TELKOMNIKA, Vol. 11, No. 7, July 2013, pp. 3491 ~ 3497
e-ISSN: 2087-278X

RBF Neural Networks Optimization Algorithm and Application on Tax Forecasting

YU Zhijun
Hefei University of Technology, Hefei 230009, China
Telp: +0086-15055700208, e-mail: 847672386@qq.com

Abstract

Tax plays a significant role in China's rapid economic growth. Therefore it is of particular importance to improve the predictability and accuracy of the tax plan. The tax data is characterized by being so highly nonlinear and coupling that it is difficult to be represented by using an analytical mathematical model in an accurate way.In this paper, a new optimization algorithm based on support vector machine and genetic algorithm for RBF neural network is presented. First the genetic algorithm is used to select the parameters automatically of support vector machine, and then support vector machine is used to help constructing the RBF neural network. The network on basis of this algorithm can be applied to nonlinear system identification like tax revenue forecasting. Case study on Chinese tax revenue during the last 30 years demonstrates that the network based on this algorithm is much more accurate than other prediction methods.

Keywords*: RBF neural network; SVM; parameter optimization; tax forecasting*

1. Introduction

Since the implementation of Reform and Opening Policy, China has achieved high economic growth and historic progress in taxation. Tax revenue has become a major source of government's revenue. With the further economic development and the continual improvement on the market mechanism, the tax policy will play an increasingly important role in the development of the market economy in China. On the one hand, the policy not only affects citizen's disposal income and consuming behavior, but also has an influence on the enterprise's financial burden and development potential. On the other hand, the state budget formulation, the implementation of macro-control and the expenditure of infrastructure construction are all depended on the total amount of tax revenue. Furthermore, it is the premise and only way, which makes the tax plan head for much more scientific and rational orientation, to construct a stable and accurate tax forecasting model.Therefore, in order to provide reliable information for tax planning and the decision of budget and economy, the tax authorities at all levels should strengthen the analysis of tax forecasting and set up a set of prediction system, which ,of course, needs to go hand in hand with the improvement on quality and efficiency of tax collection and management.

From different perspectives, scholars at home and abroad have conducted research on tax revenue forecasts and put forward quite a few forecasting methods. Although these methods have played an important guiding role in practical work, there are still some deficiencies. G.Duncan and other researchers have developed Bayesian forecasting model. The result shows that Bayesian forecasting model is superior to the single-variable multi-state Kalman filter method. Besides, the relative accuracy increases with the reduction of the length of the historical time series [1]; This support vector machine based on principal component analysis can eliminate the redundant information of each index and reduce the input dimensions of support vector machine, but it will lose some valuable information when making spatial reconstruction on supporting vector machine indicators data [2]. The Error Correction Model,which is based on the co-integration theory, contain only GDP and tax revenue, without considering other explanatory variables. In that case, it will influence the interpretation quality of the model [3-4]. The time series methods also have limitations. For example, the ARIMA model can not reflect the relationships between tax and economic elements. In addition, the description of simple time series methods on external factors is not clear enough [5-8]. The

Received January 13, 2013; Revised March 13, 2013; Accepted March 26, 2013

relationship between China's tax and economics changes a lot. Moreover, the continuing reform in tax rules leads to the changing of the tax statistic caliber. For this reason, the parameter or structure of the grey prediction model is no longer applicable [9-10]. Neural network prediction method requires a large number of samples, and the convergence speed of learning process is slow. There are some other defects, such as over-fitting, less strong capability of the model generalization and so on. The neural network training and learning are based on the causal relationship between dependent variables and independent variables which is implied in the samples. What's more, this learning cannot reflect the change of external factors and its effects on the prediction. When the environment of prediction objects changes, the prediction accuracy will be greatly reduced [11]. Rough Set Theory is a kind of more valid method to deal with the complicated systems. It can seek for the relationship between tax revenue and influencing factors by removing the redundancy information from the data directly. However, the consistency of sample classification and data characterization are closely related to the calculating speed of the method and the prediction accuracy, whether the method can guarantee higher predictive accuracy is great deal of uncertainty [12].

On the basis of the literatures above, considering the basic characteristics of China's economic operation, a new optimization algorithm based on support vector machine and genetic algorithm for RBF neural network is presented. In which genetic algorithm is chosen to select the parameters automatically for support vector machine. and then the support vector machine is used to help construct RBF neural network. According to which, a forecasting model of tax revenue is set up. This algorithm can avoid not only the shortcomings of traditional algorithm which is easy to get local minimal value, but also a large number of experiments or experiences which are needed to pre-specify network structure. In the last part of this paper, these algorithms like $ARMA$, GM, $LS-SVR$ are compared to verify the rationality and effectiveness of the method. The results show that the forecasting model based on this method can obtain better performance in tax forecasting than the other models. So it can be used as a new approach for tax forecasting.

2. Research Method

2.1. SVM Providing the Oretical Foudation for Structure and Parameters of RBF

RBF network, which is a three-layered feed forward network, maps directly input vector onto the hidden layer space by using radial basis function. The output of the network is the linear weighted sum of hidden unit's output.

The mapping of RBF network from input columns to output columns is nonlinear, while the output of the network is linear in terms of the weights, and the output of the k^{th} hidden unit is

$$\phi_k(x) = \exp(-\frac{\|x - c_k\|^2}{2\sigma_k^2}) \quad (1)$$

Where $\|\bullet\|$ is Euclidean norm, σ_k is the width of hidden layer nodes. c_k is the center of the hidden layer nodes, X_i is the i^{th} input variable. N denotes the number of the hidden units, W_k is the connection weights between output and the hidden layer nodes, then the output of RBF networks is

$$f(x) = \sum_{k=1}^{N} w_k \exp(-\frac{\|x - c_k\|^2}{2\sigma_k^2}) \quad (2)$$

In accordance with Mercer Conditions, kernel function is used to map the sample in the original space to a vector in high-dimensional feature space. The application of Gaussian kernel function used here is as follows.

$$K(x, v_i) = \exp(-\frac{\|x - v_i\|^2}{2\sigma_i^2}) \quad (3)$$

The number of the hidden units of SVM is the number of Support Vector, g represents the number of support vector. W_i is the i^{th} weights between the hidden units and the outputs. Z_i denotes Support Vector. b is the bias. SVM is the linear combination of the hidden units, then

$$f(x) = \sum_{i=1}^{g} w_i K(x, v_i) + b = \sum_{i=1}^{g} w_i \otimes \exp(-\frac{\|x - v_i\|^2}{2\sigma_i^2}) + b \quad (4)$$

The principles of RBF network and SVM construct RBF kernel space are different, however, they are comparable, there is a one-to-one correspondence among network parameters, and the output of the network are the linear weighted sum of hidden layer nodes' output.

2.2. GA Providing SVM Models Parameters

The genetic algorithm is a kind of stochastic optimization method, which can simulate natural selection and genetic variation in the process of biological evolution. It has a strong capability of global search, and this ability is not dependent on a specific solution model. This algorithm can provide an effective way to solve the selection of support vector machine model parameters. This algorithm is applied to optimize support vector machine model parameters, including the parameter σ, penalty factor C and insensitive loss function ε, and the basic steps of this algorithm are as follows:

Step 1: Select the initial population of every individual randomly;

Step 2: Evaluate the fitness function value of each individual;

Step 3: Choose a new generation of population from the prior generation by using the method based on selection operator;

Step 4: Take the evaluation, selection, crossover and mutation operation on the new population after the crossover operation and mutation operation are used on the current population, and continue.

Step 5: If the fitness function value of optimal individual is large enough or the algorithm has run sequentially a lot of generations, and the optimal fitness value of the individual can't be improved perceptibly, then we obtain insensitive loss function ε, penalty factor C, and the optimal value of kernel function parameter σ, and we can also get the optimal classifier by using the training data sets.

The genetic algorithm optimizes the fitness fuction directly, and the selection of SVM model is to optimize minimized generalization performance indicators. Therefore, it is necessary to transform the minimized generalization index to the maximum fitness function.

$$fit = 1/(T + 0.05) \quad (5)$$

Where $T = R^2/l\gamma^2$ is the testing error bound, $\gamma = 1/|w|$ is the interval value, l is the number of the samples.

After determining the fitness function, we can follow the above steps of the genetic algorithm to search more optimal model parameters for support vector regression machine, and then a support vector regression machine is obtained from the study of training samples.

2.3. SVM Providing Network Structure and Parameters for RBF

The learning of Support vector regression machine process is a quadratic programming problem with linear restrictions, and the regression machine that has been trained well could be used to determine the structure and parameters of the RBF network. In consideration of the linear regression situation, we give the following samples $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in R^n \times R$, and set the linear function to be $f(x) = w \cdot x + b$. So the optimization problem is to minimize the following function

$$R(w, \xi, \xi^*) = \frac{1}{2} w \cdot w + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^*\right) \quad (6)$$

$$s.t.\begin{cases} f(x_i) - y_i \le \xi_i^* + \varepsilon, i = 1,\ldots,n \\ y_i - f(x_i) \le \xi_i + \varepsilon, i = 1,\ldots,n \\ \xi_i, \xi_i^* \ge 0, i = 1,\ldots,n \end{cases} \tag{7}$$

We solve the parameters by using Lagrange function and get the maximum function

$$W(\alpha_i, \alpha_i^*) = -\frac{1}{2}\sum_{i=1,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \times (x_i \cdot x_j) - \varepsilon\sum_{i=1}^{n}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) \tag{8}$$

$$s.t.\begin{cases} \sum_{i=1}^{n}(\alpha_j - \alpha_j^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C, i = 1,\ldots,n \end{cases} \tag{9}$$

We solve this quadratic optimization problem to get α_i, α_j^*, then $w = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)x_i$ we can also get b by

$$\begin{cases} b = y_i - w \cdot x_i - \varepsilon \\ b = y_i - w \cdot x_i + \varepsilon \end{cases} \tag{10}$$

Thus, the regression function is

$$f(x) = (w \cdot x) + b = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)(x_i \cdot x) + b \tag{11}$$

Considering it is a non-linear regression, we use Gaussian function as the kernel function $K(x,y)$, then the regression function will be

$$f(x) = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)K(x_i, x) + b \tag{12}$$

Assuming that the number of support vectors obtained from SVM training is g $(g \le n)$, support vectors are $v_i, i = 1,\ldots,g$, the offset coefficient is b. The weight values are $w_i = \alpha_i - \alpha_i^*, i = 1,\ldots,g$, we can construct the RBF neural network by using these parameters, where Gaussian function is used as the kernel function, the number of input nodes is the dimensions of input matrices, the number of hidden units is g, the number of output node is 1, which is the same as the SVM. The radial basis function centers are $v_i, i = 1,\ldots,g$, the offset coefficient is b ,the weight values are $w_i = \alpha_i - \alpha_i^*, i = 1,\ldots,g$. Since SVM training is to solve the quadratic optimization problem and it is characterized by being highly learning efficient, global optimized, RBF network constructed based on SVM can have better performance. The flow chart of the proposed method is clearly shown in Figure 1.

3. Results and Analysis

3.1. Selection of Trained Sample Data and Index

In this paper, we select the relevant economic data from 1980 to 2011. The data from 1980 to 2008 is used as training samples, and the rest is used as test samples. The data of tax revenue is set as characteristic sequences X0, and then we select the following 7 indexes as relevant factors sequence to make an analysis according to the size of the influencing factors, the comparability of information and the requirements of prediction model.

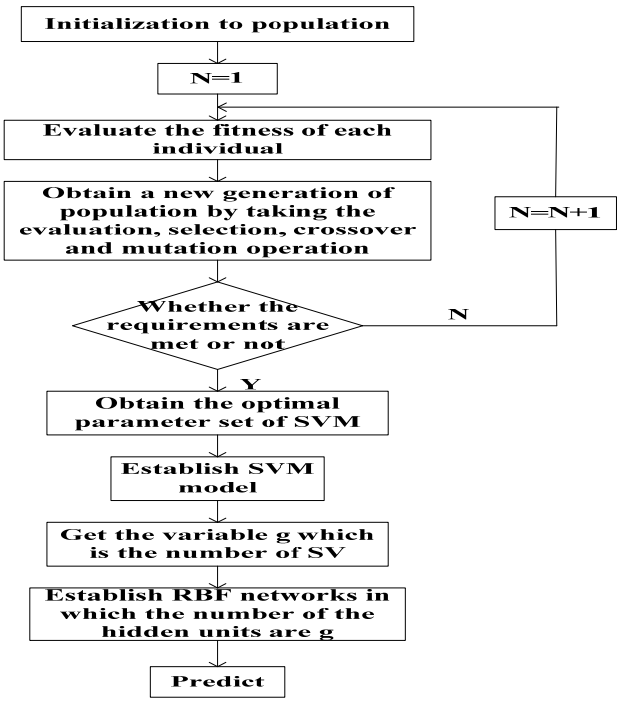

Figure 1. Proposed method's flow chart

Among these indexes, one of them is the three industries index, which have a direct impact on the level of tax revenue, such as the value-added of the first industry (X1), the value-added of the secondary industry (X2), and the value-added of the tertiary industry (X3). And there are indexes which can show the size of tax revenue directly or indirectly, including fixed asset investment across the country (X4) and total volume of foreign trade (X5). And some can show the people's living standards, and have a direct impact on the status of the total tax revenue growth, like total retail sales of social consumption goods (X6). And there is an index which can reflect the relationship between revenue growth and economic development, namely the rural and urban residents' deposit balance (X7).

3.2. Identification Results

The input layer node number and the output layer node number are 7 and 1, so we set the number of hidden units in RBF network is 6. The Gaussian function center vector are the support vectors, the width of Gaussian function is the same with the regression machine. According to $w_i = \alpha_i - \alpha_i^*, i = 1, \ldots, g$, we know the corresponding weights w = [-0.0093 0.0582 0.1209 0.2860 0.9892 11.1850]. Through standardization of data processing, the identification results are shown in Figure 2.

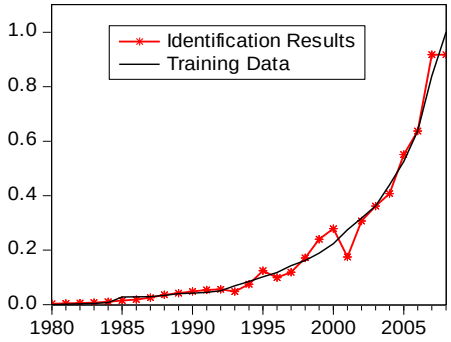

Figure 2. Identification results

3.3. Error Comparison Analysis

The validity of the forecasting method can be evaluated by using the following two error indicators.

(1) Mean absolute percentage error

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|(x_t - \hat{x}_t)/x_t\right| \quad (13)$$

(2) Mean square prediction error

$$MSPE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}[(x_t - \hat{x}_t)/x_t]^2} \quad (14)$$

Where in x_t is the actual value of x at the time of t, $\hat{x}_t$ shall be it's prediction value. Comparison of the predicted results is shown in Figure 3,and table 1 shows comparison of the forecasting errors according to the above two indicators.

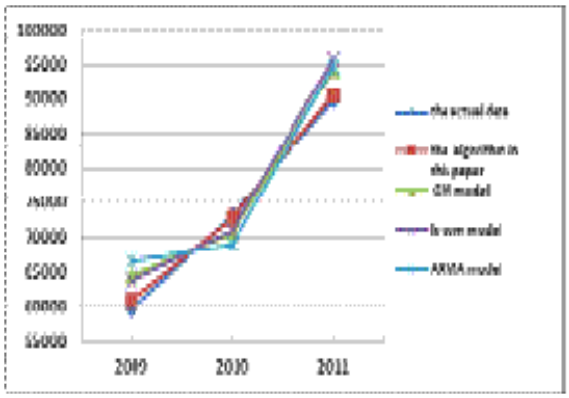

Figure 3. Comparison of the predicted results

Table 1. Comparison of the forecasting errors

The evaluation standards of forecasting performance	$MAPE$	$MSPE$
The algorithm in this paper	0.0122	0.0080
GM model	0.0557	0.0338
$LS-SVR$ model	0.0548	0.0332
$ARMA$ model	0.0767	0.0473

The results shown in Figure 3 and table 1 demonstrate that the two error index of this prediction method are lower than other single forecasting models, and it indicates that the prediction method proposed in this paper can effectively improve the prediction accuracy.

4. Conclusion

Accurate tax revenue forecasting becomes a most important management goal. However, tax revenue, in the socio-economic system, is subject to quite a few uncertain qualitative and quantitative factors, it often presents nonlinear data patterns. And for exactly that reason, it is difficult to be represented by using an analytical mathematical model in an accurate way. Therefore, a rigid forecasting approach with strong general nonlinear mapping capabilities is essential.

The algorithm proposed in this paper first use genetic algorithm to optimize the model parameters of support vector regression machine, and then this regression machine supplies RBF neural network with a superior structure and parameters. A forecasting model of tax revenue in accordance with RBF neural network is put forward, aiming at the problem of tax revenue forecast. Compared with the traditional method of tax revenue forecast, it avoids the disadvantage of traditional algorithms which are often trapped to local minimathis, what's more, this method effectively improves generalization and don't need a large number of experiments or empirical experiences to pre-specify network structure. According to case study,the method has higher precision,good generalization ability and classification ability.

References

[1] Uncang, Gorrw, Szczypulaj. Bayesian forecasting for seemingly unrelated time series: application to local government revenue forecasting .*Management Science*. 1993; 39(3): 275-293.
[2] Zhang Yu, Yin Teng–fei. Study on Tax Forecasting Base on Principal Component Analysis and Support Vector Machine (in Chinese). *Computer Simulation*. 2011; 9(28): 357-360.
[3] Zhang Shao-qiu. Research on tax prediction error correction model based on co-integration theory [J] (in Chinese). *Journal of south china normal university (Natural science edtion)*. 2006; (1): 9-14.
[4] Xue Wei, Zhang Man. Granger Cause and Co-integration Test on China's Tax Revenue (in Chinese). *Journal of Central University of Finance & Economics*. 2005; (11): 6-19.
[5] Zhang Xin-bo. The application of time series model in tax forecasting (in Chinese). *Journal of Hunan Tax College*. 2010; 8(23): 30-32.
[6] Zhang Meg-yao, Cui Jine-huan. Study on monthly central tax revenue forecasting models based on time series method. *J. Sys. Sci & Math. Scis*. 28 (11): 1383 -1390.
[7] Shang Kai, Zhang Zhi-hui. The Economic Factor Analysis of Impacts on Changes of China's Tax Revenue Growth Rate (in Chinese). *Economy and Management*. 2008; 7(22): 15-19.
[8] Mocanhn, Azads. Accuracy and rationality of state general fund revenue forecasts: evidence from pane data. *International Journal of Forecast*ing. 1995; (11): 417-427.
[9] Yu Qun, Li Wei-min, SHEN Mao-xing. Application of Gray Sequence Prediction in Tax Forecasting (in Chinese). J*ournal of System Simulation*. 2006; 8(18): 971-972.
[10] Sun Zhi-yong. A Study on Tax Forecasting Model Based on Grey Theory (in Chinese). *Journal of Chongqing University (Social Science Edition)*. 2010; (16): 41-45.
[11] Zhang Shao-qiu, Hu Yue-ming. Taxation Forecasting Model Based on BP Neural Network (in Chinese). *Journal of South China University of Technology (Natural Science Edition)*. 2006; 34(6): 55-58.
[12] Liu Yun-zhong , Xuan Hui-yu, Lin Guo-xi. Application Research on Tax Forecasting in China Based on Rough Set Theory (in Chinese). *Systems Engineering Theory & Practice*. 2004; 10: 98-103.
[13] K ayathri, N Kumarappan. Accurate fault location on EHV lines using both RBF based support vector machine and SCALCG based neural network. *Expert System with Applications*. 2010; (37): 8822-8830.
[14] Wang Ling-zhi, Wu Jian-sheng. *Application of Hybrid RBF Neural Network Ensemble Model Based on Wavelet Support Vector Machine Regression in Rainfall Time Series Forecasting*. Proceedings of the 2012 5th International Joint Conference on Computational Sciences and Optimization, CSO 2012: 867-871.
[15] Olej, Vladimír. Filipová, Jana. Modelling of Web Domain Visits by Radial Basis Function Neural Networks and Support Vector Machine Regression. *IFIP Advances in Information and Communication Technology.* 2011; 364(2): 229-239.
[16] Olej, Vladimír. Filipová, Jana. *Short time series of website visits prediction by RBF neural networks and support vector machine regression*. Lecture Notes in Computer Science (including sub series Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2012; 7267(1): 135-142.
[17] Ren Jin-xia, Yang Sai. *RBF Neural Networks Optimization Algorithm Based on Support Vector Machine and Its Application*. 2nd International Conference on Information Engineering and Computer Science Proceedings, ICIECS 2010.
[18] Wang Bing, Wang Xiaoli. Perception Neural Networks for Active Noise Control Systems. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2012; 10(7): 1815-1822.
[19] Patricia Melin, Victor Herrera, Danniela Romero, Fevrier Valdez, Oscar Castillo. Genetic Optimization of Neural Networks for Person Recognition Based on the Iris. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2012; 10(2): 309-320.