

Financial Time Series Prediction Based on EMD-SVM

Kai Tang

Department of Software Engineering, Tongda College of Nanjing University of Posts and
Telecommunications, Zhenjiang, China

1328590135@qq.com

Abstract. The market plays an essential role in the national economy and society, and people pay more and more attention to the investment in the capital market. Every valuable investment needs the guidance of scientific theory. With the improvement of China's securities market system, more and more researchers have conducted in-depth research focusing on the development of the stock market. This paper will study the stock price prediction algorithm, select the Bank of Ningbo as the research object, and propose a prediction model based on the combination of EMD and SVM algorithms. This paper first decomposes the original sequence by EMD and comprehensively considers the three related factors of the Shanghai Composite Index, Shenzhen Composite Index, and China Merchants Bank in the prediction model. Then, the SVM model is used to predict multiple decomposition sequences, respectively, and the prediction results are obtained after integrated processing. Then, regression analysis is carried out on them, and the weight is designed according to the constructed regression analysis model. The final prediction result is obtained after reconstruction. Compared with the basic SVM model, the combined prediction model constructed in this paper has a better prediction effect, and the prediction results are more objective and scientific.

Keywords: EMD; SVM; Financial time series; Combined forecasting.

1. Introduction

Today, the market plays an important role in the national economy and society, and people pay more attention to capital market investments. Every valuable investment needs the guidance of scientific theories, such as understanding the box of stock price volatility, the reasonable price, the potential and future of stocks, etc. Many securities analysts and investment advisers are keen to predict stock prices and can give investment suggestions such as recommended ratings and target stock prices. With the proper investment guidance, investors can invest with confidence. Stock price forecasting plays an important role in guiding funds, which is conducive to the rational allocation of capital, the delicate operation of the capital market, and the reduction of financial risks. Therefore, stock price forecasting has become an important research direction and has great research value. This article is produced under this background, which provides some help to meet this demand.

Meanwhile, the stock market prediction at home and abroad has a long history. Many researchers continue to introduce various models and parameters to study, which makes the research on stocks mature gradually. However, the stock price is affected by many factors, such as economic situation, political policy, social environment, market investment sentiment, etc. At the beginning of the study, researchers only regarded stock data as traditional time series [1-3]. The technology and research methods of studying stock price data through time series have been perfect. However, because many factors affect the stock price and are often unstable, the previous stock forecasting methods have considerable limitations. As people enter the era of big data analysis and intelligence, machine learning has become the hottest and most popular research direction, and the support vector machine model has also achieved remarkable results in practical application.

Stock price forecast has been paid great attention for a long time. Prediction methods can be divided into different and causal models. The difference model is mainly the ARIMA model [4-6] and grey GM (1, 1) [7-9] model, and the causal model mainly includes neural network and multiple linear regression model [10-11]. If ARIMA and GM (1, 1) models are used to predict time series data, they must be stable. If the data is unstable, the law cannot be captured. For example, stock data cannot be predicted is that stock data is unstable and often fluctuates under the influence of policies and news.

For the causal relationship model, if you want to predict the dependent variable, you need to predict the independent variable first, but sometimes it may be more challenging to predict the independent variable than the dependent variable.

The EMD-SVM algorithm in this paper is suitable for predicting unstable time series. It is less affected by noise and can predict unstable sequences. The above factors make up for the defects of the four algorithms in the last paragraph. In addition, the existing algorithms do not fit the prediction results, and there are some errors in the prediction results. The EMD-SVM algorithm in this paper has achieved this very well.

The financial time series data itself has great volatility. The whole time series is permeated with high-frequency and low-frequency data. EMD [12-14] decomposes the signal according to the time scale characteristics of the data itself. SVM is a generalized linear classifier that classifies data in a supervised learning way. Its decision boundary is the maximum margin hyperplane for solving the learning samples. Therefore, this paper uses the combined model based on EMD-SVM to predict the financial time series [15]. Taking the financial time series (Bank of Ningbo from 2007 to 2022) as the research object, this paper will predict the financial time series based on EMD-SVM. The idea is to decompose the time series by EMD-SVM to establish the prediction model and then use EMD reconstruction technology to obtain the prediction value of the original time series.

2. Methodology

2.1 EMD

EMD can decompose any signal into the sum of several IMF and a remainder [11]. Generally, signals or functions meeting the following two requirements are defined as IMF:

(1) In all statistical series, the total amount of extreme points (including maximum points and minimum points) and the total amount of zero-crossing points should be equal, or at most similar to some;

(2) At any time, the average number of the upper envelope points defined in the local overall maximum of the signal and the lower envelope points defined in the overall local minimum is zero.

The specific steps of EMD decomposition are as follows:

(1) Assuming that the signal is $x(t)$, and it takes the sequence composed of the local mean of its upper and lower envelopes as $m_1(t)$, then:

$$h_1(t) = x(t) - m_1(t) \quad (1)$$

For non-linear and non-uniform data, one processing is usually insufficient to form IMF, and some asymmetric waves may still appear. At this time, regard $h_1(t)$ as the data is to be processed, repeat the above operation k times, and you can get:

$$h_k(t) = h_{(k-1)}(t) - m_k(t) \quad (2)$$

When the IMF conditions are met, the first IMF is obtained, which is recorded as $f_1(t) = h_k(t)$.

(2) Separate the first IMF from the signal and the remaining signal $r_1(t)$ is:

$$r_1(t) = x(t) - f_1(t) \quad (3)$$

(3) As the signal is to be decomposed, repeat the steps of equations (1) to (3), and then decompose it in turn to obtain:

$$\begin{cases} r_2(t) = r_1(t) - f_2(t) \\ r_3(t) = r_2(t) - f_3(t) \\ \vdots \\ r_n(t) = r_{n-1}(t) - f_n(t) \end{cases} \quad (4)$$

Until the remaining information has little impact on the research content, it becomes a monotonic function, and IMF can no longer be selected. So far, the signal has been decomposed into n IMF, that is, the sum with a remainder:

Until the information in the remaining $r_n(t)$ has little impact on the research content, or it becomes a monotonic function, and IMF can no longer be selected. So far, the signal $x(t)$ has been decomposed into n IMFs, that is, the sum of $f_i(t)$, ($i = 1, 2, \dots, n$) and a remainder $r_n(t)$:

$$x(t) = \sum_{i=1}^n f_i(t) + r_n(t) \quad (5)$$

Equation (5) shows that the EMD decomposition of the signal has complete properties, which are determined by the analysis process itself.

2.2 SVM

SVM is one of the most used machine learning classification algorithms. Support vector machine (SVM) solves the linear indivisibility of original data sets in multidimensional space by mapping multidimensional space to higher dimensions. The main idea is to find the $N-1$ -dimensional hyperplane in the n -dimensional space so that the distance between different sample points and the hyperplane is the largest. SVM minimizes structural risk and the confidence interval on the premise of controlling empirical risk.

Assume that the training set is $(x_i, y_i) \in D$, set the error bandwidth and penalty function, and transform the maximization interval problem into an optimization problem:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s. t. } & y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \dots, m \end{aligned} \quad (6)$$

Where w and b are the two parameters of the discriminant function, ξ_i represents the relaxation variables of the sample points i , and C is the penalty coefficient of the relaxation variables. The kernel function projects the original feature data into the high-dimensional space. The required segmentation hyperplane can be obtained by solving equation 2.3, and then the corresponding label can be predicted by inputting the characteristics of the test set.

2.3 Prediction model of Bank of Ningbo based on EMD-SVM

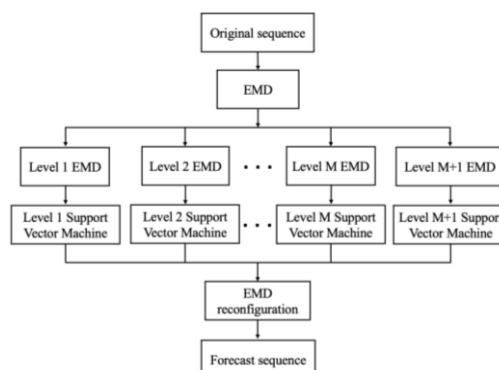


Fig. 1 EMD-SVM algorithm flow chart

The EMD-SVM decomposition prediction algorithm in this paper is different from the basic EMD-SVM algorithm, and its flow chart is shown in Figure 1. This paper cites three influencing factors and makes multiple regression to make the prediction more accurate and reasonable. The flow chart of the improved SVM decomposition algorithm in this paper is shown in Figure 2. below.

However, this paper improves based on EMD-SVM and introduces three related factors: the Shanghai Composite Index, the Shenzhen Composite Index, and the China Merchants Bank, which improves the accuracy and preciseness of prediction. The steps of the EMD-SVM algorithm in this paper are as follows:

- (1) EMD decomposition of the original sequence;
- (2) Using the Shanghai stock index, Shenzhen stock index, and China Merchants Bank to predict each decomposition sequence by SVM;
- (3) After reconstruction, the prediction sequences under their respective conditions are obtained, and they are analyzed by regression analysis respectively;
- (4) Finally, according to the original series' three and the relevant factors, the final prediction series is combined according to the proportion.

The EMD-SVM algorithm flow chart in this paper is shown in Figure 2. below:

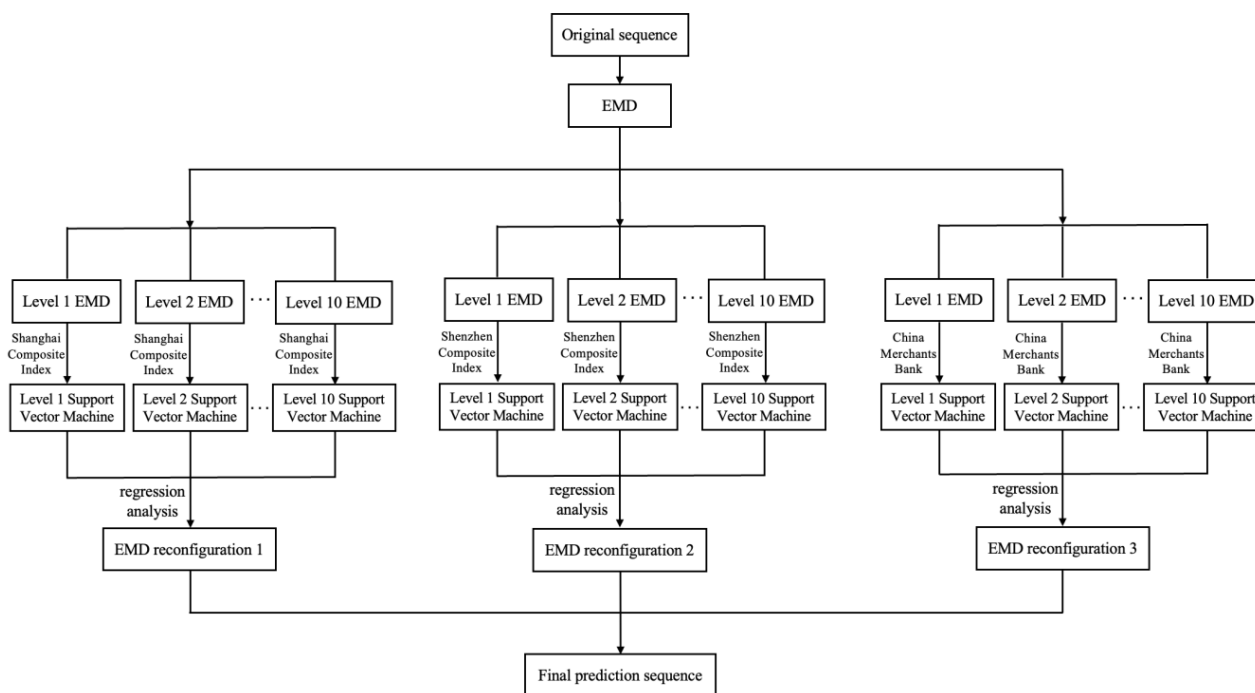


Fig. 2 EMD-SVM algorithm flow chart of this paper

3. Empirical analysis

3.1 Data acquisition and preprocessing

The logarithm yield of the daily closing price of Bank of Ningbo from July 20, 2007 (the second day after the listing of Bank of Ningbo) to February 18, 2022, in the sample of this model, excluding holidays and suspension, totals 3524 data and takes the first 70%, that is, the first 2467 data, as the initial value to establish a prediction model to predict the next 1056 values. Data source: Netease financial database.

Three factors related to the share price of the Bank of Ningbo are selected to join the SVM prediction model. This paper selects the Shanghai stock index, Shenzhen stock index, and China Merchants Bank. The reason for choosing the three influencing factors from the subjective logic is: that the Bank of Ningbo is the stock of the Shenzhen composite index, and the first thing to consider is Shenzhen composite index. At the same time, as a powerful and important sector, the trend of the

market is closely related to it, so the Shanghai Stock Exchange index was selected. Bank of Ningbo and China Merchants Bank are two major banks with relatively good equity and prosperous performance in the banking sector. They are located in the same sector and have a close linkage. Therefore, China Merchants Bank is selected. These three factors will make up for the impact of performance and investment sentiment.

Table 1. Correlation between the Share Price of Bank of Ningbo and Various Factors

Correlation between the amplitude of Bank of Ningbo and various factors			
factor	The amplitude of the Shanghai Composite Index	The amplitude of the Shenzhen Component Index	The amplitude of China Merchants Bank
correlation coefficient	0.562	0.608	0.928

Table 2. Correlation between the Share Price of Bank of Ningbo and Various Factors after Improvement

Correlation between the stock price of Bank of Ningbo and various factors after improvement			
factor	Shanghai Composite Index	Shenzhen Component Index	China Merchants Bank
correlation coefficient	0.995	0.997	0.987

3.2 Decomposition results based on EMD

The original data is decomposed into $imf1\sim9$ and a residual value res . By EMD, and their relationship is:

$$signal = \sum_{k=1}^9 imf\ k + res. \quad (7)$$

The decomposition results are shown in Figure 3 below.

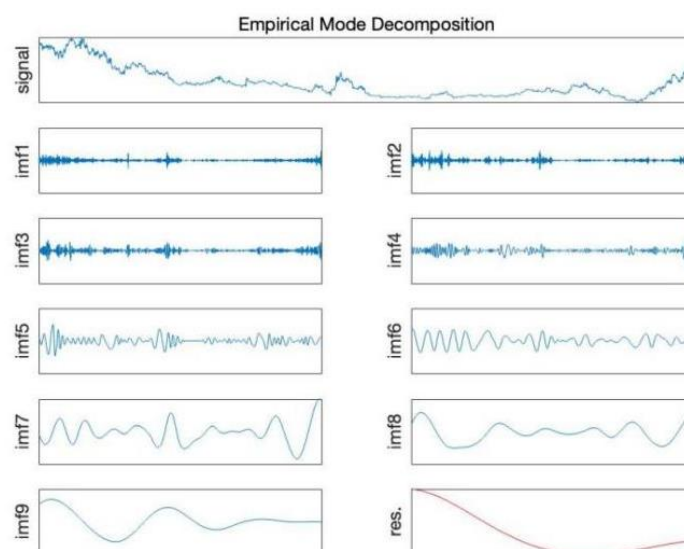


Fig. 3 EMD decomposition sequence

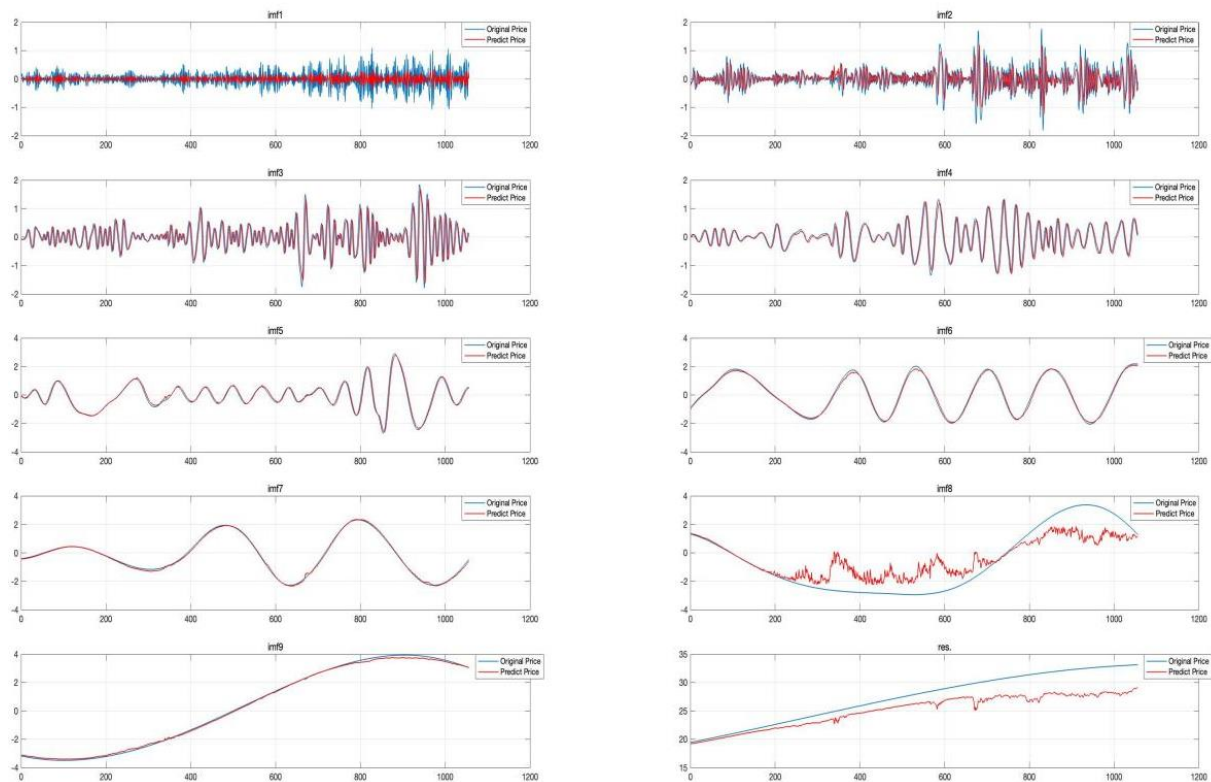


Fig. 4 EMD-SVM prediction of imf1 ~ 9 and res. under Shenzhen Composite Index

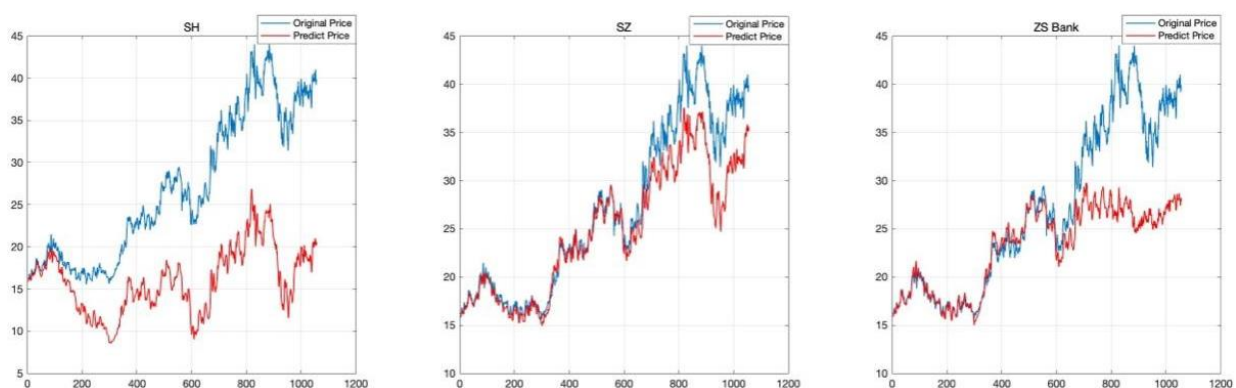


Fig. 5 EMD-SVM prediction under three related factors

Next, SVM prediction is required for each IMF and res. Here, the Shenzhen composite index is taken as an example. Figure 4 shows the EMD-SVM prediction of imf1 ~ 9 and res. under the Shenzhen Composite Index. After EMD reconstruction, the EMD-SVM prediction under the Shenzhen Composite Index is obtained.

Similarly, after EMD reconstruction, the EMD-SVM prediction results under Shanghai Stock Exchange Index and China Merchants Bank are also obtained. The EMD-SVM prediction results under the three related factors are shown in Figure 5.

After EMD Analysis of the original series, SVM prediction is carried out for each decomposed series by using Shanghai Composite Index, Shenzhen Composite Index, and China Merchants Bank. After reconstruction, the prediction series under their respective conditions can be obtained, but there is still room for improvement in the current prediction effect. Therefore, this paper considers their regression analysis to obtain the fitted Shanghai Composite Index, Prediction series under Shenzhen Composite Index, and China Merchants Bank. Finally, according to the original series' three and the relevant factors, the final prediction series is combined according to the weighted average of the proportion.

Figure 6 shows the EMD-SVM prediction results of the Shanghai Composite Index, Shenzhen Composite Index, and China Merchants Bank.

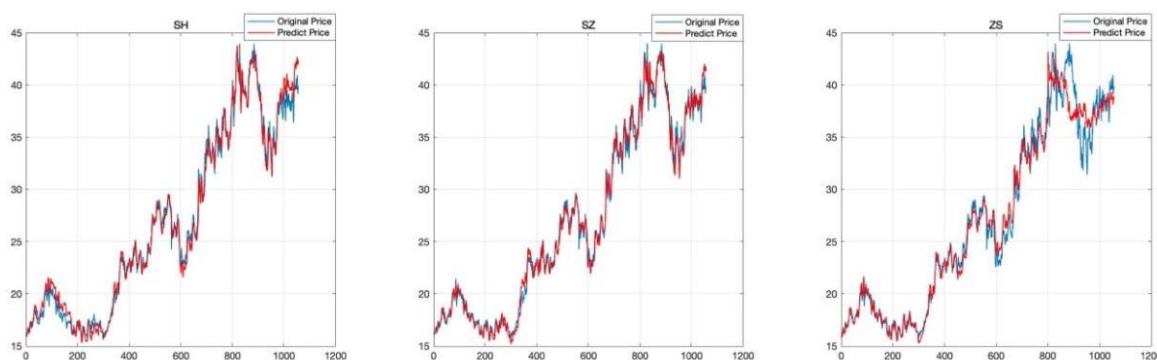


Fig. 6 EMD-SVM prediction results after fitting under three related factors

3.3 Comparison and analysis of prediction results

The prediction results using only SVM are shown in Figure 7, and the prediction effect is not very good. The final prediction result of EMD-SVM is shown in Figure 8. The prediction is relatively accurate and fits the original sequence very well.

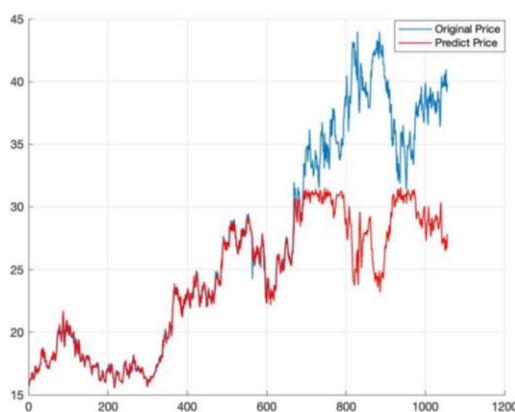


Fig. 7 Prediction results using SVM only

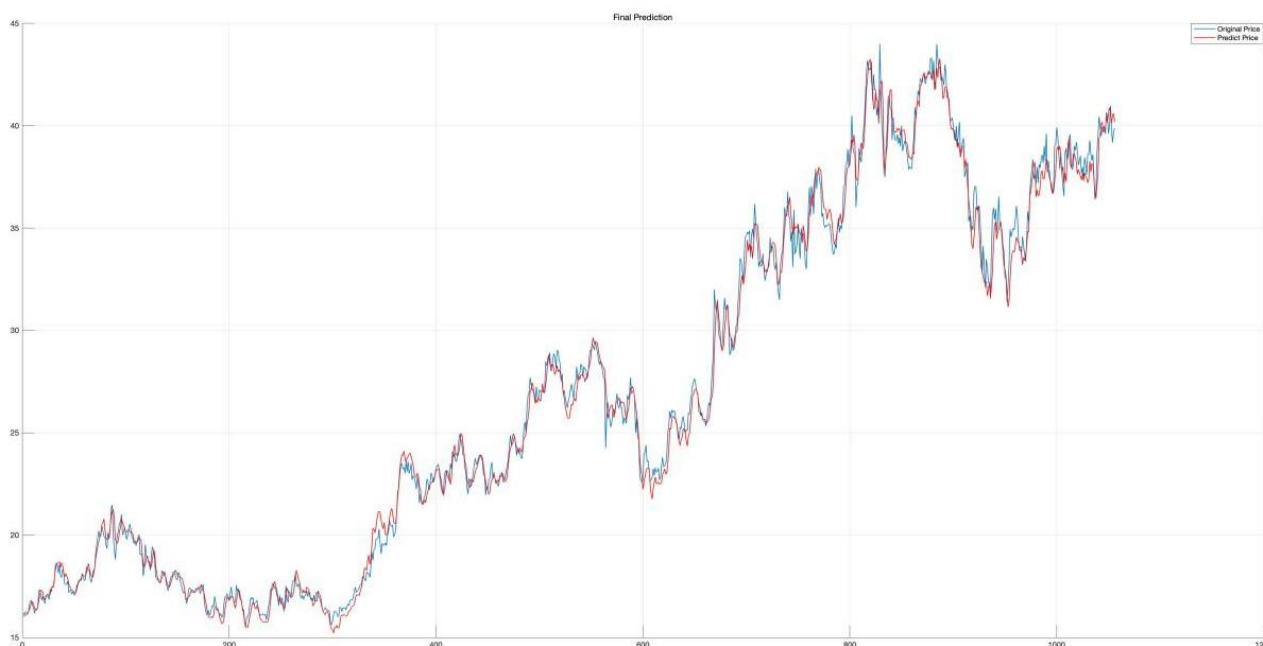


Fig. 8 Final EMD-SVM prediction results

Next, four error indicators Mae, RMSE, MAPE, and R2, will be used to analyze the error of the prediction results of SVM and EMD-SVM.

It is assumed that the predicted value is $\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$, and the true value is $y = \{y_1, y_2, \dots, y_n\}$.

3.3.1 RMSE

RMSE (Root Mean Square Error) formula is shown in (8), which represents the expected value of the square of the error. Its value range is $[0, +\infty)$. It is used to compare the prediction errors of different models of a specific dataset rather than the prediction errors between datasets because it is related to the proportion. Generally, a lower RMSE is better than a higher RMSE.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

3.3.2 MAE

MAE (Mean Absolute Error) formula is shown in (9). Its value range is $[0, +\infty)$. When the predicted value is completely consistent with the real value, it is equal to 0, that is, the perfect model; the larger the error is, the larger the value is—the smaller the Mae value, the better the prediction model's accuracy.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (9)$$

3.3.3 MAPE

The MAPE (Mean Absolute Percentage Error) formula is shown in (10). Its value range is $[0, +\infty)$. A MAPE of 0% indicates a perfect model, and a MAPE greater than 100% indicates a poor model. The smaller the MAPE value, the better the accuracy of the prediction model.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (10)$$

3.3.4 R²

R² (R-Square) formula is shown in (11), where the numerator represents the sum of the square differences between the real value and the predicted value, similar to the mean square error MSE; The denominator part represents the sum of the square difference between the true value and the mean, similar to variance. Judge the quality of the model according to the value of the R-squared. The value range is $[0, 1]$. If the result is 0, the model fitting effect is inferior; If the result is 1, there is no error in the model. Generally speaking, the larger the R-squared, the better the fitting effect of the model.

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (\bar{y}_i - y_i)^2} \quad (11)$$

Table 3. Error Indicators of SVM and EMD-SVM

Error Indicators of SVM and EMD-SVM				
Model	RMSE	MAE	MAPE	R ²
SVM	5.7544	3.1194	11.5525%	0.5463
EMD-SVM	0.6219	0.4606	1.7208%	0.9947

The error results are shown in Table 3. Through comparison, the error-index of the prediction model constructed in this paper is small.

The stock price is affected by many factors, such as performance, investment sentiment, etc. The financial time series is unstable and irregular. Then, the traditional SVM algorithm can not effectively consider many kinds of influencing factors. It has certain limitations and large errors for the intermittent and highly volatile financial time series prediction. It can be seen from figure 3.8 that the EMD-SVM model has an excellent fitting degree, which greatly fits the original sequence. The correlation coefficient is as high as 0.997, and the prediction result is good with a high fitting degree. Compared with the traditional SVM algorithm, the EMD-SVM algorithm constructed in this paper can effectively improve the prediction effect.

In this paper, SPSS software is further used for correlation analysis. After the research, the correlation coefficient between the sequence predicted by SVM and the original sequence is 0.85, and the correlation coefficient between the Shanghai Stock Exchange Index and the Bank of Ningbo is 0.562. After fitting, the correlation coefficient is increased to 0.995. The correlation coefficient between the Shenzhen Composite Index and the Bank of Ningbo is 0.608. After fitting, the correlation coefficient is increased to 0.997. The correlation coefficient between the Shenzhen Composite Index and the Bank of Ningbo is 0.928. After fitting, the correlation coefficient is increased to 0.987. The correlation coefficient between the final prediction series and Bank of Ningbo is as high as 0.997, which is in good agreement with the original series. The feasibility of applying the EMD-SVM algorithm to stock price forecasting is verified. Compared with the traditional SVM algorithm, the prediction effect of this algorithm is better.

4. Conclusion

To aim at the problem of stock price forecasting with substantial volatility in financial time series, this paper constructs a combined forecasting model based on EMD-SVM, considers a variety of influencing factors in the model, forecasts the financial time series, and makes a practical analysis with SPSS data to verify the effectiveness of this model. The main conclusions of this paper are as follows:

(1) Financial series sometimes show the characteristics of instability and irregularity. The traditional SVM algorithm has certain limitations in processing such signals, and there are large errors in the prediction of financial series. (2) Stock data is a kind of time series with noise. EMD decomposition can effectively reduce data noise, greatly save information, and significantly improve the stock price prediction algorithm. (3) EMD-SVM algorithm has specific feasibility and practicability for predicting stock price. According to the actual situation, a specific EMD-SVM model is constructed.

This paper constructs a model framework based on the EMD-SVM stock price prediction algorithm, which provides the possibility and reference for the EMD-SVM algorithm to be widely used in stock price prediction in the future. In the future, if you want to predict other stocks or modify the corresponding model according to the current situation, you only need to replace the influencing factors and carry out the corresponding regression test.

As far as investment suggestions are concerned, the share price of Bank of Ningbo is a trend of the combination of band and box. Mastering and understanding it will help to invest in it; From the aspect of management suggestions, the prediction model is applied to the prediction of financial stock model, which has strong anti-noise interference ability and greatly improved prediction ability. It is suitable for stocks with strong volatility and can provide more effective suggestions for investors.

References

- [1] Chen L, Chi Y, Guan Y, et al. A hybrid attention-based EMD-LSTM model for financial time series prediction [C] // 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD). IEEE, 2019: 113-118.
- [2] Zhaoguo Shun. Research on stock price trend prediction based on time series analysis [d] Xiamen University, 2009.

- [3] Silvio Barra, Salvatore Mario Carta, Andrea Corrigan, Alessandro Sebastian Podda, Diego Reforgiato Recupero. Deep Learning and Time Series-to-Image Encoding for Financial Forecasting [J]. IEEE/CAA Journal of Automatica Sinica, 2020, 7(03): 683-692.
- [4] Liu Song, Zhang Shuai. An empirical study on stock price prediction using ARIMA model [j] Economic Research Guide, 2021 (25): 76-78.
- [5] Chen Bowen. Prediction of stock price based on technical indicators and ARIMA Model -- Taking the adjusted closing price of Ping An insurance group of China as an example [j] Statistics and management, 2021, 36 (7): 53-57.
- [6] Chenxiaoling. Stock price prediction based on ARIMA model and neural network model [j] Economic mathematics, 2017, 34 (4): 30-34.
- [7] Kang Jianwei, Su Baoping. Application of Grey Prediction GM (1, 1) model in stock price prediction [j] Volkswagen technology, 2012, 14 (9): 20-21+23.
- [8] Liuchengle, linjinfu, linxiaoping. Using GM (1, 1) model to predict the stock price of iron and steel industry [c] // procedures of 2014 4th International Conference on Applied Social Science (icass 2014) volume 532014: 116-120.
- [9] Qin Siqian. GM (1, 1) model for stock price prediction [j] Statistics and decision making, 2006 (6): 22-23.
- [10] Van-Dai Ta, CHUAN-MING Liu, Direselign Addis Tadesse. Portfolio Optimization-Based Stock Prediction Using Long-Short Term Memory Network in Quantitative Trading [J]. Applied Sciences, 2020, 10(2).
- [11] Xiongwen Pang, Yanqiang Zhou, Pan Wang, Weiwei Lin, Victor Chang. An innovative neural network approach for stock market prediction [J]. The Journal of Supercomputing, 2020, 76(3).
- [12] Huang W, Shen Z, Huang N E, Fung Y C. Engineering analysis of biological variables: an example of blood pressure over 1 day. [J]. Proceedings of the National Academy of Sciences of the United States of America, 1998, 95(9).
- [13] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, Henry H. Liu. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis [J]. Proceedings: Mathematical, Physical and Engineering Sciences, 1998, 454(1971).
- [14] Huang, Norden E. Applications of Hilbert–Huang transform to non-stationary financial time series analysis. This article is a U.S. Government work and is in the public domain in the U.S.A. [J]. Applied Stochastic Models in Business & Industry, 2003.
- [15] Vapnik V N. Statistical Learning Theory (Adaptive and Learning Systems for Signal Processing, Communications and Control Series). 1998.