Stock price prediction of Chinese company using support vector machine

Chuchu Qi

Shanghai University, Shanghai, China

qichuchu@shu.edu.cn

Abstract. The accuracy of the prediction in the stock market has gained increasing importance in both research and investment in the market. Support vector machine which can be called SVM is one of important algorithms that plays a role dealing with classification problem and regression measure and the SVM with Pearson VII function-based universal kernel gained great accuracy of prediction of stock price in this paper with the lowest value of mean absolute error of 0.0006 among six selected measures such as random forest to make the prediction and the highest correlation coefficient (0.9999). Based on the collected data and the result of experiment, the efficiency of applying SVM in prediction of stock price has been strongly approved.

Keywords: Stock prediction; SVM; Random forest.

1. Introduction

With the accelerating development of finance especially in the market of stock, the prediction of the stock price has been playing one important role in the research region and practical application. One important reason is that the stock market is enough complicit which is influenced by a large number of factors which can be categorized as macro and micro factors such as Loan Prime Rate, Consumer Price Index, and some policy instructions. It is hard to make prediction facing with the large amount of data. More and more machine learning measures have been applied to help anticipating the stock price in the prediction research because of its character of dealing with big data. Another reason is that in order to gain the expected return from the stock market, the accuracy of the method should be increased to better seize the tendency of the market and the precise price of the stock, which has great value to research institutions and retail investors. The progress made in the research not only benefits the advancement of academic research but also helps to solve some complicit difficulties related to investment of the stock market.

Support vector machine (SVM) is considered as a great choice to classify different values as a popular algorithm [1]. Vapnik introduced support vector machine in 1964[2] and it has developed rapidly and derived a series of extended algorithms, which have been applied in various ares.SVM can be used to make regression to predict the stock price without worrying about the over-fitting problem. Many other kinds of machine learning methods had also been used to do the prediction of the trend of stock market or the price of the stock recent years. A framework of model integrating PCA into WSVM which is used to predict the signals about trading of stock has been designed by Chen and Hao [3]. It has some limitations such as the parameter should be optimized which means this selection problem needs a lot of work and various strategies lead to different outcomes of profits. Ding and Qin have invested an associated network of LSTM to predict the range of the stock price [4], but the calculation method of loss in this model just takes into count the total loss and doesn't consider the relationship between input values. In this paper, SVM is used to predict the price of stocks base on the previous data from some companies in the stock market. The method is applied to process the training data accounted for seventy percents of the population and the performance of prediction can be gained through the weka application.

The remaining of this paper is constructed as follows. Section 2 illustrates the algorithm and model of SVM which represents the function and parameters. Section 3 presents the progress of the experiment about applying SVM to make regression of different companies and get the outcome of each group. And makes the comparison of the performance of SVM with other models such as linear

regression model and Gaussian processes and make some improvement of this model by changing the weight of input parameter. Finally, the limitation of this research and expectations for future researches are given in Section 4.

2. Research Methodologies

2.1 Support vector machine

Support vector machine (SVM) is one of the popular choices of classification algorithm [5]. It was introduced by Vapnik in 1964 [2] and has been applied in a large amount of areas such as the classification of text and the portrait recognition. The algorithm can be concluded that SVM can be used to make the classification to separate the data set into two parts on a hyper-plane by maximizing the margin between different categories. Structure risk minimization principle is implemented in SVM seeking the global optimum that the over-fitting is unlikely to occur compared to other traditional neural network applying the empirical risk minimization principle [6]. A gradient descendent algorithm [7] can be applied to decrease the error when using SVM to make the regression. It is adopted in a research on a log 10-space that in oreder to find the optimal outcome, those values are combined to help find the optimized one [8]. And the kernel parameters are selected during the experiment [9]. In this paper, the Pearson VII function-based universal kernel is selected to test the performance of regression. The principle of SVM is summarized as follows.

It is known that SVM can be used to get the precise classification performance when the data can be linearly separated, and the addition of kernel parameter could help to resolve the problem of classifying the nonlinearly separable data in a higher dimensional space with the hyper-plane. [10]

The goal of SVM is to construct two planes to separate the data into positive part and negative part, and the equation could be concluded as follows:

$$y_i(w^T x_i + b) - 1 \geq 0 \quad i = 1,2,3,\ldots,n \tag{1}$$

To understand this equation, it can be separated into two different portions, the original functions are as follows:

$$w^T x_i + b = +1 \text{ for } y_i = +1 \tag{2}$$

$$w^T x_i + b = -1 \text{for } y_i = -1 \tag{3}$$

The Eq. (2) is the plane for the classification of the positive class while the Eq. (3) describes the negative class, where w indicates the weighted vector, b describes the threshlod or bias and x is the input data of training sample which is divided into two classes by the plane.

The outcome of SVM is the sum of margin which can be describes as the addition of the distance between the two separate planes and the optimal hyper-plane. The formula is the $\text{margin} = d_1 + d_2 = \frac{2}{||w||}$, and according to the classifier of SVM, the width of the margin needs to be broadened which means that the value of w need to be minimized. Eq.(1) and the maximization of the margin can be transferred into the Lagrange formula as follows:

$$\min L_p = \frac{||w||^2}{2} - \sum_{i=1}^{N} \alpha_i (y_i(w^T x_i + b) - 1) \tag{4}$$

The equation can be transformed into another formation as follows:

$$\max L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_i y_i y_j x_i^T x_j \tag{5}$$

Where the $\alpha_i > 0, \sum_{i=1}^{N} \alpha_i y_i = 0,$ the minimization of L_P leads to the calculation of the values of w, b and α. Support Vectors (SVs) can find the maximum margin through the α which has the value except zero.

2.2 Progress of prediction applying SVM

In order to get the great performance of the prediction of stock price, this paper uses the approach based on the SVM with selected vectors as follows:

Step 1: get the previous stock price data of three different public companies from the application which is called Choice. Then original data set is going to be standardized for the purpose of simplifying the complexity of the prediction.

Step 2: finding the high correlation variables related to the stock price of the next day which is also called feature selection. In this section the correlation of each feature can be calculated by the SVM functions including filters and wrappers.

Step 3: After the evaluation of the features, form the training set and test set. In this paper the training set concludes 70 percent of the population and the left is set to test the precision of the model.

Step 4: imply the SVM to get the outcome of experiment and compare the different prediction performance from other models or algorithms such as simple linear regression and Gaussian progress.

Step 5: Improve prediction algorithm by changing the weight of the input variables and repeat the previous steps to find out whether it could increase the accuracy.

3. Experiment analysis

3.1 Data collection

In order to test the effectiveness of the SVM algorithm, the data originated from Shanghai stock market which plays an important role in finance region of China. The daily data from one company has been selected with the code of 300181 from January 7, 2021 to September 2, 2022. Table 1 below shows part of the specific data of the company including 9 variables:

Table 1. Original data

date	eps	pe	ps	price	cenp	ap	ma20	SH
2022/9/2	0.37	25.6	3.52	9.58	54.56	9.59	9.59	3184.98
2022/9/1	0.37	25.18	3.46	9.42	54.56	9.52	9.52	3202.14
2022/8/31	0.37	24.8	3.41	9.28	54.56	9.48	9.48	3227.22
2022/8/30	0.34	26.27	3.45	8.82	54.56	9.34	9.34	3240.73
2022/8/29	0.34	26.71	3.51	8.97	54.56	9.34	9.34	3236.22
…	…	…	…	…	…	…	…	…
2022/8/18	0.34	27.07	3.55	9.09	54.56	9.17	9.17	3277.54

3.2 Result and analysis

In the regression using SVM with the Pearson VII function-based universal kernel which is abbreviated with Puk. The parameter omega is set with the value of 1. the following table presents the outcome of regression using SVM with Puk.

Table 2. Performance of regression using SVM with Puk with omega=1

training data (%)	Time taken to build model (second)	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error (%)	Root relative squared error (%)
10	0.49	0.9947	0.0044	0.0084	6.977	10.6319
20	0.39	0.9993	0.0017	0.0031	2.6977	3.9105
30	0.41	0.9995	0.0013	0.0024	2.0283	3.0817
40	0.4	0.9997	0.0009	0.0018	1.4054	2.3132
50	0.41	0.9998	0.0008	0.0016	1.203	1.9886
60	0.38	0.9998	0.0008	0.0015	1.2481	1.9847
70	0.39	0.9999	0.0006	0.0012	0.9394	1.4939
80	0.38	0.9999	0.0006	0.0012	0.9529	1.4439
90	0.39	1	0.0005	0.0006	0.6677	0.7523

From the Table 2, it is obvious that with the increase of the proportion of training data, there has been an increasing trend on correlation coefficient range from 0.9947 to 1, and the left variables all have the decrease trend including mean absolute error, root mean squared error, relative absolute error and root relative squared error.

When the proportion of training data is 10 percents, the listed four errors all have the maximum value of 0.0044, 0.0084, 6.977 and 10.6319. When the proportion is 90 percents they have the lowest value among the experiment 0.0005, 0.0006, 0.6677 and 0.7523. The data below shows that the prediction of the stock price applying SVM has the high accuracy.

Table 3 represents the comparison of different methods applied on the regression implied to predict the stock price, and the proportion of training set is 70 percent.

Table 3. Comparison results

algorithm	Time taken to build model	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Gaussian Processes	0.63 seconds	0.9899	0.0094	0.012	14.19%	14.58%
Linear Regression	0.01 seconds	0.998	0.004	0.0053	6.04%	6.45%
Multilayer Perception	0.33 seconds	0.9998	0.0018	0.0022	2.72%	2.72%
Random Forest	0.02 seconds	0.9992	0.0009	0.0033	1.31%	3.97%
Simple Linear Regression	0.01 seconds	0.9007	0.0326	0.0363	49.12%	44.13%
SVM	0.39 seconds	0.9999	0.0006	0.0012	0.94%	1.49%

Among all the six algorithms, SVM with Puk shows the best performance of regression. It has the highest correlation coefficient which has the value of 0.9999 compared to other measures like Random Forest which is 0.9992. the lowest value of absolute error and relative squared error of SVM indicates that the measure of prediction gains high accuracy. Take the mean absolute error as an example, SVM has the value of 0.0006 and the value of other algorithms range from 0.0009 to 0.004 which have the higher value compared to SVM.

4. Conclusion

In this paper, the stock price prediction of regression has applied SVM with Pearson VII function-based universal kernel and compared them with different models using algorithms such as random forest and multilayer perception. The data collected from Shanghai stock market with nine variables and 605 instances of one company.

In this paper, Pearson VII function-based universal kernel has been applied in the SVM. Based on the result of experiment, table 2 and table 3 shows that 90 percent of training set gained the most accurate outcome that the value of mean absolute error was 0.0005 and with the same training set proportion of 70%, SVM got the highest accuracy among all the six models which has the relative absolute error of 0.94%. As the prediction of stock price gaining increasing intention from the region of research and real-life investment, the approach used in this paper can be widely applied in many prediction research and advice for the investment and the more data sets are hoped to be used to make future research.

References

[1] Minaee S, Kalchbrenner N, Cambria E, et al. Deep learning--based text classification: a comprehensive review[J]. ACM Computing Surveys (CSUR), 2021, 54(3): 1-40.

[2] Awad M, Khanna R. Support vector machines for classification[M]//Efficient learning machines. Apress, Berkeley, CA, 2015: 39-66.

[3] Yingjun Chen, Yijie Hao, Integrating principle component analysis and weighted support vector machine for stock trading signals prediction, Neurocomputing (2018)

[4] Ding G, Qin L. Study on the prediction of stock price based on the associated network model of LSTM[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(6): 1307-1317.

[5] Minaee S, Kalchbrenner N, Cambria E, et al. Deep learning--based text classification: a comprehensive review[J]. ACM Computing Surveys (CSUR), 2021, 54(3): 1-40.

[6] Kim K. Financial time series forecasting using support vector machines[J]. Neurocomputing, 2003, 55(1-2): 307-319.

[7] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, "Choosing Multiple Parameters for Support Vector Machines," Machine Learning, vol. 46, pp. 131-159, 2002.

[8] T. Eitrich, B. Lang, "Efficient optimization of support vector machine learning parameters for unbalanced datasets," Journal of Computational and Applied Mathematics, vol. 196, issue 2, pp. 425-436, 2006.

[9] S. Boughorbel, J. P. Tarel, N. Boujemaa, "The LCCP for Optimizing Kernel Parameters for SVM," International Conference on Artificial Neural Networks, pp. 589-594, 2005.

[10] Tharwat A. Parameter investigation of support vector machine classifier with kernel functions[J]. Knowledge and Information Systems, 2019, 61(3): 1269-1302.