Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(2):38-46

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Particle swarm optimized partial least square support vector regression model for tax revenue prediction

Wu Ping

Henan Mechanical and Electrical Engineering College, Xinxiang, Henan, China

---

ABSTRACT

Chinese tax revenue is non-linear and coupled, and is influenced by many factors. Therefore, traditional forecasting methods are not sufficient to predict the value of it. In this paper, disadvantages of the existing forecasting methods are analyzed. Then partial least square support vector machine (PLS-SVR) is used to construct a tax revenue prediction model. An improved particle swarm algorithm is used to optimize the parameter set of (C, σ2), which influences the performance of this model directly. By doing so, this model can deal with the nonlinearity and multi-factors of tax revenue, and ensure stability and accuracy of support vector machine based regression. Case study on Chinese tax revenue during the last 30 years demonstrates that the optimized PLS-SVR model is much more accurate than other prediction methods.

Key words: Tax revenue prediction; PLS-SVR; PSO; Parameter set

---

INTRODUCTION

Since the reform and opening up, the economy of China has implemented growth over the past thirty years. In the meantime, the Chinese government has implemented the tax policy of various policy measures, which had significant effort to ease the economic cycle fluctuation, adjust economic structure, promote employment, price stability, and keep the development of national economy healthy. Along with the economic development and the continuous improvement of the market mechanism, the tax policy plays more and more important role in the development of China's market economy. People and the tax have more and more intimate relationship. On the one hand, the tax has effort on the resident's disposable personal income and quality of life while relates to the financial burden of enterprises and the motivation of development .On the other hand, the tax revenue is an important source of national revenue which affects the government's fiscal budget. In addition, the tax reform leads to the change of tax statistics which is new challenge to the tax work plan. It requires that the tax is not only to meet the need of government administration, but also coordinate the development and economy, so we need to base on the development characteristics of tax. This has significance for the government's tax plan, individual consumption, industry oriented and even though the whole economic development to establish a scientific, efficient and accurate prediction of tax revenue.

China's tax system is constituted by variety factors of the complex dynamic system. There are many factors affecting the tax revenue, such as the level of economic development, industrial development and so on. Many domestic and foreign scholars have focused on tax forecasting method accordingly. They have put forward many methods which play an important guiding role in the practical work. But these methods are still not satisfying as we wish. For example, multivariate regression model suitable for linear time series, but it has a lower accuracy when it is used to predict nonlinear time series [1-2].Neural network prediction method has certain requirements on the quantity and quality of samples, but also has defects of over fitting and the generalization ability is not strong[3].Greedy multilayer structure and BP algorithm of multi-level perception with BP the algorithm own their inherent defects such as it is easy to fall into local minimum values and the convergence speed is too slow[4] . The

grey theory to predict the parameters or structure model is no longer applicable, because of the relationship between China's tax and economy changing greatly, the tax system continuous reform and tax statistics data will also change [5-6]. The model which are established on the basis of traditional time series analysis method and Box-Jenkins model belong to the short-term forecasting model. With the passage of time, the forecasting result will also gradually become poor, so they are not suitable for a long period of prediction [7-9].

China's tax data has the characteristic of time sequence randomness and strong nonlinear, while the algorithm of support vector regression(SVR) construct the optimal regression function through risk minimization principle to transform the problem into solving a convex quadratic programming problem. We can use the kernel function to map the nonlinear data into high dimension space effectively, so that it can be processed by linear learning machine. However, although the thesis [10] adopts SVR algorithm, using the principal component analysis method for spatial reconstruction of index data support vector machine lost some valuable information. Although the thesis [11] adopts SVR algorithm, it cannot eliminate the influence of redundant information to predict the performance. From the algorithm of SVR, we can see that the deferent parameter selection among insensitive loss function ε, the penalty factor C and the radial basis function σ^2 will get the different support vector regression model.

Therefore, this paper combined with the basic characteristics of China's economic operation, taking advantage of the support vector machine and the particle swarm of two intelligent optimization algorithms based on the above literature. We put forward the controlling the values error ε, approximate optimize the parameters (C,σ^2) in the model of The partial least square support vector regression ($PLS-SVR$) through particle swarm algorithm, then using partial least squares support vector regression to forecast the tax revenue forecasting model system to construct a set of PSO-CV-SVR forecasting system for regression prediction to get the tax revenue forecast value. Finally, we verify the model has better prediction effect and it has guiding significance for practical production through the example analysis and comparing method, ANN , GM and $LS-SVR$.

THE BASIC MODEL OF SVR

$LS-SVR$ Expands standard SVR by optimizing the square of relaxation factors and converting the constraints of inequality to equality, so the quadratic programming problem in traditional SVR becomes linear simultaneous equations, thus the calculating difficulty reduces a lot in company with the solution high efficiency and convergence speeding up.

The basic method of SVR :

Define $x \in R^n$ and $y \in R$, let R^n be the input space, by nonlinear transformation $\phi(\cdot)$, we let in the input space x map into a high dimensional characteristic space where we use the linear function to fit sample data while making sure the generalization.

In the characteristic space, the linear estimation function is defined as:

$$y = f(x,\omega) = \omega^T \phi(x) + b \tag{1}$$

Where ω is the weight and b is the skewness.

The aim function is:

$$\min_{\omega,b,\xi} J(\omega,b,\xi) = \frac{1}{2}\omega^T\omega + \frac{1}{2}C\sum_{i=1}^{N}\xi_i^2 \tag{2}$$

s.t.

$$y_i = \phi(x_i)\omega + b + \xi_i \quad i = 1,\cdots,N \tag{3}$$

Where $\omega \in R^h$ is the weight vector and $\phi(\cdot)$ is non-linear mapping function, $\xi_i \in R^{N\times 1}$ is relaxation factor, $b \in R$ is the skewness while $C > 0$ is penalty factor.

Importing factors, $\alpha_i \in R^{N\times 1}$, we can easily get the function as:

$$L(\omega, b, \xi_i, \alpha_i) = \frac{1}{2}\|\omega\|^2 + \frac{1}{2}C\sum_{i=1}^{N}\xi_i^2 - \sum_{i=1}^{N}\alpha_i\left[\phi(x_i)\omega + b + \xi_i - y_i\right] \tag{4}$$

According to the KTT we get

$$\begin{cases} \dfrac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{N}\alpha_i\phi(x_i) = 0 \\ \dfrac{\partial L}{\partial b} = \sum_{i=1}^{N}\alpha_i = 0 \\ \dfrac{\partial L}{\partial \xi_i} = \alpha_i - C\xi_i = 0 \\ \dfrac{\partial L}{\partial \alpha_i} = \phi(x_i) + b + \xi_i - y_i = 0 \end{cases} \tag{5}$$

$$\begin{bmatrix} 0 & E^T \\ E & \phi\phi^T + C^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{6}$$

Where E is the matrix whose elements are all 1, I is a $N\times N$ identity matrix.

Inner product of regression in non-linear function can be replaced by kernel function satisfied $Mercer$. Let $\Omega_{ij} = \phi\phi^T$,

Then

$$\Omega_{ij} = \phi(x_i)^T\phi(x_j) = K(x_i, x_j) \tag{7}$$

We then have the $LS-SVR$ regression function model

$$f(x) = \sum_{i=1}^{N}\alpha_i K(x_i, x_j) + b \tag{8}$$

THE TYPES AND SELECTION OF KERNEL FUNCTION

In general, the kernel function is commonly used as the linear kernel function, polynomial kernel function, the radial basis kernel function, and sigmoid kernel function. The functions are as follows:

(1) Linear kernel function

$$K(x, x_i) = x * x_i$$

(2) Polynomial kernel function

$$K(x, x_i) = [(x * x_i) + 1]^d$$

Where the d is the order of the polynomial

(3) Radial basis kernel function

$$K(x, x_i) = \exp(-\|x - x_i\|^2 / 2\sigma^2)$$

Where the σ is width of the kernel function

(4) Sigmoid kernel function

$$K(x, x_i) = \tanh(\gamma(x * x_i) + c)$$

Commonly used kernel function can be divided into two categories: one category is the global kernel function, the other is local kernel functions: The linear kernel function, polynomial kernel function, and Sigmoid kernel function is a global common kernel function

THE PLS-SVR MODEL WITH MIXED KERNEL FUNCTION

The PLS method has been used in this research to reduce the size of the input data. Next, we will introduce the partial least squares algorithm, and use mixed kernel function and partial least square method to establish the prediction model of mixed kernel partial least squares support vector regression.

PLS Regression Model

PLS is a reasonably new Multivariate statistical method of data analysis, which is a method for constructing predictive models when the explanatory variables are many and highly collinear. Its main focus is to extract the potential components, which uses data of multiple dependent variables and independent variables for analyzing and modeling.

First, let E_0^T and F_0^T are separately the transposed matrix of E_0 and F_0, then we can obtain the eigenvector w_1 associated with the largest value of the matrix $E_0^T F_0 F_0^T E_0$, the component t_1 is:

$$w_1 = \frac{E_0^T F_0}{\left\| E_0^T F_0 \right\|}; t_1 = E_0 w_1; p_1 = \frac{E_0^T t_1}{\left\| t_1 \right\|^2}; E_1 = E_0 - t_1 p_1^T \tag{9}$$

In the same way, we can obtain the eigenvector w_h associated with the largest value of the matrix $E_0^T F_0 F_0^T E_0$, the component t_h is:

$$\begin{cases} w_h = \frac{E_{h-1}^T F_{h-1}}{\left\| E_{h-1}^T F_{h-1} \right\|}; \\ t_h = E_{h-1} w_h; \\ p_h = \frac{E_{h-1}^T t_h}{\left\| t_h \right\|^2}; \\ E_h = E_{h-1} - t_h p_h^T \end{cases} \tag{10}$$

If the rank of $X_{n\times p}$ is A, we can use cross validation method for identifying, then,

$$\begin{cases} E_0 = t_1 p_1' + \ldots + t_A p_A' \\ F_0 = t_1 r_1' + \ldots + t_A r_A' + F_A \end{cases} \tag{11}$$

Where $r_1', \ldots, r_A'$ a row vector of regression coefficient is, F_A is error matrix. Concerning the least square regression equation of F_A is

$$\hat{F}_0 = t_1 r_1 + t_2 r_2 + \ldots + t_h r_h \tag{12}$$

Because $t_1, t_2 \ldots, t_A$ can express as the linear combination of $E_{01}, E_{02} \ldots, E_{0A}$, Hence, according to the property of PLS regression:

$$t_i = E_{i-1} W_i = E_0 W_i^* \quad (i = 1, 2, \ldots, h) \tag{13}$$

Where

$$W_i^* = \prod_{k=1}^{i-1}(I - W_k P_k^T)W_i$$

Then equation (13) substitute into equation (12),

$$\begin{aligned}\hat{F}_0 &= r_1 E_0 W_1^* + r_2 E_0 W_2^* + ... + r_h E_0 W_h^* \\ &= E_0(r_1 W_1^* + r_2 W_2^* + ... + r_h W_h^*)\end{aligned} \tag{14}$$

Let $y^* = F_0$, $x_i^* = E_{0i}$, $\alpha_i = \sum_{k=1}^{h} r_k W_{ki}^* (i = 1, 2, ..., m)$

Then, equation（14）can be revert to the regression equation of standardized variable as follows

$$\hat{y}^* = \alpha_1 x_1^* + \alpha_2 x_2^* + ... + \alpha_m x_m^* \tag{15}$$

Equation (15) can be written down raw variable y, and it's PLS regression equation of estimated value $\hat{y}$ is obtained.

PLS-SVR Model

The processing of $PLS - SVR$ is divided into following steps:

Step1：PLS for feature extraction of the raw data

From compute the equation (9) and equation (10) we can contain the vector t_i, p_i and w_i.They respectively constitute the score matrix $T_{train} = [t_1, ..., t_h]$, load matrix $p = [p_1, ..., p_h]$ and correlation coefficient matrix $w = [w_1, ..., w_h]$ of training samples.

Step2：$LS - SVR$ Modeling

After h dimensions have been extracted, which can use T_{train}, y_{train} train the $LS - SVR$ model, it contain $Lagrange$ multiplier and bias term b of the optimal parameter. On this basis, the following equation can be written,

$$\begin{bmatrix} 0 & E^T \\ E & \phi\phi^T + C^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y_{train} \end{bmatrix} \tag{16}$$

Then using equation (16) can result in coefficient b and α.

Step3：PLS-SVR model prediction

Calculate the prediction value of test sample data is

$$y_{predict}(t) = \sum_{i=1}^{N} \alpha_i K_{mix}(x_i, x_j) + b \tag{17}$$

PARTICLE SWARM ALGORITHM WITH CROSS VALIDATION PARAMETER OPTIMIZATION

We can know that these three different parameters: ε of insensitive loss function, the penalty C, and σ^2 of radial basis function will get the different support vector regression model by the algorithm of SVR. Therefore this paper will dispose the parameter set (C, σ^2) through the particle swarm algorithm to optimize approximately and build the PSO-CV-PLS-SVR model for regression prediction through controlling the values error ε.

The core idea of parameter optimization in particle swarm optimization is treating the two-dimensional vector as the position of the particle and set a reasonable objective function at the same time. When each particle is searching by location, the purpose is to minimize or maximize the objective function and determine its historical best point in group or domain. At the basement of above steps, position changes.

The objective function is set to the mean square error function:
This is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{18}$$

Among them, y_i is the true value, $\hat{y}_i$ is the predicted value.

Then writing (C, σ^2) as $x = (x_1, x_2)$ .It consists of particles in groups. Then the position of particle i can be expressed as $x_i = (x_{i1}, x_{i2})$, while velocity of particle i is $v_i = (v_{i1}, v_{i2})$, its historical best point can be written as $p_i = (p_{i1}, p_{i2})$ and the whole best point can be written as $p_g = (p_{g1}, p_{g2})$ .Then the position and velocity of the particle will change with the following equation:

$$v_{ij}^{(t+1)} = wv_{ij}^{(t)} + c_1\delta_1(p_{ij}^{(t)} - x_{ij}^{(t)}) + c_2\delta_2(p_{gj}^{(t)} - x_{ij}^{(t)})$$

$$x_{ij}^{(t+1)} = x_{ij}^{(t)} + v_{ij}^{(t+1)}, \quad j = 1, 2$$

Among them, c_1 and c_2 are known as learning factors and always equal to 2. δ_1 and δ_2 are pseudo random number whose interval is $[0,1]$. w is the inertia weight, its value will influence the exploration ability and explore ability of the algorithm. We make the value of the time-varying as weights and hypothesis $w \in [w_{\min}, w_{\max}]$, $Iter\_\max$ is maximum number of iterations.

$$w_i = w_{\max} - \frac{w_{\max} - w_{\min}}{Iter\_\max} * i$$,

Among them $[w_{\min}, w_{\max}] = [0.1, 0.9]$.

Now we use the idea of cross validation to optimize the PSO-PLS-SVR model in order to find a more reasonable set of parameter (C, σ^2), so the model's error is smaller. Common CV methods is to make the whole sample set randomly divided into K groups, then making the $K-1$ groups as the training set and the rest group as a test set. After these steps we can model. When the K groups are all done a test set, construct the K models and then we use the K models with the average mean square error to select optimal parameter combination.

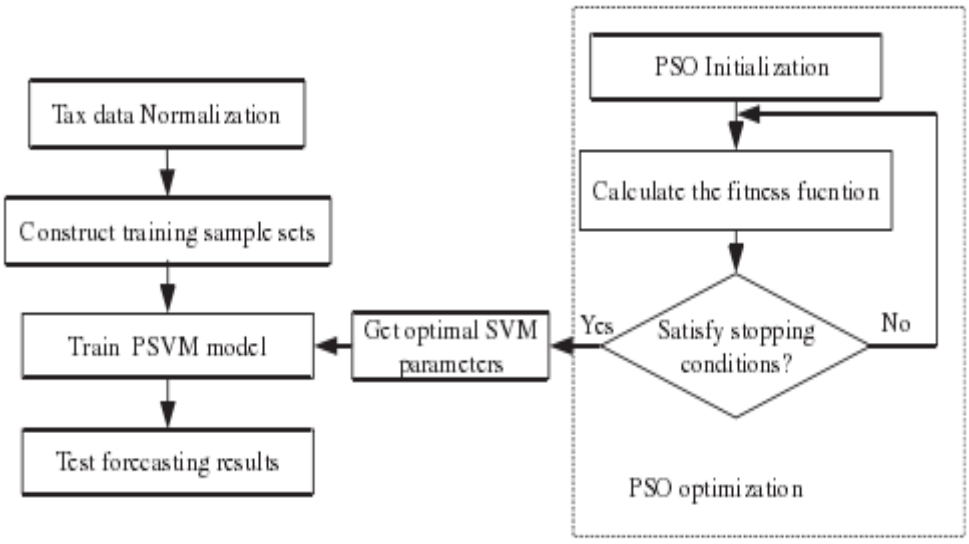

Figure 1: The construction of tax forecasting model by PSVM

In summary, the prediction accuracy of support vector machine has a great relationship with the penalty parameter C, RBF nuclear value and insensitive loss parameters. Therefore we search for the optimal combination of parameters of support vector machine in a certain region and obtain better prediction performance of support vector machine. However, PSO algorithm is to initialize a stochastic particles (stochastic solution) and then search for the optimal solution through the iterative. Particle swarm optimization algorithm for the parameter optimization process can be described in fig 1.

CASE ANALYSIS

The selection and pretreatment of training samples

The thesis uses partial least squares support vector regression system theory to optimize the relevant economic indicators of the impact of tax revenues, evaluates a variety of factors among the tax revenue and seeks a notable factor in the development of the future in order to clarify the internal relations and the development of various factors in the system Because the tax revenue forecast is related to many factors..

Table one is a tax related economic indicators of China's 1978-2008 years. According to the factors affecting, data comparable requirements, forecast model and other reasons, we choose five of the following indicators: added value of the first industry, added value of the second industry, added value of the third industry, GDP and per capita GDP. Among these indicators, there some direct impact on the three industry development status of tax revenue level indicators, such as the increasingly added values of the first, the second and the third industry; the indicators of directly or indirectly reflecting the tax size, such as GDP, per capita GDP.

Table 1. The table of related factors

Years	Tax	GDP	primary industry	secondary industry	tertiary-industry	Per capita GDP
1978	519.28	3645.2	1027.5	1745.2	872.5	381
1979	537.82	4062.6	1270.2	1913.5	878.9	419
1980	571.7	4545.6	1371.6	2192	982	463
1981	629.89	4891.6	1559.5	2255.5	1076.6	492
1982	700.02	5323.4	1777.4	2383	1163	528
1983	775.59	5962.7	1978.4	2646.2	1338.1	583
1984	947.35	7208.1	2316.1	3105.7	1786.3	695
1985	2040.79	9016	2564.4	3866.6	2585	858
1986	2090.73	10275.2	2788.7	4492.7	2993.8	963
1987	2140.36	12058.6	3233	5251.6	3574	1112
1988	2390.47	15042.8	3865.4	6587.2	4590.3	1366
1989	2727.4	16992.3	4265.9	7278	5448.4	1519
1990	2821.86	18667.8	5062	7717.4	5888.4	1644
1991	2990.17	21781.5	5342.2	9102.2	7337.1	1893
1992	3296.91	26923.5	5866.6	11699.5	9357.4	2311
1993	4255.3	35333.9	6963.8	1645434	11915.7	2998
1994	5126.88	48197.9	9572.7	22445.4	16179.8	4044
1995	6038.04	60793.7	12135.8	28679.5	19978.5	5046
1996	6909.82	71176.6	14015.4	33835	23326.2	5846
1997	8234.04	78973	14441.9	37543	26988.1	6420
1998	9262.8	84402.3	14817.6	39004.2	30580.5	6796
1999	10682.58	89677.1	14770	41033.6	33873.4	7159
2000	12581.51	99214.6	14944.7	45555.9	38714	7858
2001	15301.38	109655.2	15781.3	49512.3	44361.6	8622
2002	19636.45	120332.7	16537	53896.8	49898.9	9398
2003	20017.31	135822.8	17381.7	62436.3	56004.7	10542
2004	24165.68	159878.3	21412.7	73904.3	64561.3	12336
2005	28778.54	184937.4	22420	87595.1	74919.3	14185
2006	34804.35	216341.4	24040	103719.5	88554.9	16500
2007	45621.97	265810.3	28627	125831.4	111351.9	20169
2008	54223.79	314045.4	33702	149003.4	131340	23708

From table one, we can see that the growth of tax revenue was relatively stable from 1981 to 1984 but the tax revenue varied greatly from 1984 to 1985while the input variables changed little. This suggests that policy factors had great effect. Since 1985, tax revenue growth was relatively stable, so we abandon the sample data before 1984 and select the years from 1985 to 2008 data as a sample. In this paper, we use the data for the 1985-2008 by partial least squares support vector regression training and use CV-PSO method in the training of parameter optimization. We predict three years of tax revenue from 2009 to 2011.

Then, we give the algorithm combined with EVIEW6.0 software.

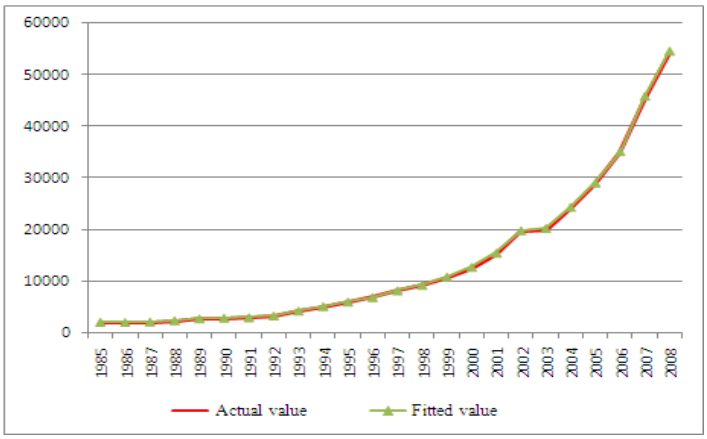

Figure 2.Schematic diagram of fitting

Error compared analysis

Choose the two of the following error index evaluation to evaluate effective prediction method in this thesis

（1）The mean absolute percentage error

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|(x_t - \hat{x}_t)/x_t\right| \tag{21}$$

（2）Mean square error percentage

$$MSPE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}[(x_t - \hat{x}_t)/x_t]^2} \tag{22}$$

x_t is the true value at the moment of t, $\hat{x}_t$ is the predicted value at the moment of t by using one prediction method. According to above-mentioned indicators, we can calculate the prediction error about the above prediction methods. We can see them from figure 3 and table 2.

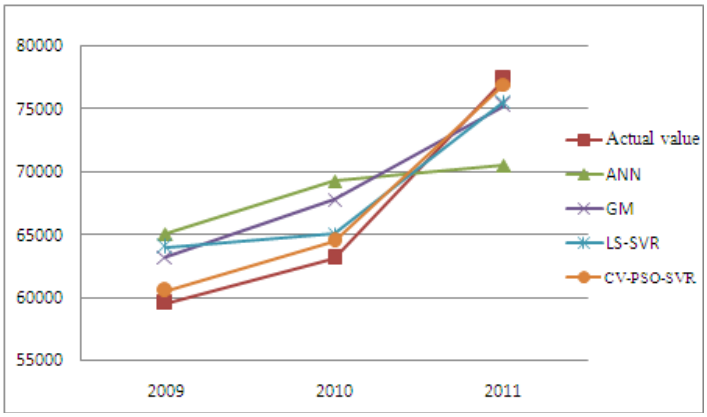

Figure 3. Comparison of the prediction error

Table.2 The comparison of the results predicted by other algorithms

Predictive index	*MAPE*	*MSPE*
Prediction method in this article	0.0104	0.0024
$LS-SVR$ model	0.0744	0.0281
GM model	0.1045	0.0416
$ARMA$ model	0.1269	0.0536

From table 2 and figure 3, we can see that the two error index prediction methods presented by this thesis are lower than other single forecasting model. This shows that the prediction method is proposed in this thesis can effectively

improve the prediction accuracy.

CONCLUSION

Taxes plays an important role in China's rapid economic growth and it is also particularly important to improve the predictability and accuracy of the tax plan. Because the tax data are complexity of highly nonlinear, coupling and multi-factor, moreover the tax forecasting model should be adapted to the change of tax economic relations, therefore the traditional prediction model is hard to meet the ideal prediction effect.

Therefore, this paper is based on the basic characteristics of China's economic operation, taking advantage of the support vector machine and the particle swarm these two intelligent optimization algorithms. We puts forward the values of control error and use the particle swarm algorithm to approximate optimization of partial set in least squares support vector regression model. After that we use partial least squares support vector regression to forecast the tax revenue, then we can establish a PSO-CV-SVR forecasting model system to regression prediction and get the tax revenue forecast value. Finally the example analysis verified the model in this thesis has better prediction effect and it has guiding significance for practical production and living.

REFERENCES

[1] Takahiko Kimura, Masaaki Ohba, Akira Shionoya. *Procedia Engineering*, 109:114-126, (**2013**)
[2] Cristina Ventura, Diogo A.R.S. Latino, Filomena Martins. *European Journal of Medicinal Chemistry*, 70:831-845, (**2013**)
[3] Guangjin Zhang, Kevin Kam Fung Yuen. *Procedia Computer Science*, 17: 441-448, (**2013**)
[4] Cristina Ventura, Diogo A.R.S. Latino, Filomena Martins. *European Journal of Medicinal Chemistry*, 11(70): 831-845. (**2013**)
[5] YananZheng, XilaiZheng, ZengwenGao, Yuxiang Zhang. *Procedia Environmental Sciences*, 18:236-242, (**2013**)
[6] Jeen Lin, Ruey-Jing Lia. *Applied Soft Computing, Volume* 13(10): 4162-4173, (**2013**)
[7] Sukhdev Singh Gangwar, Sanjay Kumar. *Communications in Nonlinear Science and Numerical Simulation*, 19,(4): 851-871,(**2014**)
[8] Hon-lun Yip, Hongqin Fan, Yat-hung Chiang. *Automation in Construction*, 38:30-38(**2014**)
[9] Zhening Zhang, JieJia, Ruifeng Ding. *Applied Mathematics and Computation*, 218(9):5580-5587, (**2012**)
[10] Fuh-YuhJu, Wei-Chiang Hong. *Applied Mathematical Modelling*, 37(23):9643-9651(**2013**)
[11] Wei-Chiang Hong, Yucheng Dong, Li-Yueh Chen, Shih-Yung Wei. *Applied Soft Computing*, 11(2):1881-1890, (**2011**)