ISSN 2616-5902 Vol. 3, Issue 6: 113-118, DOI: 10.25236/AJBM.2021.030619

Stock Prediction Methods based on Ensemble Learning

Zhiyuan Wei[1, *], Yingxu Chen[2], Meng Gao[3], Yuancen Li[4], Jianan Wan[5], Yuqi Su[6]

[1]*University of Texas at Austin, Austin, TX, United States*
[2]*Henan Experimental High School, Zhengzhou, Henan, China*
[3]*Shandong Normal University, Jinan, Shandong, China*
[4]*Hefei No.1 High School, Hefei, Anhui, China*
[5]*Zhejiang International Studies University, Hangzhou, Zhejiang, China*
[6]*Xi'an University of Technology, Xi'an, Shaanxi, China*
**Corresponding author: andywei1003@gmail.com*
These authors contributed equally to this work

Abstract: *With the rapid development of stock market, there have been large interests in stock prediction. The decision making based on rational and logical analysis as well as forecast often has a very positive supporting effect, reducing investment risk while enhancing the profits. The development of technology has led to a variety of mature machine learning models for predicting the stock market such as the support vector machine (SVM) model and support vector regression (SVR) model, which will be introduced later in the paper. In this paper, it focuses on the improvement of the existing machine learning models by comparing the deviation and coefficient of curves of different stocks. The experiment indicates that the ensemble models provide more effective and more accurate stock prediction compared with only using the SVR model.*

Keywords: *stock prediction, SVR, machine learning models, ensemble learning*

1. Introduction

Since the establishment of the stock market in China for more than 30 years, the stock market has been in a process of constant fluctuating rise and return to the fluctuating adjustment. On August 9, 2007, the total market value of Shanghai and Shenzhen's stock markets exceeded 20 trillion yuan. And for the first time, the total market value exceeded the total GDP level of 21.087 trillion yuan in 2006 [1]. After the Spring Festival in 2009, the stock market continued to develop, and in December 2012, the stock market value rose again [2]. At present, China's stock market has become an important part of the socialist market economy. At the same time, the stock market exists downturn period of oscillation. The 2007 financial crisis made the Chinese stock market experienced a rapid bubble. Although in the following years, stock market had the rational regression, in June 2013, the stock market plunged again, such a big shock showed that the Chinese stock market still exist many problems. In view of the large scale, because of the wide range and heavy significance of the stock market, it is necessary to make stock forecast. Stock forecast refers to the prediction behavior of the future development direction of the stock market and the degree of rise and fall by the securities analysts who have a deep understanding of the stock market according to the development of the stock market. This forecasting behavior is only based on the assumed factors and the preconditions. The prediction of the stock market can help to construct a more reasonable investment portfolio, and then make corresponding investment decisions, so as to reduce risks and obtain higher returns [3]. The main research problems of stock forecasting include 1) stock price forecasting; 2) Turning point prediction; 3) Quantitative investment model [4]. The stock price prediction is a kind of financial time series prediction. Through analyzing the historical stock price trend and combining with various other information, various researchers reveal the change law of stock price and make short-term or long-term prediction of its price. In the stock market, the change of stock price is related to the macroeconomic development of the country, the formulation of laws and regulations, the operation of the company, the confidence of the shareholders, the complex and changeable stock market, the evolving economic system, its non-linear and non-stable characteristics, and other factors. Therefore, the so-called prediction is difficult to predict accurately, and how to predict the stock price reasonably and effectively has become an urgent problem to be solved in related research.

ISSN 2616-5902 Vol. 3, Issue 6: 113-118, DOI: 10.25236/AJBM.2021.030619

1.1 Related work

With the development of modern information and network communication technology, the research of machine learning and other related fields has been greatly developed. A variety of machine learning models have achieved good experimental results in stock prediction. Among them, the support vector machine (SVM) model and support vector regression (the SVR) model as a machine learning model are more mature, having advantages in solving high dimensional feature classification and regression problems. It also corresponds with more features in the field of stock prediction data, therefore widely used in the study of stock prediction.

Support vector machine was proposed in 1995 [5] and has been developed for more than 20 years. Its maximized classification boundary theory has realized a number of historically influential applications and is a very important and widely used model in the field of neural network and machine learning. Support Vector Machines is a kind of binary classification model. Its basic model is defined as a linear classifier with the largest spacing in the feature space, and the interval is the most distinguishable from perceptron. SVM also includes kernel tricks, which make it essentially a nonlinear classifier. In recent years, the prediction models represented by GARCH model and SV model have been widely used in asset price prediction, but these models also have some corresponding limitations. Therefore, the innovative prediction models based on intelligent algorithms such as BP neural network and support vector machine have been widely used in financial asset price prediction. In 2013, on the basis of elaborating the theory of innovative forecasting model, Peng Wangshu used the pre-model based on neural meridian and support vector machine to forecast the stock index respectively [6]. With the rapid development of China's economy and the continuous improvement of China's stock market, more and more people participate in the stock market, eager to predict the stock price and pursue the accuracy of prediction. Later, more and more studies are devoted to adding more information into the SVM model to improve the prediction results. For example, [7][8][9] combined SVM with ARIMA model, least square method, and network public opinion respectively to obtain better prediction results of stock prices.

On the regression of stock price prediction, an important branch of SVM is the application of support vector regression, which is a widely studied machine learning model. In [10], SVR model is applied to stock forecasting research for the first time. Later, SVR, as a relatively mature technology, has been continuously improved in stock prediction research through the combination with various other algorithms, in order to obtain better prediction results. Then in [11], the differential evolution algorithm is used to solve the parameter optimization problem of support vector machine model, and the experiment proves that the optimized support vector regression with the improved differential evolution algorithm can achieve better prediction results. In [12], systematically introduces the development of the theory of stock price forecasting process, and a variety of commonly used time series prediction methods, including ARIMA, neural network, grey theory, and the latest development of Support Vector Machine, SV theory, as well as introducing the theory of Support Vector Machine in detail from the aspects of theoretical background, basic definition, principle and steps, and its advantages and disadvantages are analyzed. In addition, [12] also used Support Vector Machine (SV) and ARIMA to conduct empirical analysis, regression analysis and prediction on the stock price of China Merchants Bank in the second half of 2006, to find out the optimal parameters. By predicting the price in December, we prove that the support vector machine has a very high prediction accuracy for the short-term time series prediction, and thus the value of support vector machine is proved.

2. Methodology

2.1 Intro

We obtained historical stock data of many domestic and foreign companies through the yfinance module, and in the process of applying SVR model to make prediction of the future stock prices, we found out that for some specific stocks such as Amazon (stock code: AMZN) and Tian Chang Group Holdings Ltd (code: 2182.HK), the results SVM model simulated have large bias. The bias is reflected on the RMSE and the decisive factor R2. We let the true value been y_i , and expected value been y_i with a total of m data in the test set. Then the formula of RMSE will be the following:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2}$$

ISSN 2616-5902 Vol. 3, Issue 6: 113-118, DOI: 10.25236/AJBM.2021.030619

From a machine learning perspective, a lower RMSE will result in a less degree of dispersion for the bias, and thus a better prediction. The coefficient of determination reflects the proportion of the change in the dependent variable that can be explained by the independent variable through the regression relationship. Using the mean as the error benchmark, we will see if the predicted error is higher or lower than the mean error, and the formula is the following:

$$R2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y_i})^2}{\sum_{i=1}^{m}(y_i - \overline{y_i})^2}$$

We first analyze the reasons for the excessive deviation of the RMSE and the coefficient of determination R2 for SVR model on Amazon and other stocks. By comparing the curve of changes in the stock prices of these stocks with other stocks, we found that the range of changes in the stock prices of these stocks is very large, unlike other stocks that change relatively smoothly. This means that the accuracy of using the SVR model to model stock price forecasts is greatly affected by the distribution characteristics of the data. Using a single model to analyze different forms of data set cannot guarantee a good fitting. Thus, our improved method is to apply ensemble learning to integrate multiple models within an ensembled learner to predict stock prices.

2.2 Ensemble learning

Ensemble learning is to combine multiple individual learners, also called basic learners, with a certain strategy to form a learning committee in order to obtain a better comprehensive strong learner. If all individual learners are of the same kind, for example, decision trees or neural networks, then this kind of ensemble is called Homogeneous; conversely, if there are both decision trees and neural networks, then is called heterogeneous. When choosing individual learners, we generally pay attention to the following two criteria: (1) Accuracy: The selected individual learner must have a certain degree of accuracy. (2) Diversity: There must be certain differences between individual learners.

The idea of ensemble learning is that the errors of one learner can be corrected by another learner. Therefore, when selecting individual learners, we need to choose different types of learners, so that these learners can complement each other, and the final prediction result is better than the single learner result. The methods of learner ensemble can be divided into two categories: (1) Sequence integration, a single learner is trained and generated in sequence (for example, AdaBoost). The principle is to use the dependency relationship between the basic learners to improve the overall prediction results by assigning higher weights to the samples that were incorrectly marked in the previous training. (2) Parallel integration method, a single learner generates in parallel (for example, Random Forest). The principle is to use the independence between basic learners to reduce errors and improve prediction accuracy through averaging.

For the output results of different learners, the fusion methods mainly include the average method, the voting method, and the weighting method.

2.3 Intro to Models

According to the fact that the performance of a single learner cannot be too bias and the differences between the learners are required, we integrate the linear regression model (LR) and the k nearest neighbor model (KNN) on the basis of the support vector regression model (SVR). Next, we will introduce these three machine learning models.

The traditional SVM is a classifier. Its basic idea is to find a partitioning hyperplane in the sample space so that the interval between the positive and negative samples closest to the hyperplane is maximized. SVR believes that the absolute value of the error between the predicted regression model f(x) and the true value y is tolerable within ϵ. That is to say, a 2ϵ-wide interval is constructed with f(x) as the center. If the sample falls within this interval, then the prediction is considered correct. Through the introduction of the first chapter, we can also know that SVR is widely used in stock price forecasting.

Linear regression is to use lines or curves to fit the distribution and trajectory of data points in space, that is, to use linear combinations of features to represent the fitted lines or curves. The commonly used fitting methods for linear regression models include least squares approximation and gradient descent method. In linear regression, the data is modeled using a linear predictive function, and unknown model parameters are also estimated through the data. Linear regression modeling is usually used for data forecasting, time series models and discovering the relationship between variables. It is a simple and effective regression method.

ISSN 2616-5902 Vol. 3, Issue 6: 113-118, DOI: 10.25236/AJBM.2021.030619

The idea of the K nearest neighbor model is that in a sample data set, there are many samples with known categories and feature attributes, then for a sample to be classified, its category is determined by the category of the K samples that are most similar to it in the feature space. In other words, if most of the K nearest samples belong to a certain category, then this sample also belongs to this category. In the classification decision, this method only determines the category of the sample to be classified based on the categories of the nearest samples. The three basic elements of the K-nearest neighbor model are the selection of k value, distance measurement, and classification decision rules.

When choosing a basic learner, our main basis is the two criteria mentioned earlier, namely accuracy and diversity. In addition to the SVR model, the accuracy of the LR model and the KNN model in different stock price predictions is not low. The linear regression model is a mathematical statistical model. The classification idea of the K-nearest neighbor algorithm is based on the categories of a small number of adjacent samples. These are two completely different models. Therefore, our ensemble model is a heterogeneous ensemble model. At the same time, both the LR model and the KNN model are simple models. We know that the simpler the model, the stronger its generalization ability. This is consistent with our original intention to improve the SVR model's weaker generalization ability in stock price prediction. When integrating learners, this article uses a parallel integration method to train different models separately and combine the final predicted values together to obtain the final predicted value.

3. Experiment

3.1 Experimental Data

Through the yfinance module, we obtained the stock historical data of many domestic and foreign companies. In the experiment part, we mainly focus on the stocks with large deviations in the stock price prediction in the SVR model, such as Amazon stock, stock code AMZN and Tianyong Group, stock code 2182.HK. We used the closing price of Amazon and Tianchang Group stock prices during 2000-01-01 to 2018-12-01 as the target value for training and prediction. 70% of the acquired data were used as training data and 30% as testing data.

3.2 Experimental Result

The SVR model and the integrated model proposed in this paper are used respectively to forecast and compare the stock prices of Amazon and Tianchang Group. The experimental results are shown in Figure 1-2 below. It is obvious from the figure that the prediction result of the integrated model is better and more accurate than that of the simple SVR model, no matter for the prediction of Amazon stock or the stock price of Tianchang Group.

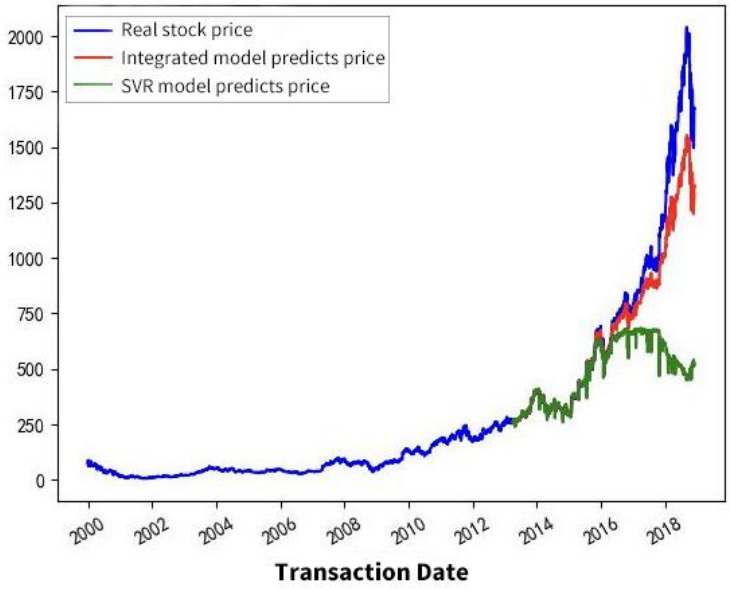

Figure 1: Amazon Stock Forecast Results

ISSN 2616-5902 Vol. 3, Issue 6: 113-118, DOI: 10.25236/AJBM.2021.030619

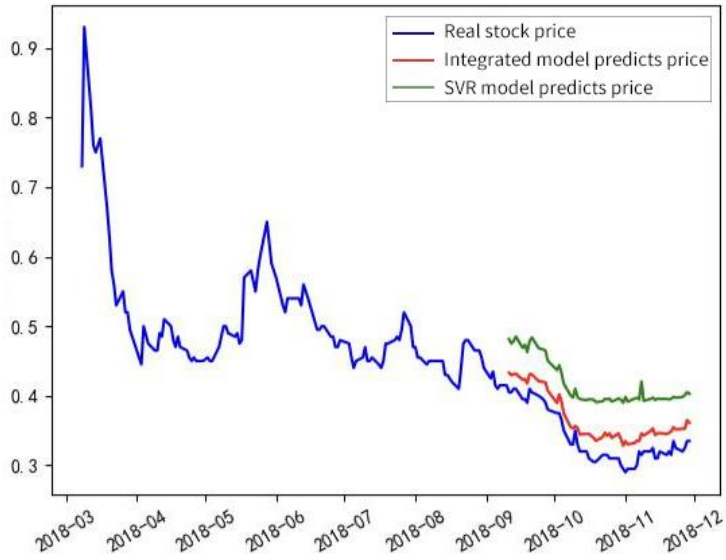

Figure 2: Tian Chang Group Stock Forecast Results

Table 1 lists the results obtained after the training of the same stock price data by the two models, respectively. The evaluation indexes include three methods, namely RMSE and determination coefficient R2 introduced previously, as well as the score of 10-fold cross-validation.

Table 1: Evaluation index results of SVR model and Integration model

	SVR model	Integrated model
RMSE	0.0685	0.0318
	273.8356	82.43232
R2	0.5341	0.8998
	0.5505	0.9592
Cross-validation-score	0.1220	0.8998
	0.5379	0.9593

In Table 1, the blank shaded table is the data of Tianchang stock, and the blue shaded table is the data of Amazon stock. It can be seen from the table data that the three evaluation indexes of the integrated model are all improved to some extent compared with the single SVR model. Specifically, the root means square error (RMSE) decreases, the correlation index increases, and the cross-validation score on the test set increases, indicating that the prediction effect of the model has been greatly improved. Among them, the evaluation index and cross verification score have been greatly improved.

4. Conclusion

Through the SVR model, the prediction experiment of multiple stock futures is predicted. It has found that the SVR model is not strong enough - the prediction deviation of the R2 and RMSE of some stocks is large, and the deviation is too large. The SVR model prediction method is improved. The improved algorithm based on integrated learning is proposed, and the integrated plurality of models are more accurately predicted. The idea of integrated learning is mainly combined with a variety of weak learning units, and the economy is short, and the effect is improved. The integrated model proposed in this article is based on the SVR model. It integrates a simple and effective linear model LR and K neighboring model KNN, using different stock data training forecast results, proves that integrated model relative to simple SVR models, in many there are better precision on the evaluation indicators. At the same time, this article still has further improved space: This article only considers the influencing factors of the closing price. If you can fuse more factors, it may get better forecast results; this article uses the SVR model, the default parameters are used. Setting, the search for the optimal parameter is also a direction that can explore.

ISSN 2616-5902 Vol. 3, Issue 6: 113-118, DOI: 10.25236/AJBM.2021.030619

References

[1] Huai Jiang, 2009. Research on the Development of China's Stock Market. International Economic Cooperation, 2009(07).
[2] Teng Wang, 2013. Analysis of the Current Situation of China's Stock Market. Chinese and Foreign Entrepreneurs, 2013(18).
[3] Zhenquan Zhao, et al, 2001. Appropriate Regulation Size of Chinese Securities Market. China 2001: Economic Situation Analysis and Forecast.
[4] Haoran Xu, et al, 2020. Analysis on Application of Machine Learning in Stock Forecasting. Computer Engineering and Applications. 56(12): 19-24.
[5] Vapnik & Vladimir, 2013. The Nature of Statistical Learning Theory. Springer Science & Business Media.
[6] Wangshu Peng, 2013. Comparison of Stock Index Prediction Models Based on BP Neural Network and Support Vector Machine. 2013 Financial Markets, 1007-9041-2013(01)-0071-02.
[7] Wenjuan Mai, et al, 2018. Stock Price Prediction Based on ARIMA-SVM Model. 2018 International Conference on Big Data and Artificial Intelligence.
[8] Wei Huang, et al, 2005. Forecasting Stock Market Movement Direction with Support Vector Machine. Computers & Operations Research 32.10 (2005): 2513-2522.
[9] Shijun Zhang, 2014. Stock Price Prediction Base on Network Public Opinion an Support Vector Machine. Nanjing University of Information Science & Technology.
[10] Trafalis, et al, 2002. Benders Decomposition Technique for Support Vector Regression. Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290). Vol. 3. IEEE.
[11] Shanqing Yang, 2018. Stock Price Prediction Based on Support Vector Regression and Differential Evolution. Nanchang University.
[12] Ersu Biao, 2007. Research on the Stock Price Prediction. Tianjin University.