World Scientific Research Journal Volume 6 Issue 11, 2020
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0009

Analysis and Forecast of Stock Index based on ARIMA-SVR Model

Qingqing Lin[1, *]

[1]College of Economics, Jinan University, Guangzhou, China.

Abstract

The main purpose of this paper is to establish ARIMA model and SVR model to analyze and forecast the stock price composite index. This paper takes CSI 300 Index as an example and takes the closing price of T+1 day as the main research variable. Firstly, the linear part of the stock index is predicted by using the traditional ARIMA model. Then, the SVR model is constructed to deal with the nonlinear factors of the stock index by considering the variables such as the closing price and trading volume on t day. Considering the shortcomings of the single model, the first mock exam is to build two models and compare the prediction results with the ARIMA-SVR models. The results show that ARIMA-SVR model improves the prediction accuracy of the model to a certain extent, and can accurately predict the trend and fluctuation of CSI 300 Index in the short term, and can provide certain guidance for investors' investment decisions.

Keywords

Stock index; ARIMA model; SVR model; kernel function.

1. INTRODUCTION

As we all know, the stock market is a market with high returns and high risks. The frequency of stock price fluctuation is high and the fluctuation range is large. It is almost impossible to predict the fluctuation of stock price completely and accurately. For ordinary investors, people often judge the expected trend of the stock according to the technical indicators such as MACD, KDJ, RSI, OBV provided by the stock software, plus their own investment experience, and carry out the corresponding trading operation. However, the stock market is a comprehensive system, the mechanism of action is very complex, affected by many external uncertainties. The difficulty of stock forecast and analysis is decided by its non-linear, non-stationary and other factors. Therefore, the use of certain methods and techniques to mine important information from the historical data of stocks, and successfully applied to the prediction and analysis of the trend, has become the focus of the majority of scholars.

Based on the consideration of the stock market as a highly complex nonlinear system, the stock index has both linear and nonlinear relations. The traditional ARIMA model is used for linear prediction, and then the support vector machine (SVM) algorithm, which is widely used in the field of classification and regression, is used to deal with the nonlinear factors. Recognizing the shortcomings of single forecasting model, this paper uses the combination model to forecast the stock index. Through the comparison of the prediction results, the combination model can improve the prediction accuracy of the model, reduce the prediction error, and provide a certain reference value for accurately studying and judging the fluctuation trend of the stock index.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0009

2. CORRELATION THEORY

2.1. Support Vector Regression (SVR)

Support vector machine (SVM) [1] is an intelligent machine learning algorithm, which is mainly based on statistical learning theory. SVM makes full use of learning mechanism and achieves high statistical prediction according to the principle of structural risk minimization. In addition to solving the classification problem, it is often used in the field of regression prediction. At this point, support vector regression (SVR) plays a key role. When SVR solves the regression fitting problem, it no longer looks for the optimal hyperplane to divide the sample types like the classification problem, but looks for an optimal hyperplane to minimize the total deviation of all training samples from the plane.

Assuming that there are a certain number of N training samples $\{(x_i, y_i), i = 1,2,\cdots,n\}$, the regression estimation function of the sample set is $f(x) = w\Phi(x) + b$, $\Phi(x)$ a nonlinear mapping function.

A linear insensitive function ε defined as :

$$L(f(x), y, \varepsilon) = \begin{cases} |y - f(x)| - \varepsilon, & |y - f(x)| > \varepsilon \\ 0, & |y - f(x)| \le \varepsilon \end{cases} \tag{1}$$

Where $f(x)$ is the predicted value of sample regression.

The mathematical model of support vector regression can be expressed as :

$$\begin{cases} \min \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) \\ s.t. \begin{cases} y_i - w\Phi(x_i) - b \le \varepsilon + \xi_i \\ w\Phi(x_i) + b - y_i \le \varepsilon + \hat{\xi}_i \\ \xi_i \ge 0, \hat{\xi}_i \ge 0 \end{cases}, i = 1,2,\cdots n \end{cases} \tag{2}$$

Where C is the penalty factor; $\xi_i, \hat{\xi}_i$ is the relaxation factor; ε represents the linear insensitive function value.

The Lagrange multiplier method is also used to solve the model :

$$f(x) = w^*\Phi(x) + b^* = \sum_{i=1}^{n}(a_i - a_i^*)K(x_i, x) + b^* \tag{3}$$

Where a_i, a_i^* is Lagrange coefficient, and $a_i, a_i^* > 0, i = 1,2,\cdots,n$; $K(x_i, x)$ is the kernel function.

2.2. Kernel Function

Aiming at the problem of linear indivisibility in practical application, the kernel function is used Φ to map the original sample R^d to a higher dimensional space H, and then the optimal hyperplane is found in the high-dimensional space.

$$\begin{aligned} &\Phi : R^d \to H \\ &K(x_i, x_j) = \Phi(x_i)\Phi(x_j) \end{aligned} \tag{4}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0009

The common kernel function types are:

(1) Linear kernel function

$$K(x, x_i) = xx_i \tag{5}$$

(2) The kernel function of order d is polynomial

$$K(x, x_i) = (xx_i + 1)^d \tag{6}$$

(3) Gaussian radial basis function kernel function

$$K(x, x_i) = \exp(-\frac{\| x - x_i \|^2}{2\sigma^2}) \tag{7}$$

2.3. Model Evaluation

(1) Root mean square error. RMSE [2] refers to the square root of the sum of the squares of the deviation between the observed value and the estimated value and the square root of the expected value. RMSE is often used to measure the standard error of models. When the root mean square error of the model is smaller, it shows that the prediction model has higher prediction accuracy and can better reflect the actual situation. The calculation formula of RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2} \tag{8}$$

(2) Square of correlation coefficient between fitting value and actual value. When the square value of the correlation coefficient between the fitting value and the actual value is closer to 1, the correlation between the two variables is greater, and the prediction result of the model is closer to the actual value. The calculation formula is as follows:

$$R^2 = \frac{\left(n\sum_{i=1}^{n}(f(x_i)y_i - \sum_{i=1}^{n}f(x_i)\sum_{i=1}^{n}y_i\right)^2}{\left(n\sum_{i=1}^{n}f(x_i)^2 - \left(\sum_{i=1}^{n}f(x)\right)^2\right)\left(n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2\right)} \tag{9}$$

3. MODEL ESTABLISHMENT

3.1. Data Source and Preprocessing

CSI 300 Index is selected as the research object. To obtain the daily frequency data of all trading days of CSI 300 Index from January 1, 2018 to August 15, 2019. The first 75% of the observed data is used as the training data set, and the last 25% is used as the test data set. The closing price on T+1 day is taken as the prediction variable, and the opening price, closing price, trading volume, trading volume, maximum price and lowest price closely related to the closing price on T+1 day are taken as the input variables of the model.

Do standard normalization for the sample data set and make them all map to the [0, 1] interval. The processing formula is shown in equation10:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0009

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \qquad i = 1, 2, \cdots, n \tag{10}$$

Among them, is the original data of the index, is the standardized data, is the maximum value in the original data, is the minimum value in the original data.

3.2. ARIMA Modeling Results

According to the ADF test results, the closing price data is a non-stationary series. The ARMA model is built for the first-order difference series. Based on the minimum information criterion, the model order is automatically identified by Eviews software, and the result is ARMA (1, 3). Therefore, the modeling result of the original sequence is ARIMA (1, 3, 1). At this time, the results of model F test are significant, and AR(1), MA(1), MA(3) all pass the t-test with significance level of 0. 01.

3.3. SVR and ARIMA-SVR Modeling Results

The volatility frequency of stock index is often high, and the fluctuation range is difficult to predict. The accuracy of forecasting the future closing price by using historical closing price is often low. Therefore, this paper takes the T-day opening price, closing price, trading volume, turnover, maximum price and lowest price into account, and establishes support vector regression model (SVR) to predict the CSI 300 Index by using the common support vector machine algorithm in machine learning algorithm. Different kernel functions are selected to analyze the effect of the model.

By selecting linear kernel function, polynomial kernel function and Gaussian radial basis function RBF, the optimal parameters c and G are selected by grid search method. According to the prediction results of the three sum functions, ARIMA-SVR (linear) model based on linear kernel function is constructed. The sample data are simulated and RMSE and R2 are obtained as follows.

Table 1. prediction effect of different kernel functions

kernel function	Training set		Test set	
	RMSE	R^2	RMSE	R^2
SVR(Linear)	58. 8140	0. 9948	45. 2618	0. 9674
SVR(Poly)	154. 2918	0. 9106	166. 1844	0. 9550
SVR(RBF)	62. 2600	0. 9936	47. 3798	0. 9629
ARIMA-SVR(Linear)	57. 6652	0. 9965	44. 3704	0. 9713

From the results, the linear kernel function has the highest prediction accuracy among the three SVR models. Both SVR(Linear) model and ARIMA-SVR have higher R2 on the training set and test set. From the values of RMSE and R2, ARIMA combined with SVR can bring higher R2 and lower prediction error.

The prediction effect of ARIMA-SVR model in training set and test set is shown in Figure 1:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202011_6(11).0009

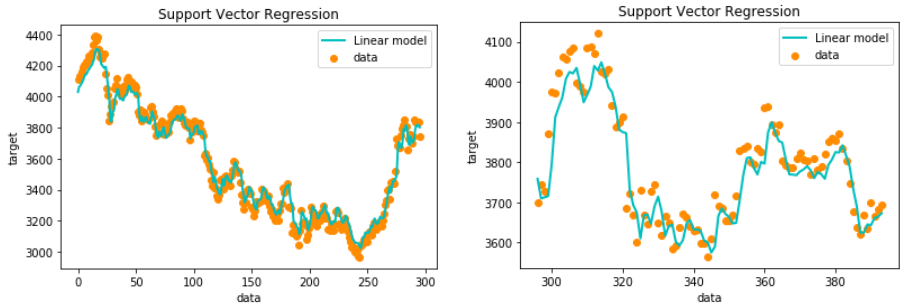

Figure 1. ARIMA-SVR prediction effect

4. CONCLUSION

Aiming at the problem of stock index analysis and prediction, this paper takes CSI 300 Index as an example, establishes prediction models by using ARIMA model and SVR model, in which the linear kernel function performs better in the samples selected in this paper. Finally, combining the two methods, the empirical results show that ARIMA-SVR model with ARIMA prediction results of T+1 closing price as explanatory variable can effectively improve the prediction ability of the model and reduce the prediction error.

REFERENCES

[1] Chenxi Zhang, Yanping Zhang, et al. Stock forecasting based on Support Vector Machine, Computer Technology and Development, Vol. 16 (2006) No. 3, p. 35-37.

[2] Wei Zhang: Research on stock index prediction based on linear and nonlinear Support Vector Machine combination model (MS. , Harbin Institute of technology, China 2017).

[3] Lian Cheng: Research on personal credit evaluation method of internet finance based on Support Vector Machine (MS. , Zhejiang University of Finance and Economics, China 2017).

[4] Yu Zhao: Air Quality Index prediction based on ARIMA and SVR combination model (MS. , Tianjin University of Commerce, China 2019).

[5] Meixia Liu: Analysis and prediction of Shenzhen Stock Index based on ARIMA model, China Urban Economy, (2011) No. 18, p. 70-71.

[6] Jinming Yuan: Application of ARMA-SVR model based on singular spectrum analysis in stock index prediction (MS. , Shandong University, China 2019).

[7] T. B. Trafalis and H. Ince: Support vector machine for regression and applications to financial forecasting, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, vol. 6 (2000) p. 348-353.