Study of Random Forest and XGBoost Quantitative Stock Selection Strategies

Youen Xu[1,*], Yang Luo[1]

[*]E-mail address: 2581433698@qq.com

University of South China, School of Computer Science[1]

Abstract. The data of all A-share stocks in the Chinese stock market with regular trading in the sample period from 01/01/2015 to 12/31/2021 were analyzed. Factor IC and other methods were used to identify factors that significantly affected the next period return of stocks, and random forest and XGBoost stock selection models were constructed based on the results of factor selection and back-tested on historical data. The two strategies were combined and back-tested using the equal-weighted portfolio and 60/40 portfolio idea methods, respectively, several tests such as annualized return, Sharpe ratio, and maximum retracement rate were compared and analyzed between the combined strategy and the single strategy. The results showed that the two portfolio strategies were more effective than the single strategy in risk control and stable returns, that had some reference value for theoretical research on quantitative stock selection methods.

Keywords: Quantitative stock selection; Machine learning; Strategy portfolio

1. Introduction

Nowadays, the domestic securities and futures markets are developing rapidly, and the traditional investment techniques are performing generally and can hardly match the complexity of financial markets. More investors are turning their eyes to a more rational investment track -quantitative investment strategies- in order to obtain more valuable stocks for investment[1]. Machine learning algorithms and big data AI continue to develop in various aspects and are now also widely used in stock market problems, providing theoretical and technical support for the establishment of quantitative stock selection models. For example, Weixing Wu[2] combined random forest with technical indicators and the cumulative return can exceed the CSI 500 index. Che Yang[3] investigated the stock selection performance of three strategies, logistic, daboost and random forest, respectively. Yuanhao Jin et al.[4] applied the K-NN algorithm to factor extraction to enhance the economic explanatory power of the model. Maojun Zhang et al[5] used decision tree method for quantitative timing strategy to obtain more robust investment performance. Yi Fang et al.[6] deep dive into more influential factor indicators and predict stock returns by eight machine learning algorithms.

Using machine learning to build stock picking strategies is a popular trend nowadays. However, compared with selecting high-quality individual stocks, good asset allocation is the general direction to grasp the benefits of investment. The study of broad asset allocation theory began in the 1930s[7], in which the traditional Constant Mix Strategy refers to keeping a fixed proportion of various types of assets in a portfolio such as stocks, futures, etc.[8]. Many

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China
Copyright © 2023 EAI
DOI 10.4108/eai.19-5-2023.2334374

scholars at home and abroad have conducted a lot of research on constant mix strategy and achieved good results[9].

It remains difficult to develop a single strategy with stable returns and small retracements by machine learning algorithms. In order to investigate the optimization effect of the combined strategy approach on a single strategy, we empirically study the all-A market stocks using random forest and XGBoost stock selection models, respectively, and used the asset allocation idea of constant hybrid strategy to combine the two machine learning stock selection strategies for backtesting, so as to improved the return of a single strategy and reduced the volatility of the strategy, and finally compared this combined stock selection strategy with a single stock selection strategy.

2 Research Design Framework

The multi-strategy combination approach is a process of combining strategies based on the prediction results of multiple single stock selection strategies by assigning different weights to each strategy and obtaining a new and better backtesting effect. The design framework of this study is shown in Figure 1.

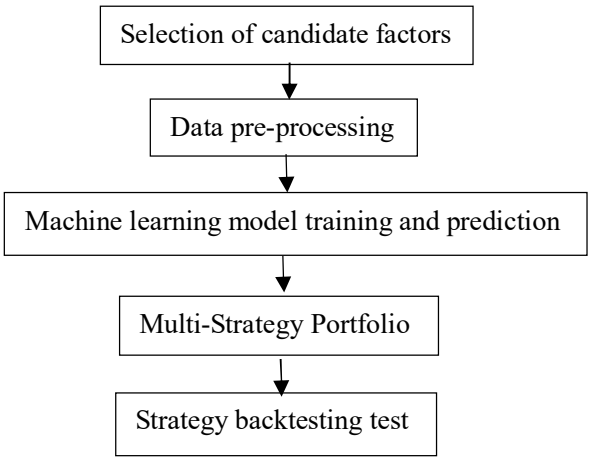

Figure 1 Design framework for multi-strategy combinations

2.1 Selection of candidate factors

In terms of factor selection, some may choose fundamental analysis factors[10-11], some may also choose technical analysis factors[12-13], or factors such as investor sentiment[14-15] to construct factors. However, in general, choosing a more effective combination of characteristic factors can directly improve the performance of the model and is the key to the success of stock selection strategy design. Based on the theory of financial economics, this paper selects three types of factors from the BigQuant factor pool, namely, quantitative, valuation, and financial, as candidate factor pools.

2.2 Data pre-processing

The raw data obtained cannot be analyzed directly, as there are usually noise, missing values and inconsistent magnitudes in these data. After obtaining the raw data, it is necessary to pre-process the "dirty data" and eliminate some unreasonable values according to the model requirements before inputting them into the model to avoid interfering with the model work. The data can be processed in many ways, such as extreme value processing, standardization processing, etc., and can be appropriately filtered and extracted according to the research needs.

2.3 Machine learning model training and prediction

Two machine learning algorithms, Random Forest and XGBoost, have good performance in stock investment applications and are the most commonly used stock selection models by many researchers. After obtaining the factor data, a standard random forest and XGBoost model are constructed respectively, and the two models are trained and learned in the training set separately, and then the new test data are put into the two trained machine learning models separately for prediction.

2.4 Multi-strategy combination

When studying quantitative investment strategies, one cannot rely on a single strategy alone; one can try to study and run multiple strategies. Using the idea of equal-weight portfolio and 60/40 portfolio in traditional asset allocation strategies, respectively, two machine learning stock selection strategies are allocated to total assets according to equal proportion and 60/40 weighting to achieve a combination of multiple strategies as a way to diversify the overall risk and smooth the return of the portfolio.

2.5 Strategy backtesting test

The BigQuant platform is used to backtest the trading of the machine learning stock picking strategy, and then the multi-strategy combination method is used to backtest the combination of the two strategies to obtain the strategy performance and test the return and risk of the multi-strategy versus the single strategy.

3 Experiment and Analysis

3.1 Validity test of the factors

Factor analysis was conducted on three types of factors in the BigQuant platform: quantitative, valuation, and financial. Through research and comparison, nine quantitative price factors, three valuation factors and two financial factors were selected respectively. The test data were selected from the whole A-share market data, and the test interval was from Jan. 01, 2015 to Dec. 31, 2020, and the factors were analyzed mainly from two perspectives of validity and stability.

To determine whether a factor is valid, we mainly look at the linear correlation between the ranking of all stocks at the beginning of the transfer cycle and the ranking of returns at the end of the transfer cycle, which is the Information Coefficient (IC)[16], and represents the factor's

ability to predict stock returns. The stability of a factor is judged by the value of the multi-period mean of IC/standard variance of IC, also known as the Information Ratio (IR)[16], which represents the factor's ability to obtain stable Alpha. The factor validity test is shown in Table 1.

3.2 Data selection and processing

In order to make the data more reasonable to reduce the adverse effects on the trading model, the full A-share market data are screened according to the rules in Figure 2.

After eliminating 1080 stocks that do not meet the conditions, the final sample pool has a total of 3149 stocks.

When constructing quantitative trading models, the quality of data is as important as the quantity. Unreasonable data is not conducive to the objectivity, validity and scientificity of the trading model. Therefore, missing values in the database are eliminated, individual extreme values are removed, and the data are standardized so that data of different magnitudes are reduced to the same interval by different proportions.

Table 1 Factor validity test table

Category	Indicator Name+	IC Mean Value	IR Value	Linear relationship
Volume and price factors	Earnings for the past 5 days	-0.05	-0.54	Significant negative correlation
	Earnings for the past 10 days	-0.06	-0.61	
	Earnings for the past 20 days	-0.07	-0.66	
	5-Day Earnings Ranking	-0.05	-0.52	
	10-Day Earnings Ranking	-0.06	-0.58	
	Average transaction value for the day	-0.1	-1.20	
	5-day average trading volume	-0.1	-1.13	
	Ranking of the day's average trading volume percentage	-0.12	-1.23	
	5-day average trading volume percentage ranking	-0.12	-1.18	
Valuation Factor	Total Market Capitalization	0.02	0.63	Positive correlation
	P/E ratio	0.06	0.83	Significant positive correlation
	EP	0.06	0.77	
Financial Factor	Net Profit	-0.03	-0.5	negative correlation
	Operating income	0.05	0.53	Significant positive correlation

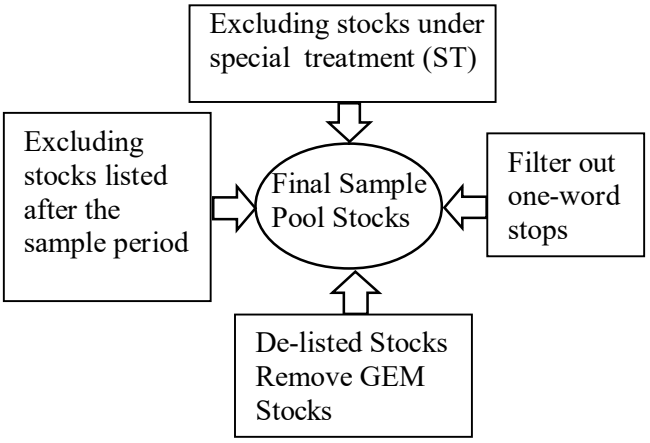

Figure 2 Data filtering rules

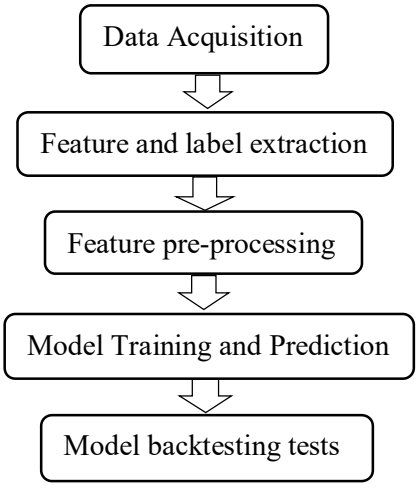

Figure 3 Diagram of the main steps of the random forest and XGBoost strategy for stock selection

3.3 Build a machine learning stock selection strategy

The data of 4229 stocks in the full A-share market from 01/01/2015 to 12/31/2021 for a total of 2556 trading days in the BigQuant platform were selected as samples. The training data set was the full A-share data from Jan. 01, 2015 to Dec. 31, 2020, and the test data set was the full A-share data from Jan. 01, 2021 to Dec. 31, 2021. The future five-day returns of individual stocks were used as labeled values, and the labeled values were divided into 20 categories according to the equal interval of returns, and random forest and XGBoost models were constructed for training and prediction. The main steps of stock selection are shown in Figure 3.

The training set data was used as a sample to train the machine learning algorithm model, and then the test set is predicted. The time period for backtesting was from 01/01/2021 to

12/31/2021. The CSI 300 stock index was used as the benchmark for measurement. The conditions for stock selection and buying and selling as well as the parameters were set as shown in Table 2.

Table 2 Table of backtest conditions for stock selection by two machine learning strategies

Backtest time	01/01/2021 to 12/31/2021
Stock Pool	All A shares
Positioning Cycle	5 trading days
Single Stocks Weight	Score sort weighted
Initial Capital	1000000yuan
Handling Fee	Buy 0.03%, Sell 0.13%
Maximum share of funds per stock	0.2
Stock Selection Rules	Top 5 highest model forecast scores

The dynamic results of the backtesting of the stock selection strategies constructed by the random forest and XGBoost methods were shown in Fig. 4 and Fig. 5, respectively. From the figures, we could see that the returns of both machine learning stock selection strategies were significantly higher than the benchmark returns, and the overall dynamic returns of the strategies showed a continuous increase.

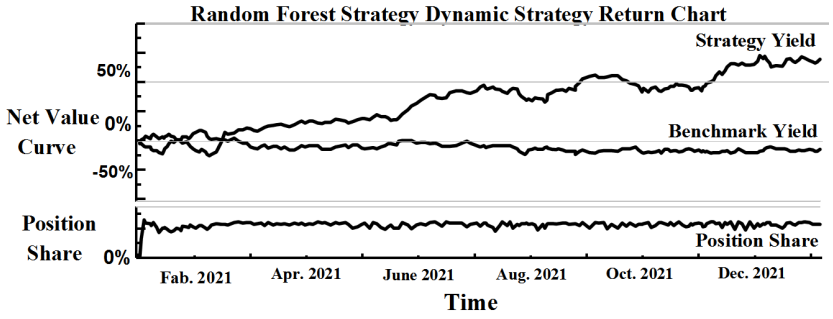

Figure 4 Dynamic strategy gain graph of random forest strategy

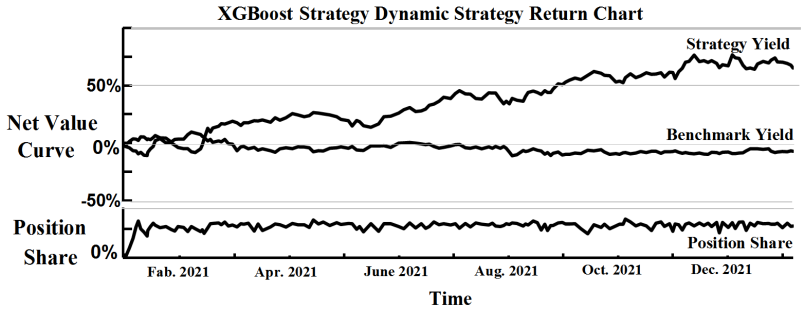

Figure 5 Dynamic strategy return graph for XGBoost strategy

The results of the backtest (Tables 3 and 4) showed that both stock selection strategies achieve high positive returns when the benchmark return was negative and perform better in several indicators such as annualized return, alpha, and Sharpe ratio in the time period from January 1, 2021 to December 31, 2021. This indicated that the constructed random forest and XGBoost models were feasible and effective for stock selection, and both were able to uncover stocks with growth potential. When comparing the results of Random Forest with XGBoost stock selection, the former had a larger Sharpe ratio, which indicates a higher return per unit of risk, and a larger value of the information ratio. The maximum retracement rate of XGBoost stock picking in the backtesting phase was 0.66% lower than that of Random Forest stock picking, indicating that the XGBoost stock picking strategy was slightly more resilient to risk.

Table 3 Summary table of returns for random forest strategy stock selection

Yield	69.61%
Annualized Yield	72.96%
Benchmark Yield	-5.2%
Alpha	0.75
Beta	0.2
Sharpe Ratio	2.29
Earnings to Loss Ratio	1.44
Earnings Volatility	23.92%
Information Ratio	0.14
Maximum retracement	12.18%

Table 4 Summary table of returns for XGBoost strategy stock selection

Yield	65.69%
Annualized Yield	68.82%
Benchmark Yield	-5.2%
Alpha	0.71
Beta	0.17
Sharpe Ratio	2.12
Earnings to Loss Ratio	1.37
Earnings Volatility	24.8%
Information Ratio	0.13
Maximum retracement	11.52%

3.4 Multi-strategy combination backtesting

In backtest trading with Random Forest and XGBoost stock picking strategy, buy and sell orders were generated according to the stock ranking predicted by Random Forest and XGBoost algorithms.

Tables 5 and 6 gave the ranking results of the top 5 stocks selected by the Random Forest and XGBoost strategies on 04/01/2021, respectively, indicating that these 5 stocks had a higher buy value. Table 7 gave the stocks at the bottom of the ranking predicted by the Random Forest and XGBoost algorithms for the stocks held on January 12, 2021, and the backtesting trade was used to generate sell orders.

Table 5 Ranking list of stock predictions for the random forest strategy

	Prediction	Date	Instrument
0	0.472941	2021.1.4	603332.SHA
1	0.200587	2021.1.4	002096.SZA
2	0.195579	2021.1.4	002858.SAZ
3	0.188570	2021.1.4	002095.SZA
4	0.181837	2021.1.4	000007.SZA

Table 6 Ranking list of XGBoost strategy stock forecasts

	Prediction	Date	Instrument
0	1.638985	2021.1.4	002862.SZA
1	1.171866	2021.1.4	002269.SZA
2	1.164781	2021.1.4	603332.SHA
3	0.840780	2021.1.4	002795.SZA
4	0.834364	2021.1.4	601798.SHA

Comparing Table 5 and Table 6, it could be found that among the five high-quality stocks selected by each of the two machine learning algorithms, only Suzhou Longjie (603332. SHA) was co-selected by both, while more of them showed differences in stock selection results. Similarly, among the sell orders in Table 7, Yonghe Zhicheng (002795.SZA) was jointly selected for sale by the Random Forest and XGBoost strategies, but the remaining three stocks eliminated by each were not the same. This showed that even for the same stock data, the stock selection results could be somewhat different when different stock selection strategies were used. Therefore relying on the results of individual strategies for investment might take higher risks.

Table 7 Partial sell orders for machine learning strategies

Time	Strategy	Stock Code	Stock Name	Buy/Sell
1/12/2021	Random Forest	002795.SZA	Yonghe Smart Control	Sell out
		002348.SZA	Clorox	Sell out
		603617.SHA	Junhe	Sell out
	XG-Boost	002095.SZA	Business treasure	Sell out
		002795.SZA	Yonghe Smart Control	Sell out
		002858.SZA	Lixson Racing	Sell out
		002269.SZA	Meibang Apparel	Sell out
		601798.SHA	Lanke High-Tech	Sell out

In order to diversify the investment risk, the equal-weight portfolio and the 60/40 portfolio were used for the asset allocation of the two machine learning stock picking strategies respectively, using the respective advantages of XGBoost and random forest stock picking strategies for the strategy combination. The combined strategies were back-tested to test the return curves of the multi-strategy portfolio, and the results were shown in Figures 6 and 7.

Based on the results of the multi-strategy portfolio backtest (Figures 6 and 7), we could see that both the portfolio strategy and the single stock picking strategy could consistently outperform the broad market, and the weighted return curves of both multi-strategy portfolios were much more stable than the return curves of individual strategies.

In order to tested the applicability and reliability of the multi-strategy combination approach, the backtesting indicators of the two combination strategies were analyzed in comparison with those of the single strategy (Table 8).

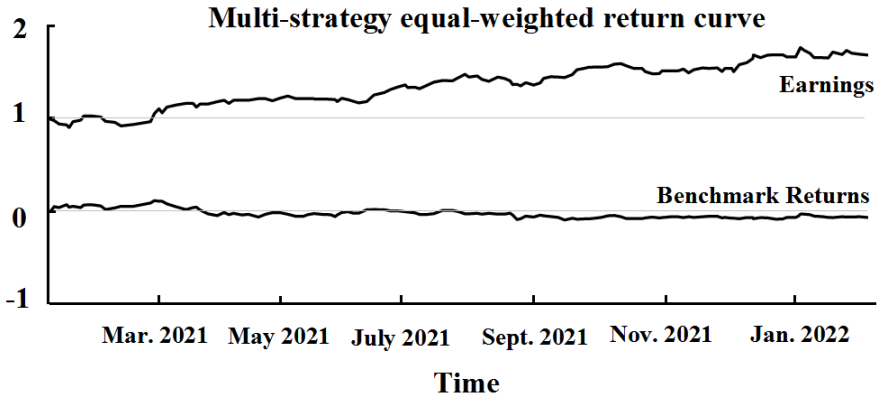

Figure 6 Multi-strategy equal-weighted return curve

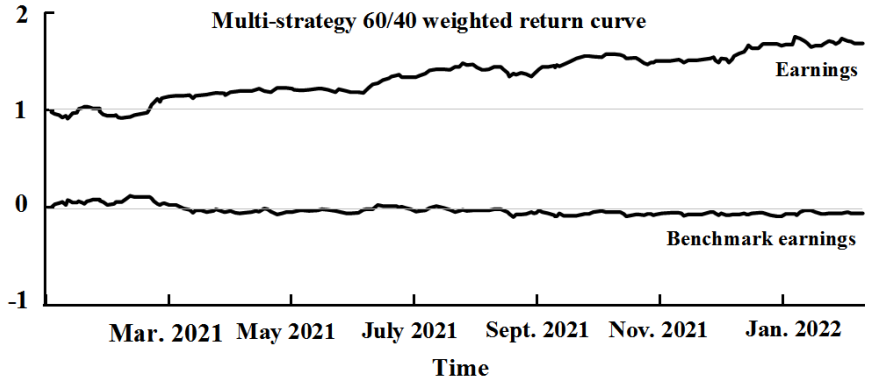

Figure 7 Multi-strategy 60/40 weighted return graph

Table 8 Comparison of backtesting effect of single stock selection and two combination strategies

Parameters	Random Forest Strategy	XGBoost Strategy	Equal-weighted weighted combination	60/40 weighted combination
Yield	69.61%	65.69%	68.12%	68.50%
Annualized Yield	72.96%	68.82%	71.39%	71.78%
Alpha	0.75	0.71	0.78	0.78
Beta	0.2	0.17	0.18	0.19
Sharpe Ratio	2.29	2.12	2.30	2.31
Earnings Volatility	23.92%	24.8%	23.09%	23.05%
Maximum Retracement	12.18%	11.52%	11.50%	11.51%

The alpha after equal-weighted and 60/40-weighted combination was 0.78 in the backtesting time period, while the alphas of Random Forest and XGBoost were 0.75 and 0.71, respectively. indicating that the multi-strategy combination was more capable of capturing excess returns. In addition, the Sharpe ratio of the 60/40 portfolio strategy was 0.02 higher than that of the random forest stock picking strategy and 0.19 higher than that of the XGBoost stock picking strategy, while the return volatility and maximum retracement of both portfolio strategies were the lowest compared to the two single strategies. The analysis of several indicators showed that both strategy combination methods could reduce the risk borne by a single strategy and obtain more robust excess returns, which was a more systematic and comprehensive quantitative investment method and worthy of in-depth study.

4 Conclusion

The BigQuant quantitative investment platform was used to backtest the historical data of all A-share stocks for the period from 01/01/2021 to 12/31/2021, and the results showed that the returns obtained by the random forest, XGBoost, and the combination strategy of both methods were significantly higher than the benchmark returns. The XGBoost strategy was more risk-averse, but its return was slightly lower than that of the random forest strategy. The strategy combination method combined two machine learning stock selection strategies with equal weight and 60/40 weight, which could effectively utilize the advantages of each single strategy and obtain an optimization method that integrates the return and risk of stock investment. Future research can consider the potential of the combination of multiple strategies for quantitative investment with a view to obtaining more robust investment returns.

References

[1] Ding P. Quantitative investment-strategies and techniques[M]. Beijing: Electronic Industry Press, 2014.

[2] Wu W X. Application of random forest in quantitative stock selection by technical indicators[D]. Sichuan: University of Electronic Science and Technology, 2018.

[3] Che Y. Research on multi-factor stock selection strategy based on machine learning methods[D]. Tianjin: Tianjin University, 2017.

[4] Jin Y H, Fang Yong, Lu Yan. Research on factor extraction method and quantification strategy based on K-NN nearest neighbor algorithm[J]. The Practice and Understanding of Mathematics, 2021, 51(19).

[5] Zhang M J, Rao Huacheng, Nan Jiangxia et al. Quantitative trading timing strategy based on decision trees [J]. Systems Engineering, 2022, 40(2).

[6] Fang Y, Chen Y Z, Wei J. Artificial intelligence and Chinese stock market - A quantitative study on portfolio based on machine learning prediction[J]. Industrial Technology Economics, 2022, 41(8).

[7] Zhang Y. The evolution of asset allocation theory[J]. Finance Expo, 2017 (19).

[8] Zhang H Q, Wu W W, Wang X F. The Revenue of the Constant Mix Strategy in Different Market[J]. Chinese Journal of Management ence. 2014, (1).

[9] Su Y M. Research on financial early warning of listed companies based on random forest and XGBoost [D]. Harbin: Harbin Institute of Technology, 2019 .

[10] Huang Y Q. Research on stock valuation based on fundamental factors[D]. Beijing: Beijing University of Technology, 2018.

[11] Hou X H, Wang B. Quantitative Investment Based on Fundamental Analysis:A Review of Research and Prospects[J]. Journal of Northeast Normal University (Philosophy and Social Science Edition), 2021(1).

[12] Ayala J, García-torres M, Noguera J, et al. Technical analysis strategy optimization using a machine learning approach in stock market indices[J]. Knowledge-Based ystems, 2021, 225(6).

[13] Alsubaie Y, Hindi K E, Alsalman H. Cost-sensitive prediction of stock price direction: selection of technical indicators[J]. IEEE Access. 2019, 7(7).

[14] Hu C H, Mei G P. Analysis of the impact of institutional investors' sentiment on stock price synchronization in China[J]. Business Times, 2014 (19).

[15] Sun A, Lachanski M, Fabozzi F J. Trade the tweet: social media text mining and sparse matrix factorization for stock market prediction[J]. International Review of Financial Analysis, 2016, 10(48).

[16] Dong X B, Chang Y Q. Multi-factor quantitative stock selection model and performance analysis based on factor IC[J]. Journal of Changchun University of Technology (Social Science Edition), 2019, 32(6).