

AN INTEGRATED MODEL FOR STOCK PRICE PREDICTION BASED ON SVM AND ARMA

¹ZUOQUAN ZHANG, ²MENGGANG LI, ³RONGQUAN BAI

¹ School of Science, Beijing Jiaotong University, Beijing 100044, P.R.CHINA

² School of Economics and Management, Beijing Jiaotong University, Beijing 100044, P.R.CHINA

³ School of Science, Beijing Jiaotong University, Beijing 100044, P.R.CHINA

E-mail: ¹zqzhang@bjtu.edu.cn, ²mgli@bjtu.edu.cn, ³452046532@qq.com

The article is sponsored by the Reseach on China's Industrial Safety Index (B09C11010020)

ABSTRACT

Changes in stock price will be influenced by many aspects of factors. When we are predicting stock price, it is difficult to build a determined mathematical model between stock prices and these complex factors. This paper first utilizes ε -SVM (ε -support vector machine) to build a stock price prediction model. By fitting the prediction error sequence, we find the law factors, which the prediction model could not include or failed to give sufficient explanation, have a lasting influence on the stock price according to the ARMA theory of the time series analysis. Through an analysis we can predict the changes in the next step of this unexplained impact. Thus we revise the SVM model in the previous step. And the final results are obtained. The empirical analysis proves: ε -SVM model has a high prediction accuracy to predict the stock price. Moreover, after the ARMA error correction model is used, the prediction accuracy is further improved.

Keywords: Stock price, ε -SVM, ARMA

1. INTRODUCTION

Traditional stock price prediction models, such as random walk model, the unit root model and the later famous ARCH model, are utilizing statistical measurement approach as support [1]. They have a relatively long history, solid theory foundation, and a simple structure. And whether types of model or computer-aided tools (spss, matlab, eviews and other related software) have been quite perfect, they are relatively easy to be carried out. But they have the following disadvantages inevitably: (1) Linear regression. Most of the modeling approaches of traditional stock price prediction models use linear regression approach. But factors of affecting stock price are complex and can't be predicted. So it can't be a precise mathematical model and a simple linear model. (2) Large sample. The traditional stock price prediction models are based on some assumptions of mathematical statistics. The models require the lots of sample data and a very long time span. Most of the long-term disciplines which models sum up are not good at predicting short-term behavior of market. However, what the investors pay attention to are mostly short-term stock price fluctuations. (3) Single factors. As

previously seen, the traditional models are mostly single-factor model that can't add influence factors or it will bring out the quantity of explosive increase and have to sacrifice lag order number after adding factors. However, the reality is that the fluctuation of the stock price can't be explained by one or two factors.

1.1 SVM

SVM is based on the VC dimension theory of statistical learning theory and the principle of structural risk minimization. According to the limited sample information, we find a best compromise solution between the complexity of model (Learning precision to specific training sample, Accuracy) and learning ability (Error-free ability to identify any sample) to obtain the best promotion ability (Also known as generalization)[3].

1.2 ARMA

1.2.1 Autoregressive Model AR (p)

P-order autoregressive model is written as AR (p), and satisfies the following equation:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + a_t, \quad (1)$$

where, parameter c is a constant; ϕ_i is the coefficient of the autoregressive model; p is the order of the autoregressive model; a_t is a white noise sequence whose mean is 0 and variance is σ^2 .

1.2.2 Moving Average Model MA (q)

Q-order moving average model is written for MA (q), and satisfies the following equation:

$$X_t = \mu + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \dots + \theta_q a_{t-q}, \quad (2)$$

where, parameter μ is a constant; θ_i is the coefficient of the moving average model; a_t is a white noise sequence whose mean is 0 and variance is σ^2 .

1.2.3 ARMA (p, q) Model

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \dots + \theta_q a_{t-q} \quad (3)$$

Obviously, this is the combination of the above models, which is known as Hybrid model and written as ARMA (p, q) [4].

2. EMPIRICAL RESEARCH

2.1 \mathcal{E} -SVM Modeling

The stock transaction data of Industrial and The stock transaction data of Industrial and Commercial Bank of China (sh601398) used in this paper is taken from Super View stock analysis software. The tool used in this part is LIBSVM package for SVM research developed by Professor Lin Chih-Jen from Taiwan University, version 2.90. Selecting the Sigmoid kernel function is the most appropriate. The predictive values compared with the real value are shown in Figure 1.

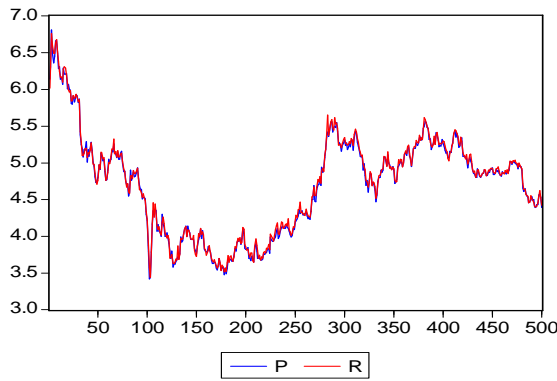


Figure 1. The predictive values compared with the real value

Where, P denotes the real closing price; R denotes the predictive result of the model.

2.2 The ARMA Model of the Error Sequence

The tool used in this part is the econometric software Eviews 5.0. The model's errors of the test set that has 500 samples are shown in Figure 2.

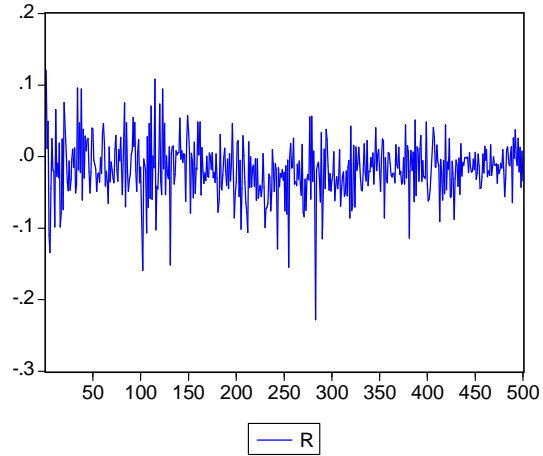


Figure 2. The Model's Errors Of The Test Set

Obviously, it is a time serial. And it can be written as $\{r_i\}$, $i = 1, 2, \dots, 500$. First, we do the center of treatment to the sequence to get $\{r'_i\}$. By Pandit-Wu modeling we fit its ARMA (2n, 2n-1) model based on the reality that it is a stationary serial [5, 6].

From the test results we know, using ARMA (6, 5) model to fit the error serial is appropriate.

2.3 Integrated Prediction

First, we take the data (opening prices, high prices, low prices, turnovers) of the 20 trading days from May 19, 2010 to June 18, 2010 as input vectors into the SVM model determined by 2.1. And we can get these 20 trading days' closing prices output sequence of prediction $\{y'_i\}$, $i = 1, 2, \dots, 20$. We compare $\{y'_i\}$ with the real value and we can get the error sequence $\{r_i^*\}$, $i = 1, 2, \dots, 20$. Because of the continuity of the time, it's proper that we merge $\{r_i^*\}$ to $\{r_i\}$ (putting r_i^* to the 501-520 terms of r_i according to the time serial). Thus we obtain the $\{r_i\}$, $i = 1, 2, \dots, 520$.

Then we make use of SPSS to fit the terms from the first 500 to the first 519 of $\{r_i\}$. And we can get



their ARMA (6, 5) models respectively. Utilizing of these models we predict the SVM model's error of the following trading day. We obtain the predictive value serial that there may be some errors of SVM's prediction from May 19, 2010 to June 18, 2010, and it's written as $\{e_i\}, i = 1, 2 \dots 20$.

At last, We add $\{y_i'\}$ and $\{e_i\}, i = 1, 2 \dots 20$, and we get the final predictive closing prices of these 20 trading days from May 19, 2010 to June 18, 2010 as Table 1.

Table 1. The Final Predictive Closing Prices

Date	Real closing price	Predictive closing price of SVM	Predictive error of ARMA	Final predictive closing price
2010-5-19	4.39	4.43858	-0.00864	4.429939
2010-5-20	4.41	4.4551	-0.01056	4.444535
2010-5-21	4.46	4.43237	-0.00605	4.426317
2010-5-24	4.55	4.55516	-0.01354	4.541624
2010-5-25	4.50	4.54082	-0.00583	4.534994
2010-5-26	4.52	4.54305	-0.00718	4.535866
2010-5-27	4.37	4.32165	-0.0142	4.307447
2010-5-28	4.33	4.35493	-0.01239	4.342539
2010-5-31	4.26	4.29399	-0.00831	4.285684
2010-6-1	4.23	4.24936	-0.01507	4.23429
2010-6-2	4.21	4.19917	-0.0118	4.187367
2010-6-3	4.22	4.23256	-0.01445	4.218108
2010-6-4	4.22	4.21557	-0.01057	4.204998
2010-6-7	4.12	4.14117	-0.01256	4.128607
2010-6-8	4.10	4.11278	-0.01054	4.102243
2010-6-9	4.20	4.1946	-0.01499	4.179614
2010-6-10	4.14	4.14944	-0.01914	4.130301
2010-6-11	4.17	4.18786	-0.0157	4.172163
2010-6-17	4.23	4.23064	-0.0166	4.214039
2010-6-18	4.17	4.22973	-0.01331	4.216419

Where, MAE has an 11.52% reduction; RMSE has an 8.79% reduction.

2.4 Analysis Of Results

By calculating we know that mean square error (MSE) is 0.000839056, and squared correlation coefficient R is 0.966657 when we utilize of SVM model only. It shows that the model's predictive ability has reached a considerable accuracy.

After we make use of the SVM-ARMA model to predict the closing prices, compared with the results before we can get the following results as Table 2.

Table 2. The Predictive Error

Evaluating indicator	SVM	SVM-ARMA
Mean absolute error	0.023591	0.020873
Root-mean-square error	0.028968	0.026421



3. CONCLUSIONS

Changes in stock price will be influenced by many aspects of factors. When we are predicting stock price, it is difficult to build a determined mathematical model between stock prices and these complex factors. There are two reasons: on the one hand, this relationship is a very complex non-linear relationship; on the other hand, because of these factors alone can't be completely understood, any model can't contain all the factors.

As to the former, this paper chooses a new machine learning approach-SVM. SVM is based on the VC dimension theory of statistical learning theory and the principle of structural risk minimization. According to the limited sample information, we find a best compromise solution between the complexity of model and learning ability to obtain the best promotion ability. This paper chooses ϵ -SVM model.

As to the second problem, this paper chooses the ARMA model of time series analysis. By fitting the prediction error sequence, we find the law factors, which the prediction model could not include or failed to give sufficient explanation, have a lasting influence on the stock price according to the ARMA theory of the time series analysis. Through an analysis we can predict the changes in the next step of this unexplained impact. Thus we revise the SVM model in the previous step. And the final results are obtained.

As to the empirical research, we take the Industrial and Commercial Bank of China (ICBC)'s data (opening prices, high prices, low prices, turnovers) of the 100 trading days as input vectors and closing prices as outputs. And we predict the following 500 trading days' closing prices utilizing ϵ -SVM model we get. Its MSE is 0.00183803 and R is 0.996671. It has achieved good results. After fitting the 500 error terms' ARMA model, we revise the predictive results. It makes that MAE has an 11.52% reduction and RMSE has an 8.79% reduction. So the model's predictive ability improves.

REFERENCES:

- [1] D.H. Jin, "Study on the fluctuation and control of Chinese stock market", *Shanghai University of Finance & Economics Press*, 2003.
- [2] J. Tian, "Stock market prediction model based on SVM", *Beijing Jiaotong University*, 2010.

- [3] N. y. Deng and Y.J. Tian, "Support vector machine---theory", *algorithm and development, Science Press*, 2009.
- [4] T. Jiang, "Arch model in Chinese stock market: an empirical research", *Hunan University*, 2007.
- [5] T. M. Gao, "Econometric analysis and modeling: applications and examples of EViews", *Tsinghua University Press*, 2006.
- [6] Z. L. Wang and Y. H. Hu, "Applied time series analysis", *Science Press*, 2007.