Machine Learning and Artificial Intelligence Applied in the Research of the Emotions Impact on Forecasting

Yaokui Jiang*, Yi Zuo and Yudong Yang
School of Economics and Management, Xidian University, Xidian University, Xi’an, China
*Corresponding author: 2003010209@st.btbu.edu.cn

Abstract. According to the behavioral finance theory, the psychology and behavior of investors in capital market have an important impact on the fluctuation of stock index. Therefore, this paper assumes that there is a certain internal mechanism between investor sentiment and stock index, which can predict the overall price change of the stock market. The text mining technology and sentiment analysis method can be used to generate a total of six category positive and negative investor sentiment time series data, with three orders; The methods such as unit root test, Granger causality test and factor analysis are used to construct the composite index of SSE investor sentiment, and then the support vector machine and neural network are adopted to predict the stock market price changes and conduct the hypothesis verification. The results indicate that the SSE investor sentiment composite index constructed by using the text data of online stock market forum and stock transaction data can improve the accuracy of stock index trend prediction, which is conducive to better decision-making of the government, online platforms, listed companies, and investment entities.

Keywords: Investor sentiment, Stock forecast, Text mining, machine learning.

1. Introduction

Financial market forecasting methods include multiple regression model and self-vector regression model classical statistics, as well as modern machine learning methods such as support vector machine (SVM), neural network, and random forest, and among them, the SVM and BP neural network are most widely used. The traditional regression analysis is based on strict assumptions and sufficient priors, so it is difficult to construct an effective financial forecasting model. Machine learning can automatically learn, repeatedly improve and optimize the algorithm, and the results are satisfactory. Many literature studies on financial market association or prediction based on text mining directly add single-dimensional emotional variables (positive or negative emotions) into the model, and fewer researcher process the nonlinear and high-noise emotional data, so it is easy to verify whether it is associated with the financial market and difficult to achieve good prediction effect. Through grasping the Oriental wealth BBS data, and drawing lessons from financial correlation analysis process of the weather or event, this study not only eliminates the neutral or noise data, but also select emotion data with strong correlation to involve in investor sentiment index design. Based on the nonlinear characteristics of the emotion data and stock index, the SVM and BP neural network
model are used for stock index prediction. It proves that there is an internal relationship between investor sentiment and stock index, and the prediction is efficient, which is expected to provide a good reference value for the decision support of investors, listed companies, and government regulatory authorities.

2. Stock index prediction based on machine learning

The stock index prediction based on machine learning includes four parts: the pre-processing and the stability test of the stock index and sentiment data, the construction and data generation of the forecast combination index, and the test of the machine learning algorithm for the two commonly used indexes prediction.

2.1. Stock index data acquisition and sentiment data preprocessing

(1) Acquisition and preprocessing of emotional data. The investor sentiment text data was derived from the East Fortune Stock Bar Forum, and Python was used to capture a total of 368586 posts, with a time span of from July 19, 2016 to December 29, 2017. A total of 217,445 posts were retained by writing post cleaning rules to eliminate topics that could not express investor sentiment. Cleaning rules included pictures (no text), links (no text), garbage characters (no meaning) and the combination of firm offer (system automatic generation); In terms of text emotion classification, a dictionary-based Chinese sentiment analysis method [28] was used to score the emotions of posts. The dictionary consisted of three kinds of words: emotion words, degree adverbs, and negative words. According to Formula (1), the comprehensive emotion score of posts was calculated. Emotional words included general emotion dictionary and special emotion words (slight decline, positive, and cut meat).

\[
PostScore = W_r \cdot W_m \left\{ \sum_{i=1}^{m} \left[ \left( \prod_{j=1}^{n} W_d \right) \cdot \left( \prod_{j=1}^{n} W_n \right) \cdot W_s \right] \right\}
\]

Where PostScore is the emotional comprehensive score, \( m \) is the number of emotional words of the title of a post, and \( n \) and \( n_n \) are the number of degree adverbs and negative adverbs in front of the \( i \)th emotional word, respectively; \( W_s, W_m \) and \( W_r \) are the score values of each sentiment word, punctuation mark, and rhetorical word corresponding to the title of the post. \( W_d \) and \( W_n \) are the scores of degree adverbs and negative adverbs in front of the corresponding affective words, respectively.

2.2. Stability test of stock index data and sentiment data

(1) Data standardization. To eliminate the dimensional relationship between stock trading data and investor forum sentiment data and improve data comparability, the two types of data need to be standardized (ZScore) according to Formula (2), where \( \mu \) is the mean value of sample data, and \( \sigma \) is the standard deviation of sample data.

\[
z = \frac{x - \mu}{\sigma}
\]

(2) Unit root test. Through the random walk test of the moment of time series, the errors of statistical data and the false regression of the model are eliminated to ensure the stability of the prediction model. If there is no unit root, the time series will be stable. In this paper, ERS (Eliot, Rothenberg and Sockets Point Optimal Test) was used to test the unit root, avoiding constant and trend variables in the test.

The test results (see Table 1) showed that the ERS test statistics of the four time series variables (SSEC, OPEN, HIGH and LOW) were greater than the critical value when the confidence was 10%. These time series variables contained unit roots and were non-stationary.
Table 1. Unit root test of time series.

| Index  | SSEC  | OPEN  | HIGH  | LOW   | VOL   | AMO   | PI    | PIH   | PIII  | NI    | NH    | NIII  |
|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| ERS value | 17.2325 | 13.5814 | 21.4395 | 15.1403 | 0.8254 | 0.5192 | 4.2104 | 2.4557 | 0.5254 | 1.8937 | 1.0911 | 0.2846 |

(3) Differential time series unit root test. After the first-order difference operation of all variables according to Formula (3), new sequence variables were obtained, which were denoted as: DSSEC, DOPEN, DHIGH, DLOW, DVOL, DAMO, DPI, DPII, DPIII, DNI, DNII, DNIII, and $X_t$ and $X_{t-1}$ are the variable values of time period $t$ and $t-1$, respectively.

$$D(X) = X_t - X_{t-1}$$ (3)

The unit root test of each time series after first-order difference (see Table 2) found that the ERS statistical values were all less than the critical value when confidence was 1%, with the maximum ERS value of 0.233, and each time series tended to stable.

2.3. Select relevant data to generate composite index data

(1) Correlation analysis. After the difference of historical trading data variables of Shanghai Composite Index, Pearson correlation analysis method is used to find that the variables influenced each other and there is a correlation among them (see Table 3), so it can be used for effective index prediction. In this paper, five variables, including DOPEN, DHIGH, DLOW, DVOL and DAMO, were selected to construct the composite index of Shanghai Stock Exchange.

Table 2. Unit root test of differential time series.

| Index  | DSSEC  | DOPEN  | DHIGH  | DLOW  | DVOL  | DAMO  | DPI   | DPII  | DPIII | DNI   | DNII  | DNIII |
|--------|--------|--------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| ERS value | 0.1836 | 0.1836 | 0.1401 | 0.1734 | 0.1257 | 0.1281 | 0.1670 | 0.0372 | 0.0437 | 0.0037 | 0.2326 | 0.0091 |

Table 3. Correlation coefficient matrix among variables.

| Differential trading index | DSSEC | DOPEN | DHIGH | DLOW | DVOL | DAMO |
|---------------------------|-------|-------|-------|------|------|------|
| DSSEC                     | 1     | 0.085 | 0.557** | 0.599** | 0.064 | 0.087 |
| DOPEN                     | 0.085 | 1     | 0.704** | 0.557** | 0.314** | 0.356** |
| DHIGH                     | 0.557** | 0.704** | 1     | 0.626** | 0.476** | 0.516** |
| DLOW                      | 0.599** | 0.557** | 0.626** | 1     | 0.052 | 0.005 |
| DVOL                      | 0.064 | 0.314** | 0.476** | 0.052 | 1     | 0.979** |
| DAMO                      | 0.087 | 0.356** | 0.516** | 0.005 | 0.979** | 1     |

(2) Granger causality test. Assuming that investors are susceptible to the sentiment of other investors and choose an irrational investment, it is necessary to conduct a Granger causality test on the Shanghai Composite Index and six groups of investor sentiment time series to analyze and verify whether the investor sentiment changes are correlated with market volatility and whether the index information can be predicted. Granger causality test is explained as follows: whether variable $x$ is the cause of variable $y$, and it can be explained by observing the extent to which current $y$ can be explained by past $x$. If the lag value of $x$ can improve the interpretation degree of $y$, it means that $x$ is conducive to the prediction of $y$, and $y$ is caused by the Granger causality of $x$. Although the results of the Granger causality test are not equal to the actual causality, the purpose of this paper is not to test the actual causality, but to test whether there is predictive information of the time series of the SSE index in the investor sentiment time series.
Table 4. Granger causality test results.

| Delay days | DPI    | DPII   | DPIII   | DNI    | DNII   | DNIII   |
|------------|--------|--------|---------|--------|--------|---------|
| 1          | 0.0313* | 0.5365 | 0.5076  | 0.1160 | 0.1114 | 0.9777  |
| 2          | 0.0392* | 0.2329 | 0.7138  | 0.0860 | 0.0788 | 0.9891  |
| 3          | 0.0215* | 0.0598 | 0.4623  | 0.0559 | 0.1404 | 0.4959  |
| 4          | 0.0424* | 0.0943 | 0.5424  | 0.0814 | 0.1033 | 0.5571  |
| 5          | 0.0470* | 0.0694 | 0.6458  | 0.1316 | 0.1576 | 0.6503  |

(3) Factor analysis and index construction. In this paper, the multi-dimensional index method was adopted to avoid the problem of proxy bias and insufficient information when investor sentiment was used as a proxy variable of a single index. Through the factor analysis of six variables (DOPEN, DHIGH, DLOW, DVOL, DAMO, and DPI), the SSEC InvestorSentiment Index (SSECISI) was obtained. To verify the high efficiency of investor sentiment on stock index prediction, DPI was removed from the SSECISI, and only five variables, including DOPEN, DHIGH, DLOW, DVOL, and DAMO, were used to construct SSEC Portfolio Index (SSECPI). Then the principal component analysis was used to first carry out the maximum orthogonal transformation of variance on the factor load matrix to obtain the factor score (Formula 4) and the variance contribution rate (see Table 5). Later, the SSECPI and SSECISI data were obtained according to the weighted average of the factor score and variance contribution rate (Formula 5).

\[ F_j^\prime = \beta_{1j}X_1 + \beta_{2j}X_2 + \cdots + \beta_{pj}X_p, j = 1, 2, \ldots, m \]  
(4)

Where \( F_j \) is the factor score of factor \( j \), and \( \beta_{pj} \) is the factor score coefficient of component \( X_p \).

\[ F = \left( F_1V_1 + F_2V_2 + \cdots + F_jV_j \right) / \sum_{j=1}^{m} F_j, j = 1, 2, \ldots, m \]  
(5)

Where \( F \) is the comprehensive score, namely, the index constructed in this paper, and \( V_j \) is the contribution rate of factor \( j \).

3. Technical analysis

3.1. Text mining technology

The unstructured network data dominated by text format is said to account for more than 80% of the total global data volume, including e-mails, documents, reports, forms, call records, press releases, blogs, microblogs, WeChat, Q&A, forums, comments, etc., while the pure digital data accounts for less. Text mining has become a new technology of business analysis, which is used to observe all kinds of business behaviors and their effects. In this paper, the preconditions for the prediction effect analysis are a series of text mining technologies, including text data collection and cleaning, text data segmentation, text sentiment dictionary construction, text data sentiment scoring, as well as sentiment data standardization.

3.2. Machine learning prediction technology

Machine learning technology is used to solve conventional nonlinear problems. In this paper, both the stock index and text data are nonlinear, so they aren’t suitable for the application of smooth class prediction model, but the BP neural network and SVM models were selected to carry out stock index prediction, and it was proved to be a more suitable model. The result showed that the SVM algorithm was superior to the BP neural network, and other application scenarios might also be the opposite. To observe the influence of duration, three groups of data with different duration were used in the prediction process, and the results showed that short-term prediction had better effects. This indicates that the prediction research based on nonlinear text data needs to investigate the multi-dimensional
situation such as method, model and duration, and more complex data can be predicted by combining machine learning with wavelet analysis. As for complex nonlinear data sources, the scientific and precise prediction needs a better benchmark database and algorithm. The artificial participation of iFlytek in the event of machine simultaneous interpretation shows that machine learning isn’t an ideal algorithm at present, and the process of personalized speech and specialized vocabulary training is ignored when the machine is required to real-time random simultaneous translation. If there is no large corpus, machine learning won’t be competent for random problems (untrained dialects, terms, loanwords, etc.). If noise reduction and optimization of raw data are allowed in advance, delayed machine learning will be better. In addition, machine learning is widely used in artificial intelligence today. The realization approach is to perfect professional database and scenario applicable algorithm, such as acceptable network translation and speech recognition and other universal services, as well as investment and financial services of multiple data sources.

4. Conclusions
By capturing the emotional text of the network forum, the financial professional vocabulary was extracted for text mining, so as to realize the specialization and precision of the text mining data. Then the correlation analysis method was used to construct the investor sentiment composite index and eliminate the bias directly using sentiment data for prediction. Later, the machine learning method was used to design a good stock index prediction model, in order to improve the accuracy of stock index trend prediction. It was proved that the SVM-based composite index model of the SSE stock index was more effective in the stock index prediction.

References
[1] Wang Wen Bo, Fei Pu Sheng, Yi Xu Ming. Prediction of Chinese stock market based on EMD and neural network [J]. Theory and practice of systems engineering, 2010, 30 (6): 1027-1033.
[2] Lin Chunmei, he Yue, Tang Bingyong, et al. Application of fuzzy cognitive map in stock market forecasting [J]. Computer applications, 2006 (01): 195-0197.
[3] Huang Runpeng, Zuo Wenming, Bi Lingyan. Stock market prediction based on micro blog sentiment information [J]. Journal of management engineering, 2015, 29 (1): 47-52.
[4] Yin Guangwei. Research on forecasting methods of Chinese stock market [D]. Tianjin University.
[5] Zhang Yong'an, Yan Bin Bin. A deep learning compound forecasting model for stock market [J]. Computer science, 2020, v.47 (11): 263-275.
[6] Zhou Zihui. Research on stock price forecasting based on independent component analysis and deep learning [D]. Southwest University of Finance and economics, 2019.