Modeling for Nonlinear Series Prediction based on the Support Vector Machine Theory

LIU Dao-wen
Xuchang University
Public Experiment Center
Xuchang Henna province China 461000
ldw_xc@126.com

***Abstract*—In order to improve the prediction accuracy, applies the Support Vector Machine (SVM) theory to the prediction of the Nonlinear Series. Based on the analysis of the basic theory for the prediction, adopts the Cross Validation method to choose the best parameters and then establishes the prediction model. For the stock index of Shanghai Stock Exchange, carries out the prediction to verify the effect of the model. Proved by the research, the method based on the Support Vector Machine theory is able to reflect the changing tendencies, and has the better prediction accuracy, at the same time the feasibility is verified by the method.**

Keywords- Nonlinear time series; Support Vector Machine; best parameters; prediction model

I. INTRODUCTION

The time series obtained from the dynamic system implies the evolution law of the system, and some implicit system characteristics are able to be found by the research and analysis on the time series, and then the evolution tendency of the system or the prediction of the system will come true by enough using the implied characteristics. Some characteristic quantity of the dynamic system is obviously nonlinear because they are influenced by many factors. At a certain period, the system has strong randomness, but from point of macroscopic view, the system has certain deterministic and regularity [1]. Modeling for the nonlinear time series to forecast the system change tendency is useful to grasp the change law and is helpful to make correct decision in the concrete fields, so it is of quite importance in the scientific research [2].

The researchers have been devoted to finding the effective method to forecast the nonlinear time series. Regression forecasting method [3] has a high precision, and is suitable to the long term prediction, but it have some obvious disadvantage, such as strict requirement for the history data, difficulties of the regression variable decision, lack of the self-learning ability. Time series forecasting method [4] has the advantage that amount of calculating work is small and suitable to the short term prediction, but it is unable to effectively deal with the regularity. Artificial neural network with self-learning and self-adaptive has strong ability of nonlinear mapping, but the method also has some disadvantage, such as slow convergence speed, easily getting into local minimum, difficult to determine the number of implicit layers [5]. The method based on the chaos theory for the time series prediction is a view on the dynamic system prediction from itself evolution regularity, without considering the complex influence factors, but it is lack of the strict theory base and requires huge history data. Support Vector Machine (SVM) is a new general machine learning method proposed by Vapnik based on the statistics learning method [7]. For the given data sample, Support Vector Machine realizes structural risk minimization of the data sample from the approximation accuracy and the approximating function. Support Vector Machine has perfect theory in solving many real problems, and is able to construct function in many kind of function set [8]. And the method is widely applied to the speech recognition, pattern recognition, analysis of time series, bioinformatics and economics, achieves some results in the aforesaid fields [9].

Natural Science Foundation of Henan Educational Committee (No. 12B510028).

II. PREDICTION THEORY OF SUPPORT VECTOR MACHINE

Suppose the training nonlinear time series sample is $\{(x_1, y_1),(x_2, y_2),\Lambda ,(x_k, y_k)\}$ $(x_i \in R^n,\ y_i \in R,\ i = 1,2,\Lambda\ k)$, k is number of training sample data. The basic idea of Support Vector Machine for time series prediction is a nonlinear mapping Φ that transfer time series to the high dimension feature space F, and then construct the optimized linear regression function in the high feature space, and the expression of the linear regression function as follows:

$$f(x) = w\phi(x) + b \quad (1)$$

In the aforesaid expression, w and $\phi(x)$ are both *m*-dimension vector, and b is the offset value. Support Vector Machine adopts the structural risk minimization principle to determine the values of w and b, namely[10]

$$\min R_{str} = \frac{1}{2}\|w\|^2 + CR_{emp} \quad (2)$$

In the expression (2), $\|w\|^2$ is the complexity of control mode. C is the weight which is used to control the punishment degree that exceeds the error sample. $R_{emp} = \frac{1}{k}\sum_{i=1}^{k} L_{\varepsilon}[x_i, y_i - f(x_i)]$ is the error control function,

which is usually measured by the $\varepsilon-$ insensitive loss function, and the insensitive loss function is defined as follows:

$$L_\varepsilon = \begin{cases} |y - f(x))| - \varepsilon & |y - f(x))| \geq \varepsilon \\ 0 & |y - f(x))| < \varepsilon \end{cases} \quad (3)$$

According to the structural risk minimization principle, considering complexity of the regression model obtained from the training set, regression based on the Support Vector Machine essentially is a solution of an optimized question, and the optimized question are in the following[11].

$$\min \quad \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{k}(\xi_i + \xi_i^*)$$

$$s.t. \quad \begin{cases} [w\phi(x_i)] + b - y_i \leq \varepsilon + \xi_i \\ y_i - [w\phi(x_i)] - b \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad (4)$$

In the question, ξ_i and ξ_i^* are slack variable，and the question is described as the origin question of the support vector machine. For the value of dimension w is huge, in order to conveniently solve the question, introduces the Lagrange multipliers α_i and α_i^* according to the duality theorem, and establishes a Lagrange function, then the optimized question is converted to the dual space, and acquires the dual question of the origin question, the formula is shown as the (5) expression.

$$\min \quad \frac{1}{2}\sum_{i,j=1}^{k}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) + \varepsilon\sum_{i=1}^{k}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{k} y_i(\alpha_i^* - \alpha_i)$$

$$s.t. \quad \begin{cases} \sum_{i=1}^{k}(\alpha_i^* - \alpha_i) = 0 \\ 0 \leq \alpha_i^*, \alpha_i \leq C(I = 1,2\Lambda\ k) \end{cases} \quad (5)$$

In the expression, $K(x_i, x_j) = [\phi(x_i) \bullet \phi(x_j)]$ is the kernel function, and the most commonly used optimized kernel function is the Gauss function, and the concrete formula is $K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2})$. Supposed the solution to the dual question of the origin question is $\overline{\alpha} = (\overline{\alpha_1}, \overline{\alpha_1^*}, \Lambda\ \overline{\alpha_i}, \overline{\alpha_i^*}\ \Lambda\ , \overline{\alpha_k}, \overline{\alpha_k^*})^T$, consequently, the regression function is expressed as follows[12].

$$f(x) = \sum_{i=1}^{k}(\overline{\alpha_{i1}^*} - \overline{\alpha_i})K(x_i, x) + b \quad (6)$$

III. PREDICTION MODEL

A. Data preprocessing

For the convenient analysis and computing, the origin time series which consists of 4736 data firstly normalizes. Supposed the origin time series is $\{y_i, i = 1,\Lambda\ , N, N = 4736\}$, then the series is normalized by the formula (7), and the normalized series is $\{x_i, i = 1,2,\Lambda\ , N, N = 4736\}$.

$$x_i = \frac{y_i - \frac{1}{N}\sum_{i=1}^{N} y_i}{y_{\max} - y_{\min}} \quad (7)$$

In the expression, $y_{\max} = \max\{y_i, i = 1,2,\Lambda\ ,4736\}$,

$y_{\min} = \min\{y_i, i = 1,2,\Lambda\ ,4736\}$.

B. The best regression parameters

The paper adopts the cross valid method to determine the best regression parameters, lets the punishment parameter c and kernel function parameter g value in a certain interval. For the given parameters c and g, and uses the K-CV method to get the corresponding accuracy of the valid. Finally, in all parameter pairs which ensure the highest valid accuracy for the training set, chooses a pair of punishment parameter c and kernel function parameter g in which the punishment parameter is minimum as the regression parameter [8]. In the actual computing, uses the function SVMcgForRegress() of the tool of the libsvm-mat-2.89-3 in Matlab (Version R2009a) to determine the best parameters, and the range of the parameters is both in the interval [-8,8], and the step is 1, sets the cross valid parameter 5, subsequently obtains the best punishment parameter c and the kernel function parameter g, and the value is (0.5, 5.569) .

C. Case modeling

The method in the paper is based on the support vector machine theory to forecast the stock index. Firstly, analyzes the important factors of the stock market, and chooses the main index to establish the testing set, and then normalizes the origin data. Secondly, trains the training set based on the support vector machine theory, uses the best regression parameters to establish the forecasting model. Finally, forecasts the stock index by the optimized forecasting model [8], and the flow is shown in figure 1. In the case, chooses the index of Shanghai stock from October 28, 1991 to March 10, 2011, which consists 4736 data. And using the opening index, maximum index, minimum index, close index, trading volume, trading turnover of the day before as the independent variable, and the opening index of the day is used as the dependent variable, finally establishes the stock index forecasting model.

IV. PREDICTION AND ANALYSIS

A. Data prediction

In the basis of the best regression parameters, trains the training set based on the SVM theory, and forecasts the stock index from March 2,2011 to March 10,2011. The prediction result is shown in figure 2, from the figure the conclusion can be obviously drown: the method base on the SVM theory is better than that base on the chaos theory, which can accurately reflect the change tendency of the stock index.

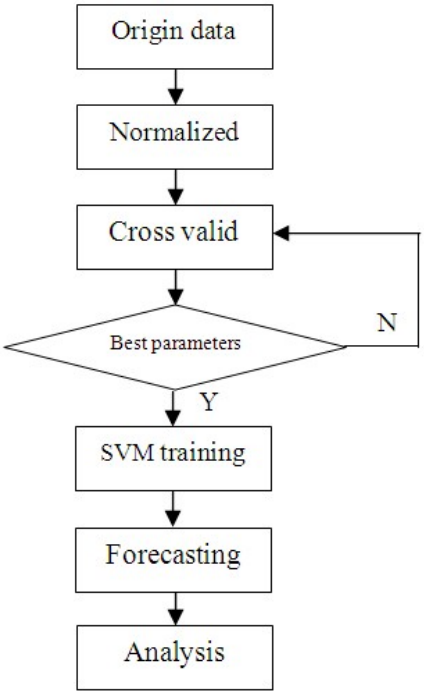

Fig.1. The flow of the prediction based on SVM

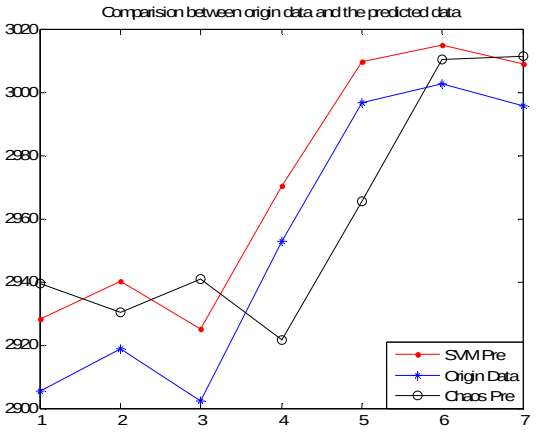

Fig.2. The prediction data and the origin data of stock index

B. Error analysis

In the paper, adopts the mean absolute percentage error and the mean square error to measure the prediction effects, and the formula are in following:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i^{'} - y_i}{y_i}\right| \times 100\%, i = 1,2,\Lambda\ N \tag{8}$$

$$RMS = \times\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{y_i^{'} - y_i}{y_i})^2}, i = 1,2,\Lambda\ N \tag{9}$$

In the formulas, $y_i^{'}$ is the prediction data, and the y_i is the origin data. In table1, the error of the different methods is compared, likewise, the conclusion can be drown from the error that the method based on SVM theory is better than that based on the chaos theory, at the same time, the method based on the SVM theory strict theoretical system, on the other hand the method based on the chaos theory is lack of strict theoretical basis.

TABLE 1. THE ERROR OF DIFFERENT METHODS

evaluation method	Method based on SVM theory	Method based on chaos theory
MAPE（%）	0.59	0.66
RMS（%）	0.89	1.29

V. CONCLUSIONS

For the nonlinear characteristic of the stock index, in order to improvement the prediction accuracy, establishes the prediction model based on the SVM theory and forecasts the Shanghai stock index. Proved by the results, the method based on the SVM theory can better reflect the change tendency of the stock index and the prediction effect is better than others, so the method is feasible in stock index prediction. But the method of the best parameters choice still needs to be improved.

REFERENCES

[1] BAO Xin-zhong, LIU Cheng, SUN Bin. A Study on Setting Parameters of Stock Index Prediction by Applying LM-BP Neural Network. Journal of Systems & Management, pp667-670, 2009.

[2] LIN Da-chao, AN Feng-ping, GUO Zhang-lin, ZHANG Li-ning. Prediction of landslide displacements through multimode support vector machine model. Rock and Soil Mechanics, Supp1:pp451-452, 2011.

[3] Zhao Hong-wei. A Short Term Load Forecasting Method Based on PAR Model. Proceedings of the CSEE, pp347-350,1997.

[4] WANG Xian, ZHANG Shao-hua. An Improved Method for Short-Term Electric Load Forecasting Using Time Series Techniques. Journal of Shanghai University(Natural Science Edition), pp133-136, 2002

[5] Marin F.J., Garcia-Lagos, F., Joya G., Sandoval, F. Global model for short-term load forecasting using artificial neural networks. Generation, Transmission and Distribution, IEE Proceedings, Volume 149, Issue 2, March 2002: pp121 – 125

[6] Hiroyuki Mori, Shouichi Urano. Short-Term Load Forecasting with Chaos Time Series Analysis. International Conference on Intelligent Systems Applications to Power Systems, ISAP'96, 28 Jan.-2 Feb. 1996:pp133-137

[7] LIANG Kun, NIE Hui-xing, XU Zong-wei. Prediction of price indices of Beijing real estate based on support vector machine. Journal of Hefei University of Technology(Natural Science), pp588-591,2011.

[8] Shi Feng, Wang Xiao-chuan, Yu Lei, Li Yang. MATLAB 30 Neural network cases analysis. Beijing, Beihang University press, 2010.

[9] SHI Yue-zhen, XU Dong-mei1. Support vector machine model based study on low-water forecast of Xiangjiang River. Water Resources and Hydropower Engineering, pp71-73,2011.

[10] YANG Hong, GU Shi-fu1, CUI Ming-dong, SUN Yu. Forecast of short-term wind speed in wind farms based on GA optimized LS-SVM. Power System Protection and Control, pp44-47,2011.

[11] WEI Jun, ZHOU Bu-xiang, LIN Nan, XING Yi. Short-term load forecasting based on MG-CACO and SVM method. wer System Protection and Control, p36-38,2009.

[12] WANG Guan-yu, GUO Yong. Study on Telecom Customer Leaving Prediction Based on Support Vector Machine. Computer Simulation , p115-118, 2011