# Stock price prediction based on ARIMA - SVM model

**Wenjuan Mei[a], Pan Xu[b], Ruochen Liu[c], and Jun Liu[d,*]**

School of Management Science and Engineering Nanjing University of Finance and Economics , No. 3 Wenyuan Road, Nanjing, China

[a]dearxp0228@163.com, [b]1563477026@qq.com, [c]965703636@qq.com, [d]9120031038@nufe.edu.cn

* corresponding author

**Keywords:** Support vector machine, ARIMA model, Stock price prediction

**Abstract:** Stock price is a complex non-stationary and non-linear time series, which is affected by economic cycle, financial policy, international environment and other factors, so the movement direction of stock price is unknown and complex. In order to accurately predict the trend of stock price, this paper proposes the ARIMA — SVM model, which is optimized and improved on the basis of the support vector machine model (SVM). Therefore, this model is able to process multi-dimensional nonlinear data. Firstly, ARIMA model was used to predict the data, and the error result obtained was used as the input variable of support vector machine (SVM). In the construction of SVM model, cross-validation method was used to traverse the search of parameter combination, and then the optimal parameter combination was determined, so as to predict the rise and fall trend and fluctuation direction of stock price. Through the empirical analysis of IBM stock, the accuracy of the model reaches 96.10%.

## 1. Introduction

At present, stock investment has become one of the ways for many people to conduct financial management, and more and more scholars are engaged in stock market analysis and stock price prediction. Stock data usually have time sequence, which can be considered as a kind of time series data with significant nonlinear and time-varying characteristics. One of the most widely used and common time series models for stock time series data is the ARIMA model, which is due to its simplicity, feasibility and flexibility [1]. It assumes that stock prices are a deterministic, linear process of change, but they are not, so ARIMA's predictions are usually less than ideal. The fluctuation of stock price is often complex, and it is difficult to describe the change characteristics of stock price with exact mathematical formulas. This information processing method is exactly what support vector machine (SVM) has. Support vector machine (SVM) is a new technology in data mining, machine learning and artificial intelligence. It belongs to nonlinear prediction model and is suitable for the modeling and prediction of stock price fluctuation system [2-4]. Francis (2011) used the support vector machine model to realize the prediction of financial time series. He took the futures data of Chicago as the research object and concluded through empirical analysis and comparison that the support vector machine model was significantly better than BP neural network in model performance, prediction accuracy and generalization ability [5]. This paper attempts to establish a combination model of smooth autoregressive moving average model (ARIMA) and support vector machine (SVM). It is organized as following: Firstly, ARIMA model and SVM model were used to predict the data respectively; Secondly, the error result obtained from ARIMA model was used as the input variable of support vector machine (SVM) to predict the closing price; Thirdly, through comparative analysis and the empirical analysis, it is found that compared with the ARIMA model and the SVM model, the ARIMA-SVM model can get higher accuracy. Experimental results showed that the hybrid models can predict stock prices more accurately.

DOI: 10.25236/icbdai.2018.008

## 2. Material and Methods

ARIMA model, known as differential autoregressive moving average model, is a famous time series prediction method proposed by Box and Jenkins in the early 1970s [6]. The so-called ARIMA model is to stabilize the non-stationary time series by d-order difference first, and then use the self-feedback for the obtained stationary time series AR(p) process and MA(q) process, and the established model was identified by the sample autocorrelation coefficient (ACF) and partial autocorrelation coefficient (PCF) data, and a set of modeling, estimation, testing and control methods were also proposed. ARIMA (p, d, q) model is called differential autoregressive moving average model, which can be expressed as:

$$x_t = \varphi_0 + \varphi_1 x_{t\text{-}1} + \cdots + \varphi_p x_{t\text{-}p} + \varepsilon_t \text{-} \theta_1 \varepsilon_{t\text{-}1} \text{-} \cdots \text{-} \theta_q \varepsilon_{t\text{-}p} \tag{1}$$

its autoregressive is AR, and the order of autoregressive term is p; The moving average is MA, and the number of moving average terms is q; The number of differences you make when you make the time series stationary is d.

Support Vector Machine (SVM) was first proposed by Corinna Cortes and Vapnik et al in 1995[7]. It has many unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition. Given sample set D=$\{(x_i, y_i)|i = 1, 2, \cdots\ x_i \in R^n, y_i \in R$. By introducing the constraint of relaxation variables $\xi$ 和$\xi^*$, and the transformation variable $\varphi$ from the input space $R^n$ to Hilbert space H, map the data sample set $(x_i, y_i), i = 1,2,\cdots,l, x_i \in R^n, y_i \in R$ to $f(\varphi(x_i), y_i, b) = 0, i = 1,2,\cdots,l,$ and the original regression problem is transformed into an optimization function:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \tag{2}$$

w is the weight vector; C is the penalty factor, and C>0. The constraint condition of the optimization function is:

$$w^1 \varphi(x_i) + b\text{-}y_i \le \varepsilon + \xi_i \tag{3}$$

$$w^1 \varphi(x_i) + b\text{-}y_i \le \varepsilon + \xi_i \tag{4}$$

$$\xi_i, \xi_i^* \ge 0, i = 1,2,\cdots,l \tag{5}$$

b is the parameter of the mapping; $\varepsilon$ is the loss function, and $\varepsilon > 0$.

From the above theoretical introduction, it can be seen that the ARIMA model can achieve good results in the processing of linear relations, but it cannot well deal with complex nonlinear data. SVR model can deal with nonlinear relationship effectively. Therefore, neither model can solve the problem of stock closing price prediction independently. In reality, Stock closing price is affected by both non-linear factors and linear factors. If we want to make a good fitting prediction of stock closing price, we need to compromise the merits of the two models and discard disadvantages of both models. Therefore, a hybrid model is constructed to combine ARIMA and SVR models for fitting prediction of stock closing price, which can achieve better results than a single model.

The stock closing price sequence is divided into two parts, namely the linear part and the nonlinear part. The relationship between the two and the closing price sequence is shown as follows:

$$\mathrm{x_t = L_t + N_t} \tag{6}$$

$x_t$ is the stock closing price sequence, $L_t$ is the linear part of the information and nonlinear information is $N_t$. The fitting and prediction steps of stock closing price using ARIMA-SVM hybrid model are shown as follows:

First, ARIMA model was used to model and predict the closing price data for the first time. The result obtained was the linear part, and the predicted value was denoted as $\overset{\wedge}{L_t}$.

Second, to calculate the residual predicted by the ARIMA model. The residual sequence is defined as:

$$e_t = x_t\text{-}\overset{\wedge}{L_t} \tag{7}$$

We regard the residual sequence $e_t$ as the sum of all the non-linear information except the linear part in the stock closing price data $x_t$.

Third, SVR model is used to model and predict the results obtained in the second step, and the predicted results are denoted as $\overset{\wedge}{e_t}$.

Fourth, adding $\overset{\wedge}{L_t}$ to $\overset{\wedge}{e_t}$ to get the sum is the prediction of the closing price of the stock.

## 3. Analysis and Discussion of Modeling and Simulation

### 3.1. Data Selection and Processing

Data selection is a very important step in empirical research. Before establishing the prediction model, the most critical link is to deal with the selection of data. Only reasonable and standard data selection can make the model computable and predictable at the same time. This paper uses the daily data of closing price as the research object, which can continuously reflect the daily stock price. Taking IBM stock as the target stock studied in this paper, the data interval was selected from January 2, 1962 to November 10, 2017, with a total of 14,059 sets of data. The closing price data of samples from January 2, 1962 to January 2, 2017 were used as the training set, and the closing price data from January 2, 2017 to November 10, 2017 were used as the test set. The source of data is Wind database. In this paper, the demonstration is mainly carried out by Python software, supplemented by Excel to obtain the calculation results used in this paper.

### 3.2. Establishment and Prediction of ARIMA Mode

#### 3.2.1. Stability Handling and Testing

For modeling with time series analysis, the first condition is that the sequence must meet the requirement of balance, that is, to stabilize the data [8]. By observing the sequence diagram of the original data, it is found that the time series has certain fluctuation. Therefore, the difference processing is carried out. First, the time sequence diagram of the new sequence after the first-order difference is observed, as shown in Figure 1. It is found that the sequence has a stationary state, and then the unit root ADF test is carried out for the sequence, and the results are shown in Table 1:

Table 1 Sequence stationarity test results.

| ADF test | Statistic | P-statistic | 1% | 5% | 10% |
|---|---|---|---|---|---|
| Original sequence | 0.11 | 0.96 | -3.43 | -2.86 | -2.56 |
| First order difference sequence | -19.01 | 0 | -3.43 | -2.86 | -2.56 |

As shown by the above result, the original sequence ADF Test result of the P value is 0.96, more than 1%, 5%, 10% under different place sexual level statistics, first-order ADF Test value is less than 1%, the new sequence can significantly decline the existence of unit root, new instructions after the first order difference sequence is stationary series, so the d = 1, can build ARIMA (P, 1, q) model.

### 3.2.2. ARIMA Model Fitting and Parameter Estimation

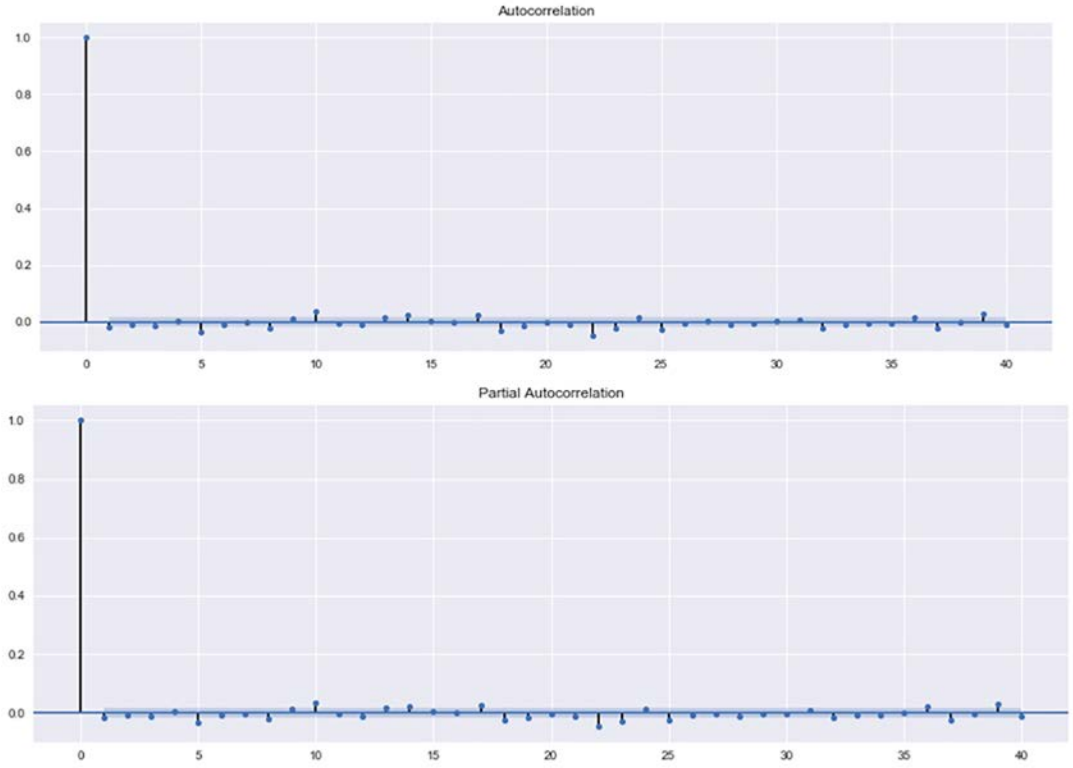

Figure 1 First order difference sequence autocorrelation graph (ACF) and partial autocorrelation function graph (PACF).

The model is fitted and estimated, and the partial correlation coefficient and autocorrelation coefficient are used for judgment and selection. Then, AIC criterion is adopted, that is, the smaller the AIC value is, the higher the accuracy is, the better the fitting is, and the model is selected as the best. It can be found from Figure 1 that both the autocorrelation coefficient and the partial correlation coefficient are basically within the confidence interval. In order to further determine the model, the AIC test statistics and log-likelihood values of each model were compared. Table 2 shows the values of p and q values of each model and the statistics of the AIC values of the corresponding model:

Table 2 Comparison of accuracy of each model.

| q<br>p | 0 | 1 | 2 |
|---|---|---|---|
| 0 | - | 39311.702030 | 39312.674345 |
| 1 | 39311.769506 | 39306.811727 | 39308.790921 |
| 2 | 39312.782035 | 39308.790328 | 39307.589660 |
| 3 | 39312.103834 | 39311.432794 | 39309.576408 |
| 4 | 39313.995002 | 39313.283605 | 39311.501943 |

Obviously, the AIC value of ARIMA(1,1,1) model is the minimum, which is better than other models. Therefore, it is more appropriate to select ARIMA(1,1,1) model for this sequence.

### 3.2.3. Model diagnostic test and prediction

The residual sequence of the model ARIMA(1,1,1) was diagnosed. As can be seen from the autocorrelation diagram named Figure 2, the sequence was a pure random white noise sequence, so the modeling was effective. Figure 3 shows the prediction effect of model ARIMA(1,1,1). Comparing the predicted value with the real value, it seems that there are still some inaccuracies in using the ARIMA model to predict the stock price. This is because the ARIMA model can achieve good results in the processing of linear relations, but it cannot deal well with complex nonlinear data.

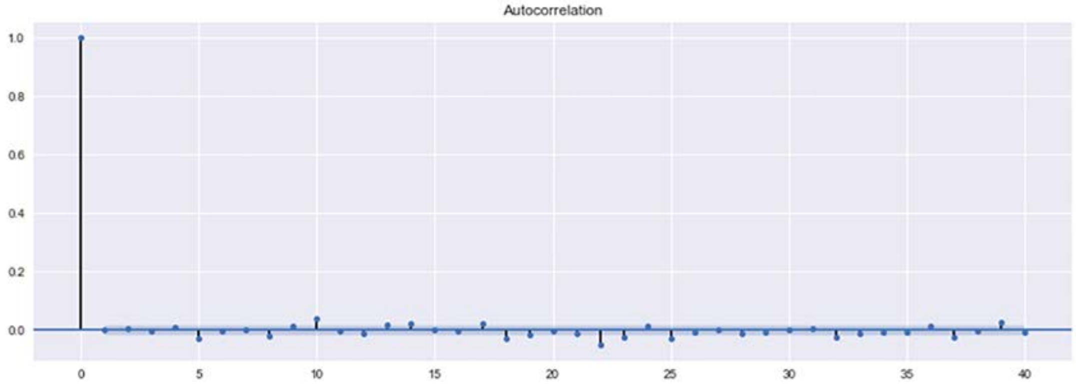

Figure 2 Residual sequence autocorrelation of model ARIMA (1,1,1).

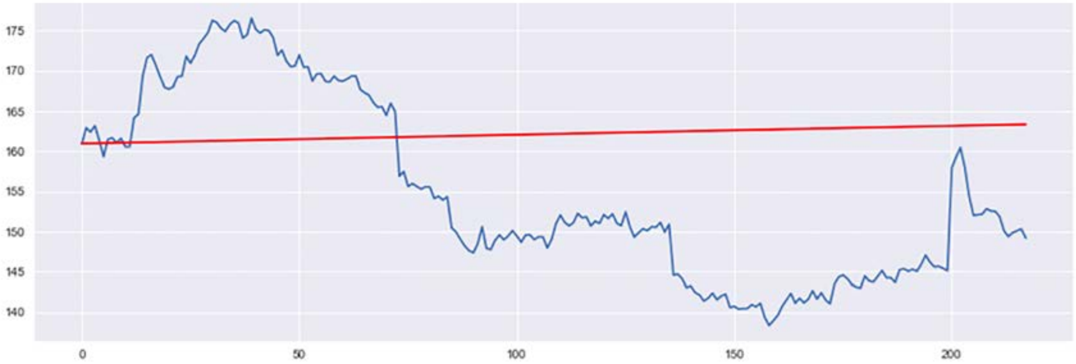

Figure 3 Prediction effect of model ARIMA(1,1,1).

### 3.3. Establishment and prediction of SVM model

For SVM model, in addition to the selection of kernel function is more important, the selection of parameters in the model is also of great significance [9-10]. This paper chooses Radial Basis Function (RBF) as kernel function and adopts the method of 10-fold cross-validation to optimize the parameters, the highest accuracy of a set of parameters, c=1, g=0.3. The accuracy of SVM model reached 88.36%. Figure 4 is the comparison diagram between the predicted stock price of SVM model and the real stock price.

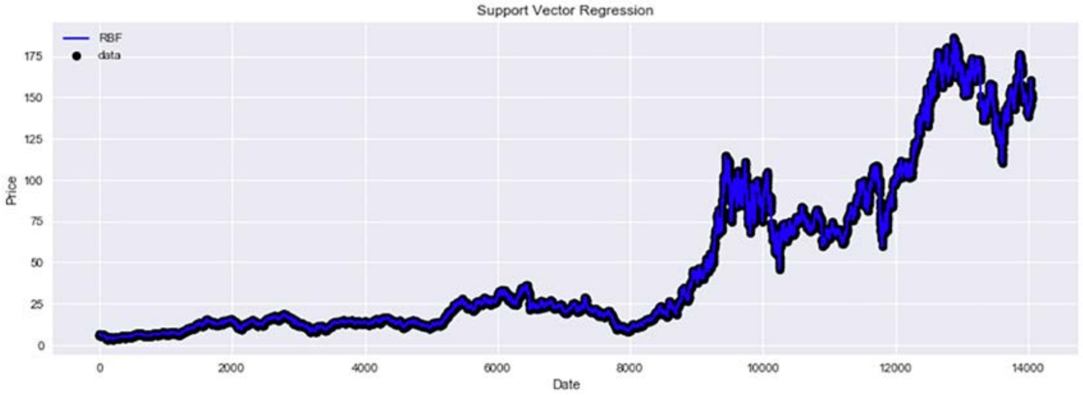

Figure 4 Comparison between stock price predicted by SVM model and real stock price.

### 3.4. Establishment and prediction of ARIMA-SVM model

The error values of 218 predicted data of IBM stock obtained from the previous ARIMA model were taken as the training set and test set of SVR model for predicting IBM error values. The first 150 samples were taken as the training set and the last 68 samples were taken as the test set. The error predicted value of SVR model was added with the closing price predicted value of ARIMA model to obtain the final predicted value. After detection, the final accuracy of the hybrid model was 96.10%. The predicted results are shown in Figure 5, and some of the results are shown in Table 3:

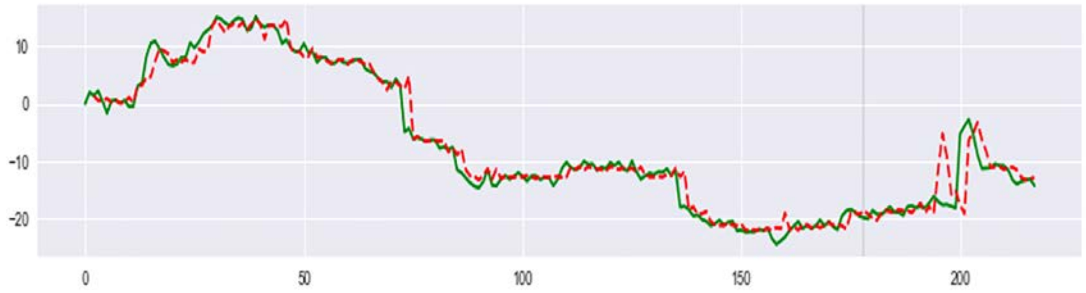

Figure 5 Prediction results of ARIMA-SVM hybrid model.

Table 3 Hybrid model IBM stock price partial forecast results.

| The actual value | The Predictive value | ARIMA error 1 | SVM error 2 | ARIMA-SVM error 3 | Total error |
|---|---|---|---|---|---|
| **160.95** | 161.42 | 0.02 | 0.1 | 0.49 | -0.47 |
| **162.94** | 165.37 | -1.96 | 1.69 | 0.47 | -2.43 |
| **162.41** | 164.69 | -1.42 | -0.09 | 0.86 | -2.28 |
| **163.2** | 165.62 | -2.2 | -2.55 | 0.22 | -2.42 |
| **161.39** | 162.29 | -0.38 | 0.233 | 0.52 | -0.9 |
| **159.34** | 157.79 | 1.68 | 0.1 | 0.13 | 1.55 |
| **161.49** | 162.02 | -0.46 | -0.1 | 0.07 | -0.53 |
| **161.67** | 163.42 | -0.63 | 0.48 | 1.12 | -1.75 |
| **161.09** | 161.21 | -0.04 | -1.63 | 0.08 | -0.12 |
| **161.62** | 165.35 | -0.56 | -0.58 | 3.17 | -3.73 |
| **160.57** | 160.53 | 0.5 | -0.1 | 0.46 | 0.04 |
| **160.58** | 160.76 | 0.5 | 0.34 | 0.68 | -0.18 |
| **164.18** | 161.54 | -3.09 | 3.71 | -5.73 | 2.64 |
| **164.64** | 161.75 | -3.54 | 5.84 | -6.43 | 2.89 |
| **169.33** | 169.9 | -8.22 | 3.48 | -7.65 | -0.57 |
| **171.63** | 173.77 | -10.51 | -0.06 | -8.37 | -2.14 |
| **171.99** | 175.18 | -10.86 | -1.05 | -7.67 | -3.19 |

### 4. Conclusion

This paper uses the historical closing price of stock as time series data, constructs the ARIMA-SVM model, and predicts the closing price of stock in the future. However, this paper also has some shortcomings: it only evaluates the fitting effect of the prediction model from the perspective of relative errors, and it will be better if it can predict and evaluate the future trend of stock price changes. Of course, stock price prediction itself cannot form a complete investment decision, and at least risk assessment and corresponding risk control methods are needed.

## Acknowledgements

This paper was supported by the JiangSu province higher education education reform research project, 2017, project number: 2017JSJG218 and Postgraduate Research & Practice Innovation Program of Jiangsu Province: KYCX18_1326.

## References

[1] Yuxia Wu, Xin Wen. (2016) Short-term stock price prediction based on ARIMA model. Statistics and decision-making.23, 83-86.

[2] Jun Deng. (2017) Analysis of stock price fluctuation based on GARCH-SVM model. Journal of economic research.6, 56-57.

[3] Yanjie Shi. (2005) Stock market prediction method based on support vector machine (SVM). Statistics and decision-making. 4, 123-125.

[4] Lijun Feng, Shuquan Li. (2005) Research on risk identification method of construction project based on SVM. Journal of management engineering. 19(s1), 11-14.

[5] Lu C J, Lee T S, Chiu C C. (2009) Financial time series forecasting using independent component analysis and support vector regression. Decision Support Systems. 47(2), 115-125.

[6] Box G E P, Jenkins G M, Reinsel G C. (1976) Time series analysis: Forecasting and control. Rev. ed. Journal of Time. 31(4), 238-242.

[7] V.N. Vapnik. (1995) The Nature of Statistical Learning Theory. Springer, New York, 8 (6), 988 – 999.

[8] Xiuqin Li, Manfa Liang. (2013) Stock market prediction based on ARIMA model. Journal of changchun Education University. 29(14), 51-53.

[9] Zhaoyue Hu, Yanping Bai. (2016) Stock price prediction based on PCA-SVM portfolio model. Shang. 2, 206-206.

[10] Chunxue Wu. (2018) Stock forecasting methods based on SVM and stock price trend. Software guide. (4).