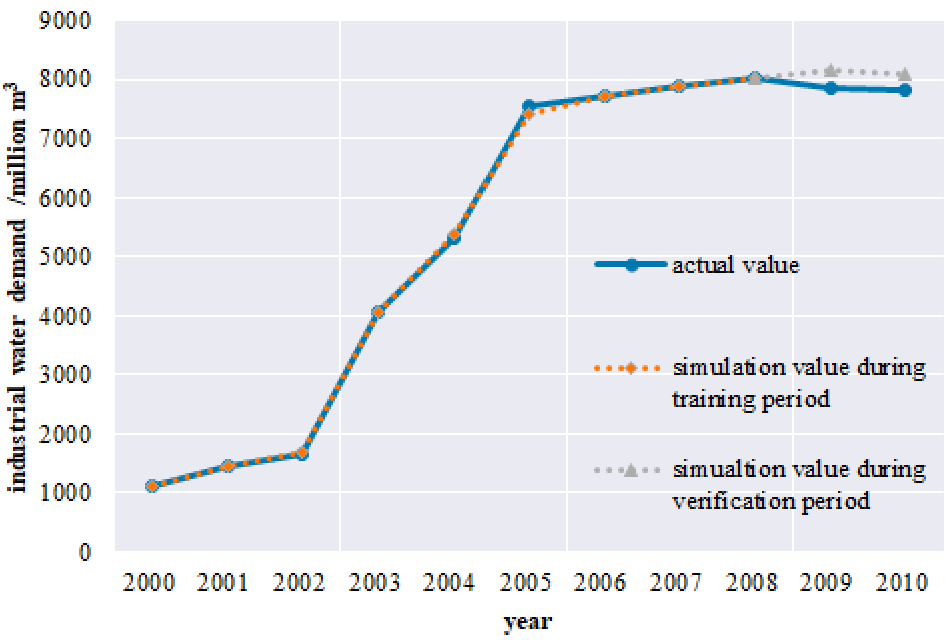

Figure 3. Fitting results of industrial water demand

From Figure 3, we can see that the simulated value of industrial water demand of the model fits well with the actual value. Combined with the results showed in Table 2, it can be seen that the absolute values of relative errors in the whole simulation period are less than 4% and the simulation accuracy is good. Among them, the maximum absolute value of relative error during training period and verification period are 2.18% and 3.94%, respectively.

3.3 Agricultural water demand

The main factors affecting agricultural water demand include precipitation, total agricultural output and planting area of crops from 2000 to 2010.

The above three factors are used as the input of PSO-LSSVM model to simulate the agricultural water demand in the study area. The relative errors of the training period and the verification period of the model are shown in Table 3.

Table 3. Relative errors of simulated agricultural water demand

Year	Actual value (million m^3)	Simulation value (million m^3)	Error (%)
2000	23480	23482	0.01
2001	24516	24517	0.00
2002	29350	29248	-0.35
2003	25001	25321	1.28
2004	28498	28497	0.00
2005	31007	31007	0.00
2006	31774	32108	1.05
2007	32402	32400	-0.01
2008	33007	33006	0.00
2009	**32268**	**32543**	**0.85**
2010	**30573**	**31163**	**1.93**

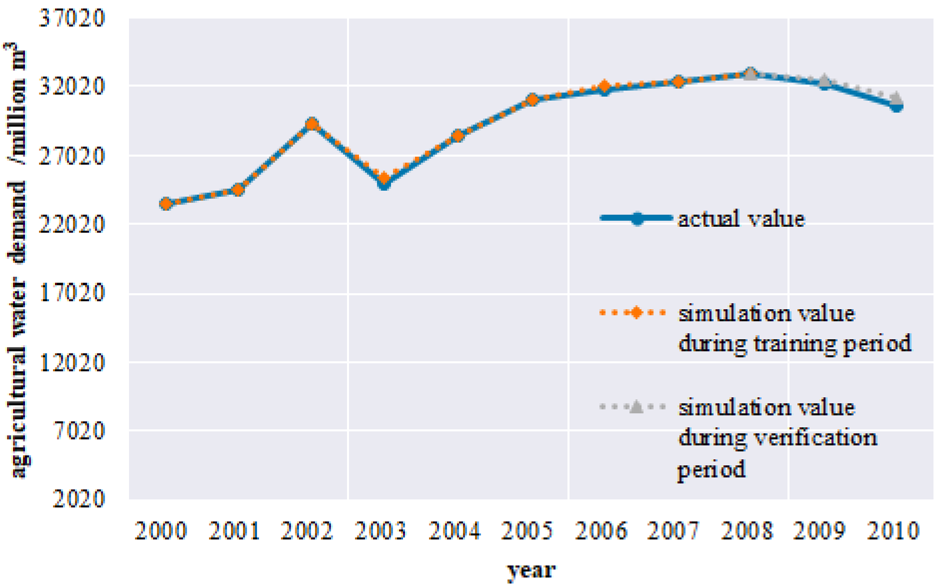

Figure 4. Fitting results of agricultural water demand

From Figure 4, we can see that the simulated value of agricultural water demand of the model fits well with the actual value. Combined with the results showed in Table 3, it can be seen that the absolute values of relative errors in the whole simulation period are less than 2% and the simulation accuracy is good. Among them, the maximum absolute value of relative error during training period and verification period are 1.28% and 1.93%, respectively.

3.4 Biological water demand forecast

The main factors affecting ecological water demand include urban green coverage, forest coverage and comprehensive index of environmental quality from 2000 to 2010.

The above three factors are used as the input of PSO-LSSVM model to simulate the ecological water demand in the study area. The relative errors of the training period and the verification period of the model are shown in Table 4.

Table 4. Relative errors of simulated ecological water demand

Year	Actual value (million m^3)	Simulation value (million m^3)	Error (%)
2000	78	77	-1.28
2001	81	81	0.00
2002	77	78	1.30
2003	83	84	1.20

2004	82	81	-1.22
2005	85	85	0.00
2006	83	83	0.00
2007	91	92	1.10
2008	93	93	0.00
2009	**89**	**91**	**2.25**
2010	**91**	**90**	**-1.10**

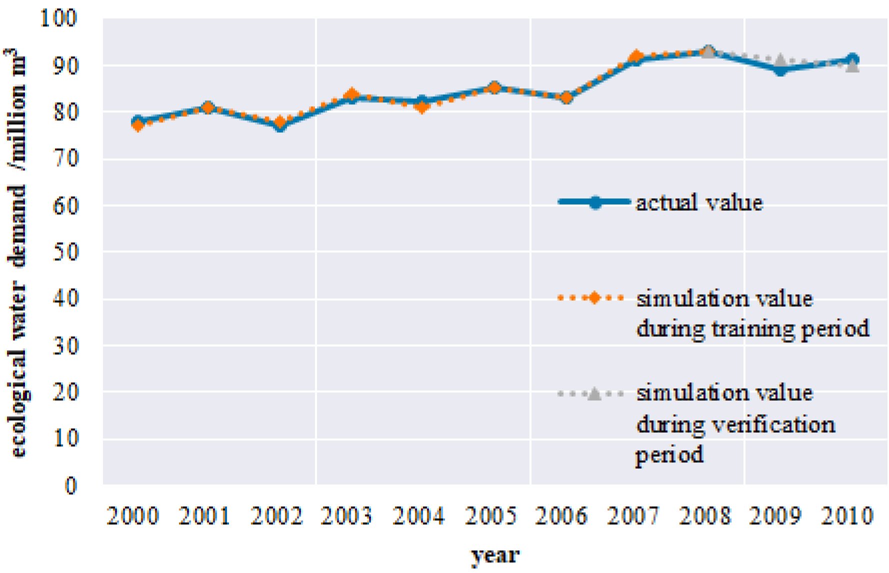

Figure 5. Fitting results of ecological water demand

From Figure 5, we can see that the simulated value of ecological water demand of the model fits well with the actual value. Combined with the results showed in Table 4, it can be seen that the absolute values of relative errors in the whole simulation period are less than 2.3% and the simulation accuracy is good. Among them, the maximum absolute value of relative error during training period and verification period are 1.30% and 2.25%, respectively.

3.5 Total water demand forecast

The total water demand in Binhai is composed of four types of users: life, industry, agriculture and ecology. The relative errors of the total water demand during training period and verification period of the model are shown in Table 5.

Table 5. Relative errors of simulated total water demand

Year	Actual value (million m^3)	Simulation value (million m^3)	Error (%)
2000	26736	26738	0.01
2001	28128	28125	-0.01
2002	33158	33097	-0.19
2003	31244	31556	1.01
2004	36022	36077	0.15
2005	40760	40631	-0.02
2006	41699	42046	0.82
2007	42500	42496	-0.01
2008	43256	43258	0.01
2009	**42344**	**42937**	**1.40**
2010	**40654**	**41495**	**2.07**

https://doi.org/10.1051/matecconf/201824601029

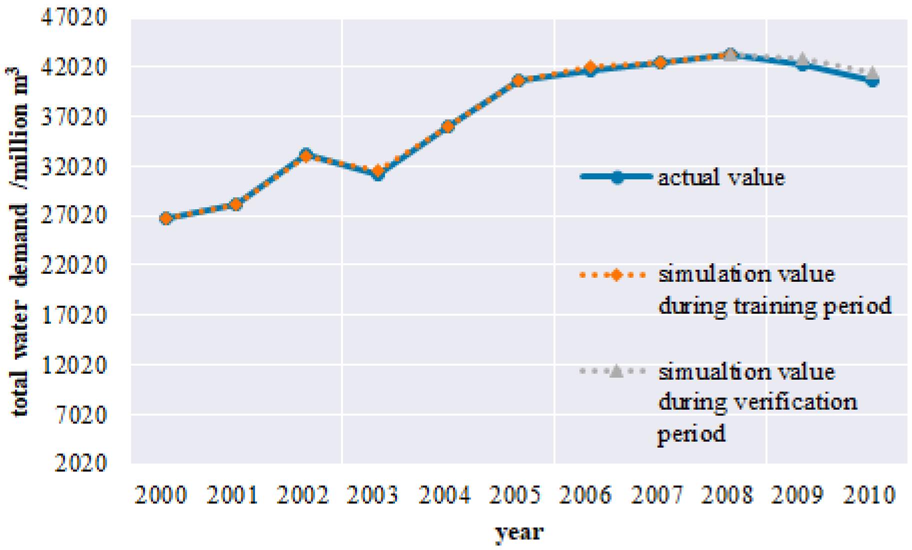

Figure 6. Fitting results of total water demand

From Figure 6, we can see that the simulated value of agricultural water demand of the model fits well with the actual value. Combined with the results showed in Table 5, it can be seen that the absolute values of relative errors in the whole simulation period are less than 2.1% and the simulation accuracy is good. Among them, the maximum absolute value of relative error during training period and verification period are 1.01% and 2.07%, respectively.

4 Conclusion

Water demand forecast is of great significance to water supply system operation and management. In this paper, PSO-LSSVM water demand forecast model is established, and a case study is carried out in an area in Binhai of Jiangsu Province. The conclusions are as follows.

(1) The PSO-LSSVM water demand forecast model can quickly find the optimization,
which is superior to the traditional method in saving the workload.

(2) The model was used to simulate and calculate the domestic, industrial, agricultural
and ecological water demand in the study area. The simulation value was in good agreement with the historical actual value, and the absolute value of the relative errors in both the training period and the verification period can be controlled within 5%. The model has high forecast precision and strong generalization ability.

(3) The example shows that the PSO-LSSVM method has good applicability and high
application value in the field of complex water demand forecast with small sample, nonlinear and high dimension.

Reference

1. Zuo, Q. T., Li, L. Prediction method of regional water demand under the condition of social and economic security [J]. Water Resources Protection, 2008,24(1):6-11.
2. Wang, Z., Yang, X. J., Liu, L. Research on water demand prediction model based on hybrid itelligent algorithm to optimize LS-SVM [C]. 2012.
3. Liu, W. L. A comparative study on several forecasting models of water demand [J]. Yangtze River, 2011,42(13):19-22.
4. Li, W. L., Li, Y. X. Hydrological prediction based on Particle Swarm Optimization Least Squares Support Vector Machine [J]. Computer Application, 2012,32(4):1188-1190.
5. Zhang, L.P. Theory and practice of Particle Swarm Optimization algorithm [D]. Zhejiang University, 2005.
6. Kennedy J, Eberhart R. Particle swarm optimization [C]. 1942-1948.
7. Yuan, X. H., Wang, C., Zhang, Y. C., et al. Application of Particle Swarm Optimization in power system [J]. Grid technology, 2004,28(19):14-19.
8. Gao, Y., Xie, S. L. Particle Swarm Optimization algorithm based on simulated annealing[J]. Computer Engineering and Application, 2004,40(1):47-50.
9. Zhang, Y. M., Jiang, S. J., Chen, R. Y., et al. Class integration test sequence determination method based on Particle Swarm Optimization algorithm [J]. Computer journal, 2018(4).
10. Gu, Y. P., Zhao, W. J., Wu. Z. S., et al. Research on algorithm of Least Squares Support Vector Machine [J]. Journal of Tsinghua university: natural science edition, 2010(7):1063-1066.
11. Suykens, Gestel J V, Brabanter T D, et al. Support Vector Machines : Least Squares Approaches and Extensions [J]. Advances in Learning Theory Methods Models & Applications Nato Science, 2003.
12. Keerthi S S. Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms [J]. Neural Networks IEEE

Transactions on, 2002,13(5):1225-1229.
13. Chapelle O, Vapnik V, Bousquet O, et al. Choosing Multiple Parameters for Support Vector Machines [J]. Machine Learning, 2001,46(1-3):131-159.
14. Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression [J]. Neural Networks the Official Journal of the International Neural Network Society, 2004,17(1):113-126.
15. Rong, H. N., Zhang, G.X., Jin, W. D. Research on support Vector Machine Kernel function and its parameters in system Identification [J]. Journal of system Simulation, 2006,18(11):3204-3208.
16. Duan K, Keerthi S S, Poo A N. Evaluation of simple performance measures for tuning SVM parameters [J]. Neurocomputing, 2003,51(2):41-59.