Advances in Social Science, Education and Humanities Research, volume 290

2nd International Conference on Economic Development and Education Management (ICEDEM 2018)

Research on Stock Prediction in China based on Social Network and SVM Algorithm

Li Tang
Department of Information Science and Technology
Tianjin University of Finance and Economics
Tianjin, China

Shuhua ZHANG
Coordinated Innovation Center for Computable Modeling in Management Science
Tianjin University of Finance and Economics
Tianjin, China

Li He
Department of Information Science and Technology
Tianjin University of Finance and Economics
Tianjin, China

Huiyu Fan
Department of Information Science and Technology
Tianjin University of Finance and Economics
Tianjin, China

***Abstract*—To some extent, the information from social network can have an impact on the investment decisions of people. Based on behavioral finance theory, this thesis focuses on the influence of social networks on the stock market and forecast the stock price. Firstly, it collects data from the social network platform, and explores the relationship between social network information and stock trading volume and stock fluctuations. Secondly by means of SVM algorithm, then a stock price forecasting model is established to predict the stock price, which verifies that there is a positive correlation between stock market and social network information. Finally, the experimental results prove that the accuracy of SVM algorithm is high.**

***Keywords*—Social Network; Stock Prediction; Support Vector Machine (SVM) ; Regression**

I. INTRODUCTION

The stock market of China has been developed for more than 20 years, which plays an increasingly important role in the world financial system. The analysis and prediction of stock market has also become a focus. Behavioral finance theory suggests that market price, income and resource allocation will be affected by psychological, social, cognitive and emotional factors of investors, and the stock price in the stock market depends not only on its intrinsic value, but also on the influence of various factors. With the popularization and development of the Internet, many social network platforms have emerged, such as Weibo, blog, forum, space, community and so on. People can use these web platforms to express their views freely, and also get opinions from others. These comments and opinions from social network information will influence people to make some investment decisions in the stock market. To some extent, the amount of visitors to a stock on the Internet platform can reflect the degree of interest in stocks by social media. The positive and negative reviews of a stock can reflect the emotional attitude towards the stock market. The potential relationship between the social network data and stock price fluctuation is of great practical significance to stock investment.

Based on the view of data mining, this paper uses SVM to analyze the potential relationship and influence between social network information diffusion and stock market. It is of great practical significance to the economic development of China, and provides an effective basis and support for investors to make better predictions and decisions. A stock named China medium term (000996) was selected as object. We collect the historical data on all posts from the forum of East Wealth Stock Bar and the stock index of China Medium-Term in 2017 from the NetEase Financial website. The SVM model is established to explore the relationship between the information of social network and the stock trading volume and the stock price. Then the SVM algorithm is used to establish the stock price prediction model to predict the stock price. Finally, the experimental results are obtained and analyzed. The organization structure of this paper is as follows. The second part is related research status. The third part is the research methods. The fourth part is the research results and analysis. The fifth part is the conclusion.

II. RELATED WORKS

Social network information has a certain relationship with the behavior and decision-making of people, and there are more and more researches to use data mining to analyze the online behavior of Internet users. Daniel Gruhl successfully predicted the book sales by analyzing content in blogs [1]. Sitaram Asur gets comments on Twitter to predict the sales of movie [2]. Huang Jiuming uses AC-Trie technology to study some hot topics and emergencies [3]. Many researchers have developed a variety of data visualization software, and users can easily transform the complex big data into a chart straightly in order to analyze the structure of social network better.

At present, there are many studies to prove that there is a correlation between the stock information on social networks and changes in stock market. There is a lot of literature on the relationship between the behavior of online user and the stock market. Tetlock uses the method of value at risk to compile

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

and analyze the relationship between the comments of the Wall Street Journal on stocks and the stock market [4]. The research results show that abnormal emotional index will affect the trading volume of the stock market on the same day. Sabherwal proposes a cross-section analysis method. It proves there is a positive correlation between the number of comments on the Internet and the return on the stocks of the same day [5]. Zimbra uses the stock forum on Wal-Mart to predict the volatility of stock price [6]. Bollen et al. find that the three-day cooldown index in the sentiment index of Twitter big data was linked to the Dow Jones Industrial Average index (DJIA) consistently with the measure of user bias, positive and negative emotions [7].

In recent years, with the rapid development of Internet financial industry in China, the research on Chinese stock forecasting has increased. In investor sentiment measurement, Guo Xiaofi et al. find that the correlation between investor sentiment index and financial asset price trend by random forest method is higher than that of ridge regression method [8]. Pang Lei et al. used the Sina Weibo platform to analyze comments on stock topics posted by users on the Internet platform and to predict sentiment trends of investor in the future [9]. The results show that this method has a good recognition effect on investor sentiment.

In general, the research on the relationship between social network information and stock market is a relatively hot topic. Many studies have proved the feasibility and effectiveness of using social network platform to explore stock market prediction.

III. Stock Prediction in China Based on Social Network and SVM Algorithm

A. Data collection

The data source of social network in this paper is from the stock bar forum of East Wealth Network. East Wealth Network is the largest and most influential financial website in China. The website covers all aspects of finance and economics. In addition, there is a separate region in the forum of East Wealth Network for each stock in the market. When the users want information on a particular stock, they can search the stock in the forum. In this paper, we collect the information of all posts from January 1, 2017 to December 31, 2017. The selected attributes include number of readings, number of comments, titles of posts, author and date of publication. The collected data is saved into the excel form and 14803 pieces of data are obtained. In addition, the data of stock index used in this paper include the opening price, closing price, the highest price, the lowest price, rate of rise and fall, and so on, from the NetEase Finance and Economics website. It summarizes the stock index data for each day of the year, the number of posts per day, the number of reading times for each post, and the number of comments for each post. There are 244 trading days removing the holidays and weekends, and there are 244 samples.

By drawing the broken line diagram with the volume of trading volume and the amount of posts, the relationship between the parameters can be reflected more intuitively by the broken line graph. It can be found that the rise and fall of the trading volume often accompany with the change of the quantity of posts in the same direction. On February 17, 2017, the trading volume reached its maximum, and the number of posts on the same day is also the highest in the year. On September 8, 2017, it was the second most traded volume in the stock market, and the number of posts on the stock bar also increased significantly during that time. Therefore, it is concluded that there is a positive relationship between the volume of transaction and the amount of posts on the network. The larger the number of posts is, the higher the attention of network is, and the more volume of transaction is, as shown in figure 1.

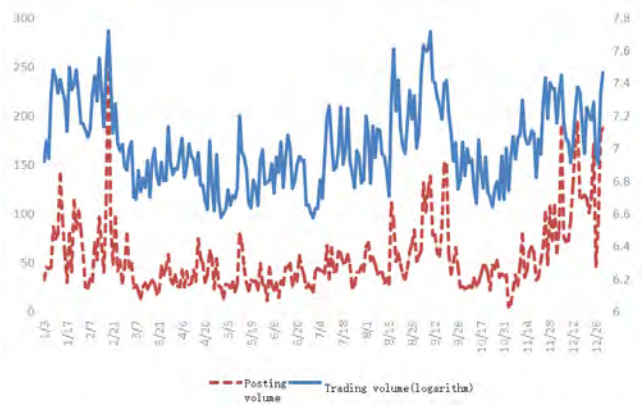

Fig. 1 Time series diagram of post and transaction volume

B. Data pretreatment

The data set is preprocessed, including data format conversion and data normalization. Firstly, the format of the data set is converted, and the macro command in FormatDataLibSVM is executed. FormatDataToLibSVM is used to standardize the data in the excel table and converts the format of data.

Secondly, because the variable data units involved are all different, we consider the data normalization. When the eigenvalue is only a positive number, the data is normalized to the interval of [0, 1], and when the eigenvalue is a positive number or a negative number, the data is normalized to the interval of [-1, 1]. There are two advantages of data normalization. Firstly, it can improve the training speed and efficiency of the model. Secondly, each eigenvalue can be compared numerically and improve the accuracy of classification model. The formula of data normalization is (1).

$$D'(i) = (U - L) * \frac{D(i) - \min(D)}{\max(D) - \min(D)} + L \qquad (1)$$

Here, U and L are upper bound and lower bound of the range, D'(i) and D (i) are normalized data and pre-normalized data respectively. In the folder of the libsvm toolkit, there is a svm-scale.exe program that can be executed with svm-scale command. Execute the svm-scale command to get the normalized data.

C. Algorithm Realization

The basic condition of modern financial theory is the hypothesis of rational market and the assumption that people are rational. However, in real life, investors and markets are not rational, and the economic behavior of investors reflects the rational deviation. Behavioral finance theory breaks the rational hypothesis and explains financial phenomenon in terms of the activities of individual participants in micro market and the psychological changes that produce these behaviors. Behavioral finance theory has been applied to stock investment by many scholars. The theory holds that the psychology and behavior of investors have a significant impact on the change of stock price. Changes in stock prices are considered nonlinear and can be used as a nonlinear classification or regression problem. Support vector machine (SVM) is widely used in the classification and prediction of nonlinear systems because of its good learning ability and generalization ability. Therefore, SVM algorithm is selected for classification and prediction in this paper.

The accuracy of SVM model for classification problem is very high. The more accurate model can be trained by selecting proper kernel function. SVM algorithm is based on the linear partitioning. It maps the low-dimensional points to high-dimensional spaces so that they can be linearly partitioned. In the field of machine learning, SVM can be easily used in classification and regression analysis. We select the training data set at first. For the classification problem, the SVM algorithm establishes a model according to the training set. Then the test set is assigned as the corresponding categories according to the model. For regression problems, the SVM algorithm will determine the target value of the test data. In this paper, LIBSVM is used to establish SVM prediction model, and the test software is Matlab.

Firstly, this paper analyzes the correlation between the information of social network and stock trading volume. It classifies the training set with SVM algorithm, establishes the model and predicts the data sets. Secondly, it analyzes the correlation between the information of social network and the stock price, and establishes the model by using SVM. Finally, the SVM regression algorithm is used to establish prediction model to predict the stock price.

IV. Experimental Results and Analysis

A. Analysis on the correlation between social network information and stock trading volume

The experiment uses 200 data from January 1, 2017 to October 30, 2017 as the training data set, and uses SVM model to train the model of stock trading volume and network information. The 44 data from October 31, 2017 to December 29, 2017 are taken as test data sets, and the accuracy of the model is calculated. The label marker for trading volume greater than 10,000,000 vectors is 1, and the label marker is -1 for trading volume less than 10,000,000 vectors. In terms of posts, readings, comments, opening prices, closing prices, highest prices, lowest prices, the fluctuation rate, the turnover rate, transaction amount are taken as the features.

The training set and SVM train function are used to construct the model, and the parameters are selected as C-SVC model and RBF kernel function. Then, the SVM predict function and the established model are used to predict the test set, and the prediction accuracy is 90.9091%, which shows that the information of social network has a significant impact on the trading volume of stock.

B. Analysis on the correlation between social network information and stock price

Use 200 data from January 1, 2017 to October 30, 2017 as the training set, and use 44 data from October 31, 2017 to December 29, 2017 as the test set. The posts, readings, comments, opening prices, closing prices, maximum prices, the lowest price, trading volume, turnover rate and the transaction amount are taken as the characteristic. The stock rise and fall of next day are taken as the target value. If the stock of next day is rising, then the definition of target value is 1; if the stock falls on the next day, the target is defined as zero. We also select the C-SVC and RBF kernel functions to build the model and get the prediction results. The accuracy was 65.1163%. This shows that the impact of social network information on stock prices is not significant.

C. Stock Price Prediction based on SVM

In this paper, it predicts the opening price of Chinese stocks in the next day, and selects the opening price, the closing price, the highest price, the lowest price, trading volume, the number of posts, the number of reading and the number of comments as the characteristics. The samples of first 200 days of 244 trading days are also selected as the training set to forecast the opening price for 44 days. Because the target is the opening price of the next day, the data of last day is discarded and 43 forecast values are available.

The prediction of stock price can be regarded as a regression problem. In this experiment, the type of SVM model is epsilon-SVR, chosen to solve the regression problem. The kernel function selects the RBF kernel function, and in the regression process, SVM train and SVM predict function are used to predict the results as shown in Table 1.

TABLE I PREDICTION OF STOCK PRICES BY SVM ALGORITHM

date	initial price	forecast price
2017/10/31	13.45	13.5697
2017/11/1	13.78	13.8348
2017/11/2	13.5	13.8115
2017/11/3	13.8	13.854
2017/11/6	14.3	14.179
2017/11/7	14.36	14.4164
2017/11/8	14.34	14.5006
2017/11/9	14.59	14.5536
2017/11/10	14.84	14.9453
2017/11/13	14.7	14.806
2017/11/14	14.76	14.7478
2017/11/15	14.63	14.8494

Table I, cont

2017/11/16	14.26	14.5159
2017/11/17	13.76	14.108
2017/11/20	13.68	13.8062
2017/11/21	14.12	14.0375
2017/11/22	14.26	14.1745
2017/11/23	14.3	14.5117
2017/11/24	14.95	14.8326
2017/11/27	14.73	14.8182
2017/11/28	15.35	15.1572
2017/11/29	15.61	15.4075
2017/11/30	14.7	15.1692
2017/12/1	15.1	14.8716
2017/12/4	14.49	15.0955
2017/12/5	13.58	14.2671
2017/12/6	13.66	13.7746
2017/12/7	13.79	13.9246
2017/12/8	14.11	14.0822
2017/12/11	14	14.2863
2017/12/12	14.24	14.2401
2017/12/13	13.86	14.4297
2017/12/14	14.61	14.5161
2017/12/15	14.52	14.8481
2017/12/18	14.76	14.789
2017/12/19	14.63	14.8339
2017/12/20	14.6	14.8904
2017/12/21	14.89	14.8658
2017/12/22	14.58	14.9085
2017/12/25	13.89	14.406
2017/12/26	14.11	14.1261
2017/12/27	14.78	14.2525
2017/12/28	14.68	14.842

According to the results of table 1, the average error rate is 1.42%. The formula for the average error rate is expressed as: (original price-forecast price) / forecast price. Thus, the use of SVM to predict the stock price reflects a good performance.

This paper mainly studies the correlation between stock trading volume, stock fluctuation and social network information. It establishes a prediction model of stock price based on stock index and social network information. According to the experimental results, we can conclude that there is a positive correlation between the social network information and stock market. The greater the trading volume of a stock is, the more online financial data is, including the posts, comments, and readings. The stock price forecasting model established by epsilon-SVR model in LIBSVM tool can obtain high accuracy, and LIBSVM can be used for model training and model prediction by choosing the best parameter. The predicted value can approximate the true value. The prediction effect is good, and the operation is simple.

V. CONCLUSION

Firstly, based on SVM model, it analyzes the relationship between social network information and stock market in this paper. It mainly analyzes the influence of social network information on stock trading volume and stock fluctuation. Secondly, a forecasting model of stock price is established to predict the short-term price of a stock in order to observe the change of stock market in a relatively stable period of time. Experimental results show that the SVM algorithm can achieve better prediction results.

There are still some shortcomings in this study, such as some invalid comments on the network, which may interfere with the constructed model. In addition, only social network posting data of one year in 2017, data samples are small. Future research can be considered to increase the number of web site and the collection of multi-year posts. In this paper, it selects the opening price, closing price, the highest price, the lowest price, trading volume, number of posts, reading quantity as the characteristics. In the future research, we can consider to add more financial indicators such as net earnings per share of the stock company, rate of return on total assets and rate of return on net assets, to improve the accuracy of the model prediction.

ACKNOWLEDGMENT

The research is supported by the Tianjin Science and Technology Development Strategy Research Plan of China (18ZLZXZF00550), 2018 Key Research and Development Projects in Tianjin (18YFZGX00670), Tianjin Natural Science Foundation of China (17JCYBJC43300) and Tianjin Cultural and Arts Social Science Plan Project (d16005).

REFERENCES

[1] D. Gruhl, Bender. “Information hiding to foil the casual counterfeiter,” Lecture Notes in Computer Science, vol. 1525, pp. 1-15, 1998.

[2] S. Asur. “Predicting the Future with Social Media,” IEEE/WIC/ACM International Conference on Web Intelligence & Intelligent Agent Technology, 2010.

[3] J. M. Huang, Q. Y. Wu, S. D. Zhang, et al. “Mining hot phrases in online social network text streams based on ACTrie,” Journal of Electronics, vol. 44, no. 10, pp. 2466-2470, 2016.

[4] P. C. Tetlock. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market,” The Journal of Finance, vol. 62, no. 3, pp. 30, 2007.

[5] S. Sabherwal, S. K. Sarkar, Y. Zhang. “Do Internet Stock Message Boards Influence Trading? Evidence from Heavily Discussed Stocks with No Fundamental News,” Journal of Business Finance & Accounting, vol. 38, pp. 1209–1237, 2011.

[6] D. Zimbra. Stakeholder and sentiment analysis in web forums,” Stakeholder & Sentiment Analysis in Web Forums, 2012.

[7] J. Bollen, H. Mao, X. Zeng. “Twitter mood predicts the stock market,” Journal of Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[8] X. F. Guo. “Some questions about the implementation of equity incentive mechanism in China,” Market Modernization, vol. 21, pp. 91-91, 2007.

[9] L. Pang, S. S. LI, G. D. Zhou. “Sentiment Classification Method of Chinese Micro-blog Based on Emotional Knowledge,” Computer Engineering, vol. 38, no. 13, pp. 156-158, 2012.