A Stock Price Analysis and Prediction Method Based on Machine Learning

Runqi Wang(✉)

Shandong University of Finance and Economics, Jinan, China
1287098621@qq.com

Abstract. Changes in the stock market have a significant impact on individual investment management and are closely related to economic research and market development trends in the country as a whole. If the trend of stock prices can be predicted more correctly, investors will make more scientific investment decisions based on it, which is of great significance to promote the effective allocation of resources and improve market efficiency from a macro perspective. In recent years, big data and artificial intelligence technology have begun to rise, machine learning, as an important technology in the field of artificial intelligence, has excellent performance in simulating the specific characteristics of objects and processing complex and large amounts of data. Therefore, this paper takes the daily closing price of stocks S1 and S2 as the essential data, and uses the support vector machine (a machine learning model) of python to analyze, prove and predict the stock quotations in China. The analysis results show that the accuracy of using the support vector classifier to predict stocks is as high as 90%, We performed parameter optimization for the second time on the basis of this model, and the accuracy of the support vector machine for stock prediction is as high as 90%. This conclusion indicates that the stock prediction method based on machine learning model has high accuracy, and has certain reference value for practical application.

Keywords: machine learning · the stock price forecast · stock prices · support vector machine (SVM)

1 Introduction

The changes in the stock market are inextricably linked with the progress of the national market, and the healthy development of the stock market has a very important impact on the continuous growth of the national economy. The future direction of stock prices has always been a central concern for investors, Correct judgment of stock price trends is not only good for investors to make correct investment decisions, but also of great significance to promote the effective allocation of resources and enhance the effectiveness of the market. Nowadays, big data and artificial intelligence technology have begun to rise. Machine learning, as an important technology in the field of artificial intelligence, has an excellent performance in simulating the specific characteristics of objects and

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 1471–1477, 2023.
https://doi.org/10.2991/978-94-6463-198-2_152

processing complex and large amounts of data, which can not only create corresponding economic strength for the country, but also promote the improvement of the comprehensive national strength of the country. Combining machine learning with the stock market can greatly improve the accuracy of stock market prediction.

By referring to the research experience of domestic and foreign scholars at the national level, this paper forecasts and analyzes the stock market situation of China combined with some basic indicators. Selecting daily closing price as the index of stock quotation prediction, this paper measures the development trend of the stock of our country, and puts forward solutions to the development trend of the stock of our country through the analysis of the current situation and the results of market quotation measurement. At the same time, it puts forward some suggestions to perfect the stock market of our country and reduce the uncertainty of the stock market.

2 Related Work

Liu Qingxia [1] (2017) verified that the improved back propagation network based on principal component analysis can be well adapted to stock data technology through learning and training, and has a good prediction effect. Shen Jinrong [2] (2017) took financial indicators as the analysis object, used the improved CART decision tree and stepwise regression for measurement, and concluded that the stepwise regression model based on the decision tree can reduce the financial indicators that affect the target variable and improve the prediction accuracy of the model. Li Dan [3] (2018) studied the stock prediction problem from the perspective and conducted empirical analysis, analyzing the optimal network structure, prediction results and experimental results of SVFD-BPNN, MVFDIF-BPNN and MVFDIL-BPNN. Hu Di and Huang Wei [4] (2019) empirically analyzed the stock correlation based on the combination algorithm of SVM and the affinity propagation clustering algorithm of clustering stock prediction, and verified that the combination of the AP algorithm and other algorithms can improve the accuracy of stock prediction. Zhang Jinghua and Gan Yujian [5] (2019) proposed that deep learning support vector machine optimized the configuration of model parameters, and used this model to conduct simulation experiments. The results showed that deep learning SVM had significant improvement in prediction accuracy compared with existing SVM.

There are more studies on stock forecasting abroad than in China.

Sam Nelson built on his view of the market and eventually developed Dow Theory. W. D. Gann studied the importance of time and proposed the concept of "price-time equivalence". Frank Rosenblatt (1957) invented a linear classifier called perceptron. Corinna Cortes and Vapnik [6] proposed an SVM based on statistical learning with many unique advantages in the face of nonlinear, small sample size, and high-dimensional pattern recognition problems in the mid-1990s. Lerner and Vapnik (1963) introduced the maximum interval classification algorithm. Soft interval classifiers were introduced by Cortes and Vapnik (1995), the same year that SVM was extended to regression models. Kim [7] (2003) proved through experimental analysis that support vector machine is obviously better than other existing forecasting methods, and used SVR to make regression prediction of asset prices and found that the prediction effect is good. Funatsu and Kaneko [8] (2013) proposed to use of an online support vector machine based on time series to

study the adaptive software perception prediction model. In addition, they also studied the window size and appropriate hyperparameter Settings, and obtained the regression reliability prediction.

To sum up, although great progress has been made in forecasting the stock market, there are still many places to explore in the depth and scope of the theory. At present, it is not only an important period of the development of the national economy, but also an important period for the development of the change of the securities investment in our country. Therefore, how to find a method that can effectively improve the current defects on the basis of the advanced theories put forward by domestic and foreign scholars has become the key. Based on the above considerations, this paper chooses SVC in SVM with strong generalization ability as the core model for predicting stock prices.

3 Design of Methods

3.1 How Support Vector Machines Work

Support vector machine (SVM) is a machine learning method based on dimension theory and structural risk minimization theory. It is a generalized linear classifier, which can classify data information according to the supervised learning method, and can also analyze non-linear classification through kernel method research. It breaks through the phenomenon of small-scale data over-fitting which is easy to occur in traditional machine learning based on empirical risk minimization theory.

3.1.1 Kernel Function of Support Vector Machine.

In the feature space, we want the samples to be linearly separable. But if we don't know these feature maps, we can't know exactly which kernel is appropriate. Therefore, the correct choice of kernel function is very important for the quality of the support vector machine model. Here are a few common kernel functions:

linear is the simplest kind of kernel, The calculation method is: $K(X_i, X_j) = X_i^T X_j$. **Polynomial kernel function (poly)** is a non-standard kernel function, which is very suitable for orthogonal normalization data sets, The calculation method is: $K(X_i, X_j) = (X_i^T X_j)^d, d \geq 1$. **Gaussian kernel function (rbf)** has better anti-interference ability in dealing with data noise. The calculation method is:

$$K(X_i, X_j) = -exp\frac{(\|x_i - x_j\|D)^2}{2\delta^2}, \delta > 0 \tag{1}$$

3.1.2 The Parameters of the Support Vector Machine.

The correct selection of SVM parameters has a great impact on the classification management effect. In general, the parameters to be optimized are the C penalty parameter and the σ kernel parameter, respectively. However, at present, there is no good theory to guide the optimization of parameters, and the commonly used methods include experiment, grid, gradient descent, etc. In this paper, the grid method is used to optimize

the management of C, which simplifies the operational activity process of parameter selection and improves the classification performance of SVM based on the selected parameters.

3.2 Selection of Samples

In the empirical analysis of stock prediction in this paper, considering that the stock market is A very unstable dynamic process, and its future development trend is also affected by the government's macro-control, considering that the government's control may have a great impact on medicine in the next few years, we select S1, which is greatly affected, and S2, whose price is stable, as the research objects in data selection. It is intended to compare the two types of stock prediction results to verify the credibility of SVM.

4 Empirical Analysis of Stock Forecasting Based on Support Vector Machines

4.1 Data Preprocessing

This paper selects the data of S1 and S2 from January 1, 2018 to March 1, 2020, with a total of 523 data. At the same time, in order to test the data training of python, this paper establishes a large sample data training set and a small sample data training set for S1 and S2. The large sample data adopts the whole sample data, and the small sample data adopts the data from 2019.06.01 to 2020.01.01.

4.2 Process of Operation

Python was used to collect the historical data of the two stocks from January 1, 2018 to March 1, 2020 online, and then preliminarily sort out the data of the two stocks. In the specific experimental operation, value (today's closing price minus yesterday's closing price) is used to represent the rise and fall of the stock. A value greater than 0 is an increase. It is assigned a value of 1; a value less than 0 is a decrease. It is assigned a value of 0.

The first 80% of the data is taken as the training set, and the second 20% of the data is taken as the test set. Then the sample data is normalized. The kernel function is used to make a periodic prediction, and one value is predicted forward each time. 'ploy', 'linear' and 'rbf' are three alternative parameters that could be selected to classify the predicted value respectively. Finally, the accuracy rate in the test set is calculated, and the actual value and predicted value of the output value are shown in Table 1 and Table 2:

The above is the prediction analysis based on the parameters under the default condition. It can be seen that among the three kernel functions, the accuracy is about 90%, both for large samples and small samples. It can be seen that the prediction effect of SVM is ideal. However, since SVM parameters have an important impact on the model prediction effect and 'rbf' accuracy is relatively low, this paper chooses 'rbf' with large samples for parameter optimization:

Table 1. Compared precision results over the large sample dataset

SVM kernel parameter	Accuracy of S1 dataset	Accuracy of S2 dataset
'ploy'	91.43%	93.33%
'linear'	96.19%	96.19%
'rbf'	90.48%	93.33%

Table 2. Compared precision results over the small sample dataset

SVM kernel parameter	Accuracy of S1 dataset	Accuracy of S2 dataset
'ploy'	93.10%	96.55%
'linear'	93.10%	93.10%
'rbf'	89.66%	96.55%

The SVM parameter is selected as C = 1 000 000.0, so this paper changes the C parameter in 'rbf' from the default value of 1.0 to the optimal parameter of 1 000 000.0, and the accuracy results are improved to 98.10%, which is much better than the previous 90.48%. It can be seen that the modified parameter has a positive effect on the prediction effect.

Confusion matrices are error matrices that we can use to evaluate the performance of supervised learning algorithms. In the confusion matrix, the more values appear in the second and fourth quadrants, the better; Conversely, the fewer values that appear in the first and third quadrants, the better. It can be seen from the output results in Table 3, the values appearing in the second and fourth quadrants are 38 and 42 respectively, which are much larger than those of 16 and 9, indicating that the prediction effect of the model is considerable.

Since the statistics of the confusion matrix are only numbers, in the face of a large amount of data, it is difficult to measure the quality of the model only by numbers. Therefore, several indicators are extended based on the basis of the basic statistical results: Accuracy = the number of outcomes correctly predicted by the model/the number of all outcomes predicted by the model; Sensitivity = the number of outcomes

Table 3. Confusion matrix

predicted value	True value	
	Y = 0	Y = 1
Y = 0	38	9
Y = 1	16	42
Total number of samples	54	51

Table 4. Evaluation results of the model in this paper

	precision	recall	F1-score	support
0.0	0.81	0.70	0.75	54
1.0	0.72	0.82	0.77	51
accuracy			0.76	105
macro avg	0.77	0.76	0.76	105
weighted avg	0.77	0.76	0.76	105

correctly predicted by the model/the number of actual values; F1-score = 2 * precision * recall/(precision + recall). It is the output result of combining precision and recall. Its value ranges from 0 to 1. 1 represents the best prediction model, and 0 represents the worst prediction model; Support refers to the number of original data categories.

It can be seen from Table 4 that the data of each index are greater than 70%, and the location is close to 1, so it can be concluded that the prediction effect of the model is ideal. It is also ideal to carry out the same operation for S2 to obtain the prediction effect of its model.

5 Conclusion

This paper applies the machine learning model SVM to the prediction of China's stock market. SVM is used to select and optimize the parameters of kernel functions, and then find the optimal model for measuring stock trends. The main conclusions are as follows:

1. Because of the fast convergence speed and high accuracy of the SVM model, the SVM model can well predict the stock data, making the prediction result very close to the actual value.
2. The selection of kernel function and kernel parameters has a very important impact on the learning and prediction performance of SVM. Different kernel functions and kernel parameters are directly related to the accuracy of the calculation results.
3. SVM has good accuracy in predicting stock prices and provides a meaningful analytical tool for the majority of investors.

Reducing the uncertainty of investors and fund raisers can make the development of the stock market more stable. This paper suggests perfecting our country's stock market roughly from the following aspects in order to reduce the uncertainty of the stock market.

First, through extensive social investigation and discussion, determine the stage goal of the development of the stock market economy in our country. Second, the stock market quality dynamic monitoring system should be constructed as soon as possible in order to achieve the goal of timely and accurate evaluation and mastery of the stock market quality. Third, on the basis of the above, the regulatory authorities should pay close attention to the changes in market quality in order to achieve the stability of the stock market and reduce the investment risks of investors and fund-raisers.

References

1. Liu QingXia. Stock Price Prediction Based on Principal Component Analysis and BP neural Network [D] Suzhou: Soochow University,2017.
2. Shen JinRong. Stepwise regression algorithm based on decision tree and its application in stock prediction [D]. Guangzhou: Guangdong University of Technology, 2017.
3. LI Dan. Research on Stock Prediction Based on Multi-perspective Feature Data [D]. Xi 'an: Northwest University, 2018.
4. Hu Di, Huang Wei. Stock Price Prediction Based on AP. SVM group Table Model [J]. Journal of Wuhan Institute of Technology. 2019, (6) : 297–301.
5. Zhang Jinghua, Gan Yujian. Prediction of Shanghai Stock Exchange Moves Based on Deep Learning Support Vector Machine [U]. Statistics and Decision. 2019, (2): 178.
6. C Cortes. V Vapnik. Support-Vector Networks.[J] Machine Learning.1995.20(3):273–297.
7. Kyoung Jae Kim. Financial Timeseries Forecasting Using Support Vector Machines[J].Neurocomputing,55(1):307–319.2003
8. Hiromasa Kaneko, Kimito Funatsu. Adaptive Soft Sensor Model Using Online upport Vector Regression with Time Variable and Discussion of Appropriate Hyperparameter Settings and Window Size[J] Computers and Chemical Engineering.2013.58(11):288–29