The construction and application of RBF-SVM-EL short-term stock price prediction algorithm

Zhiyuan Ma[1, #], Zeyu Zhang[2, #], Xue Zhu[3, #, *]

[1]School of Business Dalian University of Technology 116081, Dalian, China

[2]College of Architecture and Urban-Rural Planning Sichuan Agricultural University 611830, Chengdu, China

[3]Jiangsu University of Science and Technology 212000, Zhenjiang, China

*Corresponding author. Email: zhuxue220304@163.com
[#]Equally contributed to this works

Keywords: Machine learning, RBF, SVM, Bagging ensemble learning, short-term stock price forecast.

Abstract: Today, stock investing is increasingly becoming one of the main ways people invest in finance. Therefore, reasonable short-term stock price forecasts are of great significance to investment decisions. This paper combines radial basis functions (RBFs) and support vector machines (SVMs) with bagging ensemble learning (EL) to form a new ensemble model for predicting short-term stock prices from stock market indicators. This article considers the research object of the Shanghai Stock Exchange Index closing price from January 4, 2021, to January 31, 2021. The conclusions are as follows: (1) The Bagging ensemble learning model based on RBF neural network and SVM support vector machine has a correlation coefficient R2 closer to 1 and a small mean absolute error MAE and root mean square error compared with a single model. The fitting effect is good performance, the relative error of the prediction result is low, and it is more suitable for predicting short-term stock prices. (2) In short-term stock price forecasting, the SVM support vector machine has a better general forecasting effect than RBF neural network, its correlation coefficient R2 is closer to 1, its mean absolute error MAE and root mean square error RMSE is smaller, and the prediction results of the RBF neural network are unstable.

1. Introduction

The global financial crisis has also followed with the continuous improvement of world economic integration. The arrival of each financial crisis is accompanied by economic recession problems such as stock price decline and exchange-rate turbulence. The unpredictable world economy makes the stock market more turbulent and uneasy. People also need to bear huge risks while benefiting. The stock market gradually grows with the continuous development of the economy, but it is not perfect. The stock market is easy to be manipulated by major shareholders, and the stock price fluctuates greatly. The slightest cause investors to lose their homes and property and the listed companies to go bankrupt. The severest cause is the national economic turmoil. Therefore, short-term stock price prediction is essential, and risks cannot be avoided. However, the risk can be minimized to the greatest extent through reasonable stock price prediction, or the maximum benefit can be obtained in the same degree of risk to maximize the investment benefit.

Financial markets always cover complex nonlinear behavior, which greatly affects investment decision-making. In addition, the human brain cannot make decisions quickly in the face of vast amounts of transaction data. With the advent of big data, more and more people will focus on machine learning. The application of machine learning in quantitative investment can not only provide fuzzy processing of nonlinear relationships, but also use these algorithms to improve the speed of operation, greatly improve the efficiency of machine learning for data, extract adequate information from numerous financial data, and help quantitative investment strategies to obtain better portfolios.

Through reasonable quantitative investment strategy, to enhance the majority of investors' discipline, systematic, timeliness, and accuracy.

With the continuous development of the global economy and the opening and integration of financial markets in various countries, financial globalization has brought rapid economic development to various countries. However, at the same time, financial globalization also has excellent potential risks. China's comprehensive strength is increasing, the rapid development of the economy [1], the increasing number of idle funds of national residents, the stock market is booming, stock investment has become one of the main ways of contemporary Chinese financial management. Stock investment means to benefit but also to bear certain risks. In the short term, the use of relevant indicators to reasonably predict stock prices and trends has been a research hotspot of many scholars [2]. In 1952, Markowitz [3] proposed the mean-variance model, which laid a theoretical foundation for the use of linear models to predict stock returns. For the first time in history, he explained the relationship between the return on risky asset investment and the size of risk, representing the arrival of the era of quantitative investment. In 1976[4], Ross proposed the arbitrage pricing model. This study pointed out that due to factors such as GDP growth and inflation, there will be short arbitrage opportunities in the stock market, so the expected return of stocks is also affected by macro factors. In 1986, to use this program to send effective trading signals, Irwin et al. established an automated trading system [5], which provided relevant materials for the short-term operation of futures funds to help research and understand. In 1995, Vapnik proposed the SVM model, making up for the limitations of traditional linear prediction methods for linear system analysis [6]. Peng Lifang [7] and others used the SVM algorithm to predict the stock's closing price, combined with time series to build a regression model. The experimental results show that the prediction model combined with the machine learning algorithm has higher prediction accuracy than the traditional linear time series.

However, the above linear analysis method can only consistently estimate the established relationship between variables. Under the influence of emergencies in financial markets, the established relationship between financial variables will continue to change. The hypothesis of the linear model will make the stock return series data show a low signal-to-noise ratio and thick tail distribution, so the accuracy of prediction is low. In this paper, machine learning and stock price prediction are closely integrated. RBF neural network and SVM support vector machine is introduced, and Bagging ensemble learning is used to combine the two to form a new integrated model applied to short-term stock price prediction. In the aspect of data selection, this paper first collects the closing price of the Shanghai Stock Index from January 4, 2021, to January 31, 2021, as the research object, in which the daily closing price of the stock is the output data of the ANN-RBF and SVR model, and the input data of the ANN-RBF and SVR is the technical stock index, and the collected data are normalized. Secondly, the opening price, the highest price, the lowest price, the turnover, and the growth rate are selected to construct the RBF neural network and SVM support vector machine. The Bagging ensemble learning model of RBF and SVM is constructed to predict the stock's closing price. The results of the ensemble learning model are compared with the single results of RBF and SVM, and the generalization and robustness of the model are finally tested. The method used in this paper improves the accuracy of the prediction results and can better help investors make decisions.

2. Sub-model construction

2.1 RBF neural network

In 1988, Broomhead and Low introduced RBF (Radical Basis Function) into the neural network design according to the characteristic that biological neurons have a local response and produced RBF (Radical Basis Function), namely RBF neural network. In 1989, Jackson demonstrated the consistent approximation performance of the RBF neural network to nonlinear continuous functions. RBF neural network is a forward network with good performance, which has the performance of optimal approximation and overcoming the local minimum problem. RBF neural network is a kind of artificial

neural network (ANN) with three neural network layers, including input layer, hidden layer, and output layer. [8] See Figure 1.

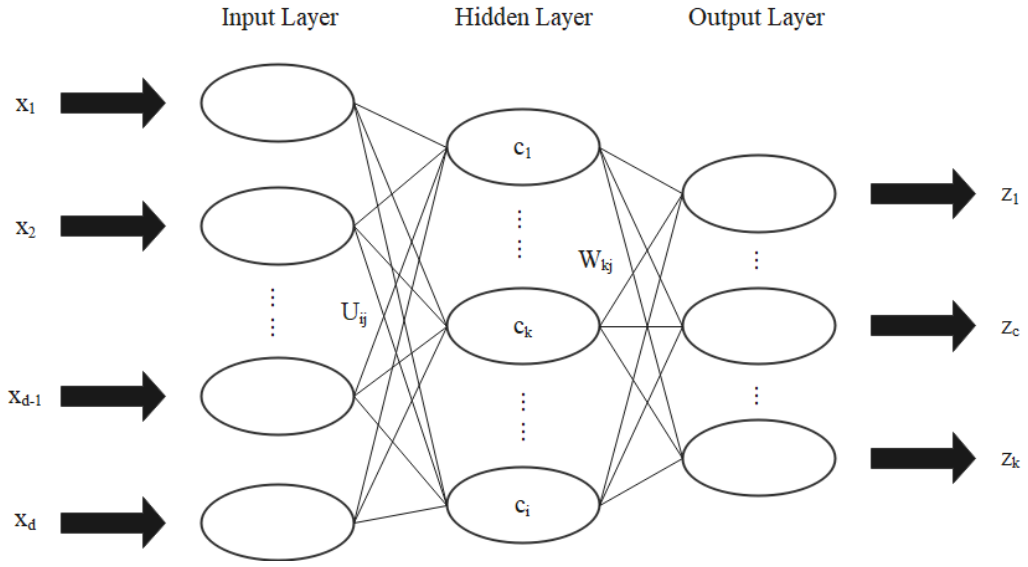

Figure 1. RBF neural network structure

The input layer is composed of signal source nodes, which play the role of signal transmission. The hidden layer is a nonlinear mapping of network input, and the mapping function is the radial basis function (RBF). The output layer is the linear weighted summation of hidden layer neurons.

The basic idea of the radial basis function neural network is to use the radial basis function as the hidden unit of the ' base ' to constitute the hidden layer space. The hidden layer transforms the input vector and the low-dimensional pattern input data into the high-dimensional space so that the linear inseparable problem in the low-dimensional space is linearly separable in the high-dimensional space. From the input signal of the input layer to the hidden layer, the basis function in the hidden layer node responds locally to the input signal. When the input signal approaches the central range of the basis function, the hidden layer node will produce a considerable output value, thus producing the effect of local approximation [9].

In general, Gaussian function is selected as the basis function, and its expression is:

$$R_i(x) = e^{-\frac{\|x-c_i\|^2}{2\sigma^2}} \tag{1}$$

c_i is the center of the i-th base function, and σ ^ 2 is the normalized parameter. In mapping the input layer to the hidden layer through the radial basis function, it is necessary to determine the network weight vector through the learning algorithm. The linear equation can express the output model of the network:

$$Y(x) = \sum w_{ij}\delta_j\left(\left\|x_p - c_i\right\|\right) \tag{2}$$

2.2 Support vector machine

Support vector machine (SVM) is a machine learning method proposed in 1995, which can realize nonlinear and high-dimensional problem analysis on the premise of small samples and solve the problems of structural selection and local minimum (over-fitting and under-fitting) of neural network. After years of development, SVM has many functions such as classification, recognition, and regression prediction. Moreover, great success has been achieved with the application of SVM regression prediction in the field of the financial economy.

Its mathematical expression [10]:

$$\min \frac{1}{2}\sum_{\mathrm{i,j=1}}^{n}(\overset{*}{a}_i - a_i)(\overset{*}{a}_j - a_j)K(x_i, x_j) + \varepsilon\sum_{i=1}^{n}(\overset{*}{a}_i - a_i) - \sum_{i=1}^{n}(\overset{*}{a}_i - a_i) \tag{3}$$

$$s.t.\begin{cases} \sum_{i=1}^{n}(\overset{*}{a}_i - a_i) = 0 \\ 0 \le a_i, a_i \le \dfrac{c}{n} \\ K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2} \\ i = 1, 2, \cdots, n \end{cases} \tag{4}$$

The training samples are xi, i = 1,2,3, n; k is a kernel function; we choose Gaussian radial kernel function; C is the penalty function, and the larger C is, the greater the penalty of the relative error ε is.

If the optimal solution is $a = (\overline{a}_1, \overset{*}{a}_1, \cdots, \overline{a}_n, \overset{*}{a}_n)^T$, The decision function of support vector machine regression is

$$f(x) = \sum_{i=1}^{n}(\overset{*}{a}_i - a_i)K(x_i, x) + \overline{b} \tag{5}$$

3. Bagging integration learning

3.1 Introduction to Bagging

The Bagging algorithm (Boostrap Aggregating), initially proposed by Leo Breiman, determines the prediction value by randomly sampling from the training data set, training the model with the extracted samples, and voting by all models together.

The main idea is to use multiple independent random samples from the initial dataset to produce multiple independent datasets. Given one or more weak learning algorithms, the generated multiple training sample sets are learned by this algorithm, and a prediction function hi is obtained after each training, and a total of T prediction functions are obtained in T training rounds. The sequence of sub-prediction functions is used to predict the sample sets, and then the final prediction results are obtained by majority voting or averaging.

The schematic diagram of Bagging is as follows:

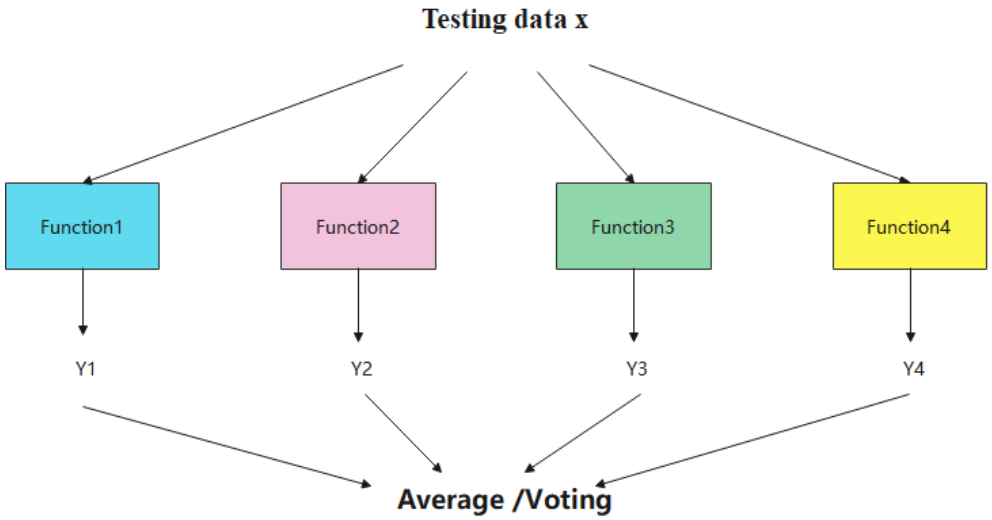

Figure 2. Bagging diagram

Bagging increases the diversity among learners through the perturbation of samples, so the learner of Bagging should be an unstable learning algorithm that is more sensitive to the training set, and the error of the RBF neural network SVM support vector machine in this paper is reduced through Bagging integrated learning to improve the prediction accuracy and stability.

The basic algorithm is as follows.

In the first step, the data are drawn from the original sample set Φ in a put-back manner. It is repeated B times to obtain B Bootstrap resampling data sets $\Phi_l^* = \{(x_{1l}^*, Y_{1l}^*), \cdots, (x_{nl}^*, Y_{nl}^*)\}, l = 1, \cdots, B$

In the second step, one model has obtained for each Bootstrap resampled dataset. The results in a total of B models for classification or regression.

In the third step, the B models are integrated. For the classification problem, the B models obtained in the second step are voted to obtain the classification result, i.e., the category with the most votes is used as the final model output; for the regression problem, the average of these B models is calculated as the final result [11].

3.2 Precision analysis indicators

The following three indicators are used for the accuracy analysis and comparative analysis of the model.

1) Correlation coefficient(R^2), whose expression is:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{6}$$

In the formula,$\hat{y}_i$ is the predicted value,y_i is the measured value, $\bar{y}$ is the mean value of the measured value, and N is the number of samples.

R2 measures the fitting degree of the predicted value to the actual value, the closer R2 is to 1, indicating that the model has a better fitting effect on the predicted value. Conversely, the smaller the R2, the worse the fitting effect.

2) Mean Absolute Error (MAE), whose expression is:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |(\hat{y}_i - y_i)| \tag{7}$$

MAE is a loss function for the regression model. It is the sum of absolute values of the difference between actual and predicted values. It only considers the average module length of the predicted error without considering the positive and negative directions.

3) Root Mean Square Error (RMSE), whose expression is:

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

The root means the square error is the square root of the deviation between the predicted value and the actual value and the number of observations N, that is, the square root of MSE, which is used to measure the deviation between the observed value and the actual value.

4. Authentic proof analysis

4.1 Data collection and data normalization

This paper selects the closing price of the Shanghai Stock Index from January 4, 2021, to January 31, 2021, as the research object, a total of 243 days of data. The input data of RBF and SVM are stock technical indicators, which are the opening price, the highest price, the lowest price, trading volume, and growth rate. The output data are the daily closing stock price.

To eliminate the influence of the dimension of different indexes of the original data and the large difference between different indexes of the original data, and to improve the calculation ability and accuracy of the model. The closing price, opening price, the highest price, the lowest price, volume, and growth rate of six types of data were normalized, the normalized data were recorded as close, open, high, low, vol, rate.

The normalization formula is as follows:

$$x_{norm} = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{9}$$

4.2 Price Forecast

In this paper, three consecutive sets of 15 days of data are randomly selected from 243 days of data as the study data, with the first 10 days of each set of data as training data and the last 5 days of data as test data. Bagging integrated learning based on RBF-ANN and SVM is performed to predict the stock price in the last 5 days of each data set. The test flow is shown in Figure 3.

In Bagging integrated learning, since the sub-training set is generated by randomly selecting more than 7 days of data from 10 days of data, the training set of a single sub-model has a total of 176 (C_{10}^{7}, C_{10}^{8}, C_{10}^{9}, C_{10}^{10}). Because there are two sub-models, RBF and SVM, there are 352 training sets for Bagging integrated learning. Among the 352 sub-learners, R2 is used to judge the learning effect of the sub-learners. The sub-learners with R2 are lower than 0.8 are excluded, and the final results are obtained by averaging among the remaining sublearners [11].

The structure of the RBF-ANN neural network is n × 10 × 1. The input layer is n nodes, which varies with the number of elements n in the training set, for the five technical indicators of the SSE index for n days in the training set, 10 nodes in the middle layer, and 1 node in the output layer, for the closing price of the SSE index on the next day.

The structure of the SVM support vector machine regression prediction model uses a total of five principal components data of the SSE stock technical index to build a support vector machine regression model $y = \omega_1\overline{F}_1 + \omega_2\overline{F}_2 + \omega_3\overline{F}_3 + \omega_4\overline{F}_4 + \omega_5\overline{F}_5 + b$, The kernel function is a Gaussian radial kernel, and the parameters are taken as C=1, ε= 0.001, γ= 0.008 [8].

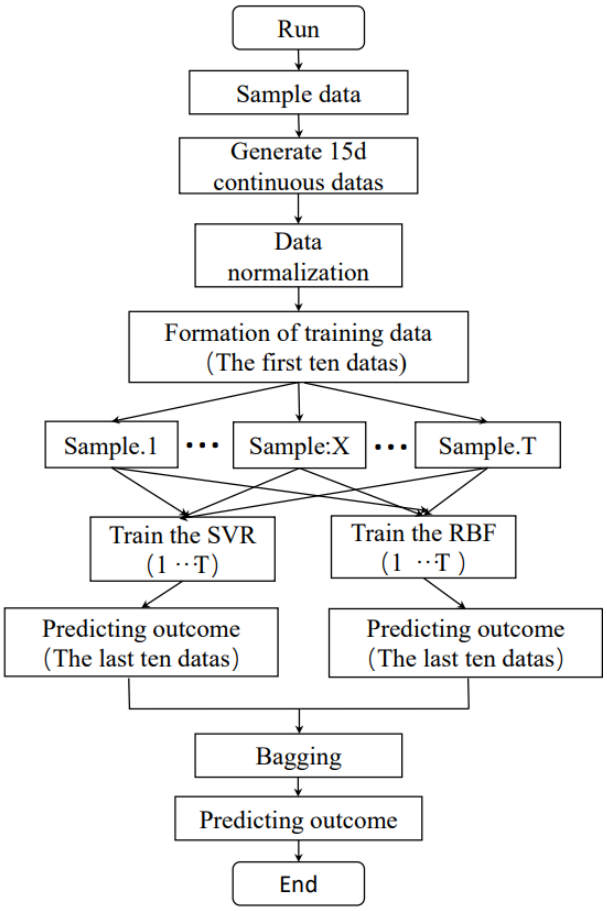

Figure 3. Bagging integration strategy diagram of RBF and SVM

The three sets of 15d continuous data obtained by random selection were March 30-April 20, June 15-June 25, and December 1-December 21. The first 10d data of each group are used as the training set, and the last 5d data are used as the test set, respectively. Based on the above principles, the integrated learning model program of RBF and SVM with Bagging, the regression prediction model program of SVM alone, and the regression prediction model program of RBF-ANN alone are written. The analysis plots of each prediction model's predicted and actual values after 5d for three randomized trials were obtained by Matlab implementation and are shown in Figure 4.

The comparison of the actual values of 5d after three randomized trials and the prediction results of each prediction model is shown in Table 1. It is easy to find from the table that the maximum relative errors of RBF, SVM, and Bagging in the first trial are 1.83%, 1.71%, and 1.70%, respectively, and the maximum relative errors of the second trial results are 2.69%, 2.31%, and 2.21%, respectively; the maximum results of the third trial relative errors chalked as 1.92%, 1.70%, and 1.59%. The Bagging integrated prediction model has the smallest relative error, RBF the largest, and SVM the second. Meanwhile, the overall analysis of the relative error of each model's daily prediction in Table 1 shows that the Bagging prediction model has a greater possibility of reducing the relative error of prediction based on SVM and RBF, which is more evident in Test.2 and Test.3.

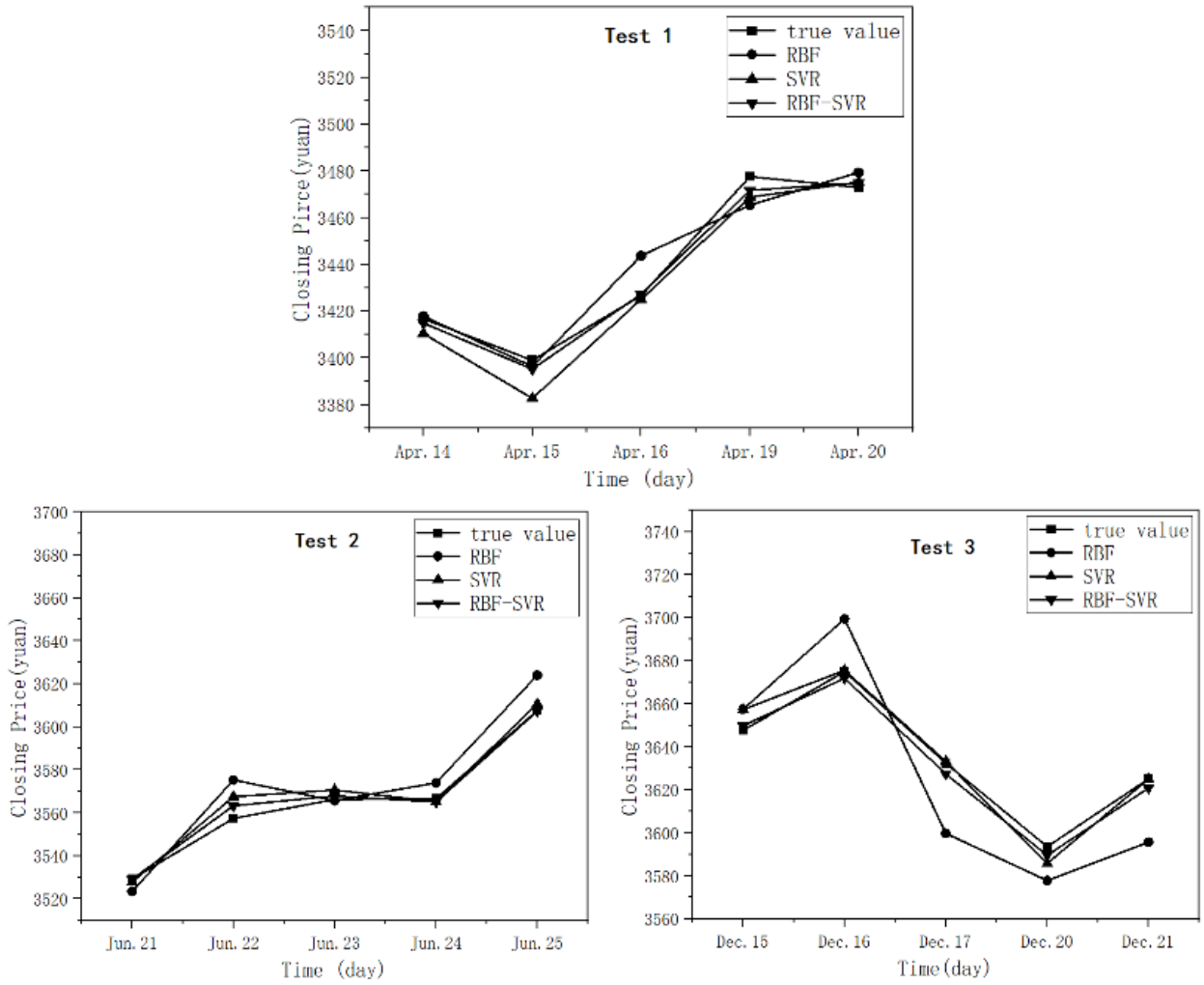

Figure 4. Prediction analysis of each prediction model for three randomized trials

Table 1. Stock price prediction results of three randomized trials

The Relative Error. of Prediction Results (%)														
Test.1					Test.2					Test.3				
Date	Ture	RBF	SVM	Bagging	Date	Ture	RBF	SVM	Bagging	Date	Ture	RBF	SVM	Bagging
4.14	3417	-0.03	0.18	0.05	6.21	3529	0.16	0.04	0.00	12.15	3648	-0.27	-0.27	-0.06
4.15	3399	0.60	1.00	0.63	6.22	3557	-1.31	-1.09	-0.97	12.16	3675	-1.42	-0.77	-0.67
4.16	3427	-0.79	-0.24	-0.31	6.23	3566	-1.03	-1.18	-1.10	12.17	3632	1.31	0.40	0.55
4.19	3478	-1.42	-1.52	-1.61	6.24	3567	-1.27	-1.02	-1.02	12.20	3594	1.92	1.70	1.59
4.20	3473	-1.83	-1.71	-1.70	6.25	3608	-2.69	-2.31	-2.21	12.21	3625	1.43	0.62	0.73

4.3 Accuracy Analysis

The application of evaluation metrics R2, MAE, and RMSE to test the accuracy of each prediction model for each randomized test is shown in Table 2.

(1) The closer the value of the R2 parameter is to 1, the better the model fits the predicted value. In the three tests in Table 2, the R2 of the Bagging integrated learning model is on average 36.42% higher than that of the RBF model and 4.10% higher than that of the SVM model, which proves that the Bagging integrated learning model fits better, here due to the weighted optimization of that set of machine learning by ensemble learning, resulting in lower error.

(2) The closer the value of the mean absolute error (MAE) indicator is to zero, the smaller the error between the predicted value and the actual value. In the three experiments of Table 2, the MAE value of the Bagging integrated learning model is reduced by 343.34% compared to the RBF model and

97.84% compared to the SVM model, which proves that the prediction result of the Bagging integrated learning model is closer to the actual value (3) root mean square error (RMSE). It is explained, and given the coupled nature of ensemble learning, the robustness of the sub-prediction model is further integrated, increasing the accuracy of the coupled model.

(3) Root mean square error (RMSE) is the degree of fit between the measured value and the actual value curve, which is used to measure the accuracy of the measurement. The smaller the RMSE value, the higher the measurement accuracy. From the results, it can be seen that the algorithm after ensemble learning has strong robustness and generalization.

Table 2. Comparison of prediction accuracy of three randomized trial prediction models

Precision index	Test.1			Test.2			Test.3		
	RBF	SVM	Bagging	RBF	SVM	Bagging	RBF	SVM	Bagging
R^2	0.900	0.920	0.988	0.786	0.957	0.987	0.196	0.956	0.979
MAE	9.887	8.834	3.360	11.620	5.224	2.817	23.996	5.589	3.850
RMSE	53.656	12.497	8.606	25.984	11.680	6.299	22.109	20.041	7.279

5. Conclusion

This paper combines RBF neural network and SVM support vector machine using Bagging integrated learning to form a new integrated model for short-term stock price prediction and the closing price of the SSE index from January 4, 2021, to January 31, 2021, with a total of 243 days of data, is selected as the research object. Using the five indicators of the opening price, high price, low price, volume, and growth rate, two sub-models of RBF neural network and SVM support vector machine were constructed, respectively, and an integrated learning model of RBF and SVM with Bagging was constructed to predict the closing price of stocks through three randomized trials. The prediction results of the three models were compared and analyzed comprehensively. Through the above study, the following conclusions were obtained.

(1) The constructed Bagging integrated learning model has a correlation coefficient R2 closer to 1, which is based on RBF neural network and SVM support vector machine. A smaller mean absolute error MAE and root mean square error RMSE than the individual models, it's fitting effect is good, and the relative error of the prediction results is lower, which is more suitable for predicting short-term stock prices.

(2) In short-term stock price prediction, the general prediction effect of SVM support vector machine is better than that of RBF neural network, its correlation coefficient R2 is closer to 1, its mean absolute error MAE and root mean square error RMSE is small, and the prediction results of RBF neural network are unstable.

(3) The machine learning model constructed in this paper can predict stock price changes through stock market-related indicators, which have particular reference significance and help investors discover investment opportunities and improve investment returns.

References

[1] Zhao Lijun, Wang Junnan & Cheng Jianhua. (2021). Stock price fluctuation prediction analysis based on integrated long-term and short-term memory neural network model. Journal of Anhui University (Natural Science Edition) (04), 17 - 26.

[2] Zhao Qingguo, Kong Xiangyue, Liu Liming & Yang Longqian. (2020). Construction of time series weighted mean model for short-term stock price forecasting. Journal of Shenyang Aerospace University (04), 81 - 89.

[3] Zhang and Liu. (2019). Markov chain model for short-term stock price prediction. Volkswagen Investment Guide (12), 272.

[4] Adults. (2018). Research on short-term stock price prediction based on long-term and short-term memory networks (Master's degree thesis, Chongqing University).

[5] Guo Jianfeng, Li Yu & Anton. (2017). Short-term stock price forecasting based on LM genetic neural network. Computer technology and development (01), 152-155 + 159.

[6] Yang Jie. (2016). High-temperature human thermal reaction simulation and experimental research based on human-clothing-environment (doctoral dissertation, Tsinghua University).

[7] Wu Wei, Chen Weiqiang & Liu Bo. (2001). Using BP neural network to predict stock market fluctuations. (eds.) Optimization Method, Tectonophysics and Risk Management-Proceedings of CCAST Workshop (pp.7-27).

[8] Peng Lifang, Meng Zhiqing, Jiang Hua & Tian Mi. (2006). Application of support vector machine based on time series in stock forecasting. Computing technology and automation (03), 88 - 91.

[9] Li Minjie & Wang Jian. (2020). Prediction of cold chain logistics demand for aquatic products based on RBF neural network. Agricultural resources and regionalization in China (06), 100 - 109.

[10] Table 2 (2007). Stock price forecasting methods (Master's degree thesis, Tianjin University).

[11] Niu Hongli & Zhao Yazhi. (2020). Prediction of stock price index using Bagging algorithm and GRU model. Computer engineering and application (1-8).