Stock Index Forecasting Based on Hybrid ARIMA and LSSVM Methodology

Lei Yuan

Centre for Applied Economics and Finance, Harbin Institute of Technology
Shenzhen Graduate School, China

Email: yuanlei1130@163.com

Keywords: ARIMA; LSSVM; Hybrid model; Time series forecasting; Stock index

Abstract. Both theoretical and empirical findings have indicated that integration of different models can be an effective way of improving the forecasting accuracy of time series, especially when there is a big difference between the combined models. Autoregressive Integrated Moving Average (ARIMA) model is one of the most popular linear models for time series forecasting. However, ARIMA model cannot effectively capture nonlinear patterns hidden in a time series. As a nonlinear model, Least Squares Support Vector Machine (LSSVM) can be applied to time series forecasting with a high degree of accuracy. Combining ARIMA model and LSSVM may further improve the prediction performance. It can helps investors making investment decisions to forecast stock index effectively. In this paper, a hybridization of ARIMA and LSSVM is proposed to forecast the daily closing price of SSE 180 stock index. The empirical results indicate that when linear and nonlinear models were hybridized properly, the forecasting performance of the hybrid model proposed in this paper outperforms the ARIMA model, LSSVM model and other hybrid models.

Introduction

Since forecasting plays an important role in various application fields, more and more scholars have conducted deep researches on time series forecasting. The forecasting of the Internet traffic can help telecom operators to improve their service quality; the forecasting of the epidemic trend of infectious diseases can help medical and health departments to take preventive measures in advance; the forecasting of financial time sequence can help investors to make investment decisions.

In the literature, Autoregressive Integrated Moving Average (ARIMA) model is one of the most popular linear models for time series forecasting due to its attractive theoretical properties and empirical evidence in its support. ARIMA model assumes that the present value is a linear function of the past values and past errors. Stationarity is a necessary condition in fitting an ARIMA model used for forecasting. The ARIMA model has enjoyed useful application in forecasting, such as inventory demand [1], sugar prices [2]. However, as a linear model, ARIMA can only be used to establish a model of linear patterns of the time series. In order to overcome the limitations of the ARIMA model, considering the existed nonlinear patterns of an actual time series, a nonlinear model was then proposed. Support Vector Machine (SVM) is a nonlinear model, which is based on statistical learning theory and structural risk minimization principle, having the ability to overcome the over fitting problem to some extent [3]. The forecasting of time series by using the SVM has also been widely applied to many fields, such as tax revenue forecasting [4], electric power consumption forecasting [5], stock market trend forecasting [6], and so on. In 1999, Suykens et al. proposed the Least Squares Support Vector Machine (LSSVM) [7], which improved the speed of solving optimization problems in the model compared with standard SVM.

However, it is not appropriate to apply the ARIMA model to complex nonlinear problems, or apply the LSSVM to linear problems, or apply the two models respectively to the complex problems involving both linear and nonlinear patterns. In order to overcome the limitations of single models and improve the accuracy of forecasting, the use of a hybrid model for forecasting has become a consensus. In addition, because it is difficult to completely determine the characteristics of the data in the actual problems, so in practice, a hybrid model with both linear and nonlinear fitting ability is the best choice. Zhang put forward to using a hybrid model of ARIMA and ANN in the forecasting of time series [8]. Zhang assumed that there were nonlinear components in the residuals of the ARIMA model, and ANN was used to extract the nonlinear components in the residuals, finally the forecasted values of both ARIMA model and ANN model were added to obtain the forecasted values of the hybrid model. On the basis of Zhang, Khashei and Bijari presented a new hybrid model integrating both ANN and ARIMA models, where the linear and nonlinear components of time series were no longer assumed to be in the form of addition, and where the relationship between the two components was denoted by a nonlinear function [9]. In this paper, we adopt the ideology of the construction of a hybrid model presented by Khashei and Bijari, constructing an ARIMA-LSSVM hybrid model to forecast the SSE 180 stock index, of which 180 sampled stocks have the characteristics of large scale, high liquidity. Its performance can reflect the overall conditions of big high-quality companies in the Shanghai stock market. The forecasting of the SSE 180 stock index will provide a reference for investors to make investment decisions.

Methodology

ARIMA Model. The Autoregressive Integrated Moving Average (ARIMA) model has the form

$$(1-\phi_1 B-\cdots-\phi_p B^p)\ \nabla^d x_t=(1-\theta_1 B-\cdots-\theta_q B^q)\ \varepsilon_t \tag{1}$$

where $\nabla^d=(1-B)^d$ is the difference operator with an order of *d*. The difference operator aims to make a stationary time series, because stationarity is a necessary condition in building an ARIMA model used for forecasting.

When the ARIMA model is proposed to fit the time series, firstly it has to be determined whether the series is stationary, by using the ADF stationary test in this paper. If the series is not stationary, it may need a difference operation for the original sequence, until the sequence is stationary. When fitting the difference stationary series with the ARMA model, except for the need to test the significance of parameters, we should also carry out the significance testing of the model. The significance testing of the model is carried out to conduct a white noise test on the residuals, thus ensuring that there is no longer linear correlation relationship between the residuals.

Although the ARIMA model can capture the linear patterns of stock index well and is relatively easy to use, it is not adequate for stock index forecasting for its deficiency in capture the nonlinear and irregular in stock index. Therefore, the LSSVM is used in this paper to capture the nonlinear patterns in the stock index.

LSSVM Model. Support Vector Machine (SVM) was proposed by Vapnik in 1995 [10] and it exhibit many unique advantages in solving small sample, nonlinear and high dimensional pattern recognition. It can also promote the use of machine learning to other problems such as function fitting. Suykens proposed an improved SVM known as least squares support vector machine (LSSVM) in 1999 [7], which can formulates the training process by solving linear problem quicker than SVM through quadratic programming.

Consider a given dataset $\{(x_i, y_i), i=1,2,\cdots n\}$, in which $x_i \in R^m$ is the input data and $y_i \in R$ is the output data. LSSVM defines the regression function as

$$\min J(\omega,e)=\frac{1}{2}\omega^T\omega+\frac{1}{2}\gamma\sum_{i=1}^{n}e_i^2 \tag{2}$$

subject to

$$y_i = \omega^T \phi(x_i) + b + e_i, i = 1, 2, 3 \cdots n \ , \tag{3}$$

where *w* is the weight vector; γ is the regularization parameter, determining the trade-off between the training error minimization and smoothness of the estimated function; e_i is estimated error; $\phi(\cdot)$ is the nonlinear function and *b* is the bias term. In order to solve the programming problem above, the corresponding Lagrange function is as Eq. 4.

$$L(\omega, e, \alpha, b) = J(\omega, e) - \sum_{i=1}^{n} \alpha_i (\omega^T \phi(x_i) + b + e_i - y_i) \ , \tag{4}$$

where α_i is the Lagrange multiplier. The solution can be obtained by using the Karush-Kuhn-Tucker (KKT) condition:

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

The LSSVM for nonlinear fitting function can be obtained from

$$y(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b \, , \tag{6}$$

where $K(x, x_i)$ is the kernel function.

The forecasting performance of LSSVM is not always better than ARIMA model. ARIMA model and LSSVM is complementary in time series forecasting. When we combine the ARIMA and LSSVM model appropriately, it may improve the forecasting precision of stock index.

Hybrid Models. Like most of time series, stock index is rarely purely linear or nonlinear. Due to the high complexity of the investors' behavior, the stock index often contains both linear and nonlinear patterns. The performances of single models are not good in forecasting when the time series has high complexity. Neither ARIMA nor LSSVM can sufficiently forecast the stock index because the single models cannot model both linear and nonlinear patterns in the stock index.

In the literature, different hybrids model which contain linear and nonlinear model have been proposed in order to forecast the complex time series. The advantage of the hybrid models are that they can integrate the advantages of single models and they can deal with different patterns in complex time series. The aim of using hybrid models in time series forecasting is to reduce the risk of model specification error. Zhang proposed a hybrid ARIMA-ANN model [8] for time series forecasting in 2003. According to the hybrid model, a time series can be composed of a linear autocorrelation structure and a nonlinear component. That is,

$$y_t = L_t + N_t \tag{7}$$

where L_t denotes the linear component and N_t denotes the nonlinear component. The hybrid model takes advantage of the unique strength of ARIMA and LSSVM models in linear and nonlinear modeling.

In 2011, Khashei and Bijari [9] proposed a novel form of hybrid model which assumed that a time series is considered as function of a linear and a nonlinear pattern instead of the sum of them.

$$y_t = f(L_t, N_t) \tag{8}$$

In this paper, according to the idea of Khashei and Bijari, we proposed a hybrid model as Eq. 9

$$y_t = f(y_{t-1}, y_{t-2} \cdots y_{t-p}, \varepsilon_{t-1}) \tag{9}$$

where $\varepsilon_{t-1} = y_{t-1} - \hat{L}_{t-1}$ is the residual of ARIMA model which represents the nonlinear pattern in the time series. The LSSVM is used to fit the function of a linear and nonlinear component.

Forecasting of Stock Index

The Data Description. This paper empirically analyzes SSE 180 stock index daily closing prices, which was from Wind Info and contained a total of 243 observations starting from December 3, 2013 to November 28, 2014. The first 225 observations of them were taken as the in-sample training set and used to construct the model, and the last 18 observations were taken as the out-of-sample and used to verify the forecasting effects.

Forecasting Evaluation Criteria. In order to measure the forecasting performance between different models, RMSE and MAPE are computed from the Eq. 10 and Eq. 11

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2} \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2} \tag{11}$$

RMSE measures the size of the absolute errors, and MAPE measures the size of the relative errors. The smaller the two indicators are, the closer are the actual values and the forecasted values, which illustrates a higher accuracy of the forecasting.

ARIMA Model. This paper established an ARIMA model of the SSE 180 index closing prices by using SAS 9.2 software. The original sequence of the stock index is not stationary. When conducting a difference operation as Eq. 12, the difference sequence becomes stationary.

$$z_t = y_t - y_{t-1},\ t = 1, 2, 3\cdots \tag{12}$$

If the difference sequence $\{z_t, t = 1, 2, 3\cdots\}$ is white noise, there is no need to fit the ARIMA model, because the best model is ARIMA (0,1,0) which the time series is a random walk. The result of autocorrelation check for white noise of $\{z_t, t = 1, 2, 3\cdots\}$ is show as Table 1.

Table 1 Autocorrelation check for white noise

To Lag	Chi-Square	DF	Pr> ChiSq	Autocorrelations					
6	11.47	6	0.0749	0.133	0.127	-0.018	0.071	-0.026	-0.082
12	17.78	12	0.1224	-0.007	-0.004	0.039	0.096	0.044	-0.110
18	20.32	18	0.3149	-0.005	0.042	0.042	0.038	0.046	0.000
24	21.78	24	0.5924	0.037	-0.044	-0.005	0.033	-0.031	-0.009

The test show that $\{z_t, t=1,2,3\cdots\}$ is white noise and the stock index is random walk which is in accordance with effective market hypothesis (EMH) [11]. The hypothesis indicates that the best forecasting value for the next daily closing price is the current value of the stock index.

LSSVM Model. Technical analysis generally uses 5 moving average, 10 day moving average or 20 day moving average as the analysis object. In this paper, only the one-step-ahead forecasting is considered. So we specify the number of input is 4, namely forecasting the next trading closing price by using 4 trading closing price before.

$$\hat{y}_t = f_1(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}) \tag{13}$$

For the LSSVM model, the LSSVMLab toolkit developed by Suykens on MATLAB 2010a platform was used to build the nonlinear function. As usual, The RBF function was chosen as kernel function of LSSVM. We use 10 folds cross - validation to optimize the parameters γ and σ^2 of the LSSVM, with the results of $\gamma = 4617.47$, $\sigma^2 = 80.97$.

Hybrid Model. Due to the idea of Khashei, we propose the hybrid model which contains both linear and nonlinear patterns as Eq. 14.

$$\hat{y}_t = f_2(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, \varepsilon_{t-1}) \tag{14}$$

10 folds cross-validation was used to optimize the parameters γ and σ^2 of LSSVM, with the results of $\gamma = 4490.42$, σ^2=4742.98.

In order to compare with other hybrid model, the hybrid model proposed by Zhang was chosen to be the representative. The structure of Zhang's hybrid model is as Eq.15 and Eq. 16.

$$\hat{y}_t = \hat{L}_t + \hat{\varepsilon}_t \tag{15}$$

$$\hat{\varepsilon}_t = f_3(\varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-3}, \varepsilon_{t-4}) \tag{16}$$

where $\hat{L}_t$ is the forecasting value of ARIMA. Similarly, the function f_3 (.) was determined by LSSVM. The determination of the parameters is similar to the above model and the parameters are γ=0.1758, σ^2=2.7040.

Experimental Results

The SSE 180 stock index is applied to test the four models above and the indicators RMSE and MAPE are used to evaluate the forecasting performance of the models. The comparison of the forecasting precision is show in Table 2 and Table 3.

Table 2 Comparison of the performance of the proposed model with other models

Model	RMSE	MAPE
ARIMA	79.012	0.0103
LSSVM	71.1632	0.0095
Zhang's hybrid model	71.2806	0.0091
Proposed model	64.8819	0.0084

Table 3 Percentage improvement of the proposed model in comparison with other models

Model	RMSE(%)	MAPE(%)
ARIMA	17.88	18.45
LSSVM	8.83	11.58
Zhang’s hybrid model	8.98	7.69

Among different forecasting models, the forecasting accuracy of hybrid model proposed in this paper is higher than that of ARIMA, LSSVM model and Zhang’s hybrid model. When considering the MAPE, the performance of Zhang’s hybrid model outperforms the ARIMA and LSSVM model. It shows that the hybrid model has advantage in time series forecasting over single models.

Conclusion

Linear ARIMA model and nonlinear LSSVM model cannot individually model the stock index accurately. Hybrid techniques that decompose the time series into linear and nonlinear patterns are one of most popular methods in time series forecasting. Hybrid models containing the strengths of ARIMA and LSSVM exploit the advantage of both types of models simultaneously. In this paper, we proposed a hybrid model for stock index forecasting. The empirical results show that the hybrid model proposed in this paper produce the lowest RMSE and MAPE in out-of-sample dataset. It exceeds the individual models and Zhang’s hybrid model. However, not all hybrid models can outperform single models [12,13]. Only the single models are hybridized appropriately, the hybrid models will have more excellent performance in forecasting.

References

[1] M.Z. Babai, M.M. Ali, J.E. Boylan and A.A. Syntetos: International Journal of Production Economics, Vol. 143 (2013) No.2, p.463.

[2] K.K. Suresh and S.K. Priya: Sugar Tech, Vol. 13 (2011) No.1, p.23.

[3] V. Vapnik, S.E. Golowich and A. Smola: Advances in Neural Information Processing Systems (1997), p.281.

[4] L.X. Liu, Y.Q. Zhuang and X.Y. Liu: Expert Systems with Applications, Vol. 38 (2011) No.1, p.116.

[5] K. Kavaklioglu: Applied Energy, Vol. 88 (2011) No.1, p.368.

[6] Y. Sai, Y. Zheng and K. Gao: *Proc. IEEE International Conference on Granular Computing* (2007). p.659.

[7] J.A. Suykens and J. Vandewalle: Neural Processing Letters, Vol. 9 (1999) No.3, p.293.

[8] G.P. Zhang: Neurocomputing, Vol. 50 (2003), p.159.

[9] M. Khashei and M. Bijari: Applied Soft Computing, Vol. 11 (2011) No.2, p.2664.

[10] V. Vapnik: *The Nature of Statistical Learning Theory* (Springer, New York 1995).

[11] C.C. Lee, J.D. Lee and C.C. Lee: Japan and the world economy, Vol. 22 (2010) No.1, p.49.

[12] M. Khashei and M. Bijari: Expert Systems with Applications, Vol. 39 (2012) No.4, p.4344.

[13] U. Yolcu, E. Egrioglu and C.H. Aladag: Decision Support Systems, Vol. 54 (2013) No.3, p.1340.