Forecasting China's Military Industry Index: Based on Decision Tree, Random Forest and Time Series Models

Xiaoyan Cheng[1], Ziyan Liu[2(✉)], Zhijie Zhang[3], and Zhiyue Zhu[4]

[1] School of Information, Xi'an University of Finance and Economics, Xi'an 710100, China
[2] School of Innis, University of Toronto, Toronto M5S 2E8, Canada
zyan.liu@mail.utoronto.ca
[3] School of Materials, University of Manchester, Manchester M13 9PL, UK
[4] School of Foreign Language, Shandong University of Finance and Economics, Jinan 250002, China

Abstract. IncrEasing uncertainty about geopolitical conflicts and downward economic pressure have contributed to increased stock price volatility in the military industry sector as a result of the ongoing Russia-Ukraine conflict which has gradually developed into a protracted tug-of-war and a war of attrition, as well as the previous financial crises. To strengthen the role of investment profitability, this paper intends to conduct more research on the index of the military industry sector. To predict the trend of sector index, a decision tree, random forest model, time series-based ARIMA model, and neural network model are used. The sector indices are forecasted using the ARIMA model and the neural network model after the correlation test is completed with the random forest model. It is predicted that the sector index will continue to rise with possible fluctuations in the future. By using the random forest, ARIMA model, and neural network model, investors are able to avoid military industry sector risks and gain stable benefits.

Keywords: Portfolio Selection · Random Forest · Time Series

1 Introduction

Since the middle of 2020, the military industry sector in China has performed well in the whole market [1], especially the aerospace military industry sector is the strongest, mainly because the military industry is affected by short-term economic fluctuations, the annual task of military protection is basically unchanged, and the epidemic has not brought a big impact and change to the demand and fundamentals of the whole industry, then it also creates the high performance certainty of the whole sector, and the stage from the beginning to the middle of last year The valuation of the whole plate is low, which attracts all kinds of funds to allocate the military industry plate, and the public active fund has a historic overweight to the aerospace military industry plate. And the 100-day chart of the military industry index is shown below Fig. 1.

X. Cheng, Z. Liu, Z. Zhang and Z. Zhu—These authors contributed equally.

V. Gaikar et al. (Eds.): FMET 2022, AEBMR 227, pp. 357–369, 2023.
https://doi.org/10.2991/978-94-6463-054-1_40

Fig. 1. Military Industry Index.

Investing in quality stocks is an essential component of stock investing, as well as essential for portfolio investment. High-quality stocks are those with strong growth, good investment returns, and good risk tolerance [2]. All investors and institutions desire to invest in quality stocks. However, the stock market is affected by a vast array of factors and is characterized by "irregular" random movements, which makes selecting quality stocks nearly impossible.

According to technical definitions, stock selection involves a multi-factor analysis that affects stock prices, in which each factor is a dimensional indicator and the price of a stock is determined by multidimensional systems [3]. Stock selection contributes statistical analysis of a large amount of information, which is a multidimensional classification problem. An important part of selecting quality stocks from a large number of stocks is analyzing a large amount of information, which is essentially a classification problem. The classification problem involves two aspects: the selection of dimensions affecting stock prices, i.e., constructing the indicator system, and the selection of the classification algorithm, i.e., the determining of the selection model classification algorithm [4]. To solve the practical problem of selecting high-quality stocks for investment, this paper uses the random forest algorithm to construct a quantitative stock selection model to predict the index of a single sector.

Developed from machine learning, decision tree algorithms are approximation methods for classification functions [5]. Based on a top-down recursive approach, the algorithm is greedy. In 1966, Hunt et al. proposed Conceptual Learning System (CLS), and subsequent decision tree algorithms evolved from or improved upon CLS. A variety of sectors, from health to finance and technology, use these algorithms for data analysis and machine learning.

The random forest method is a combination of decision trees and bagging [6], which is a relatively new method for integrating nonlinear tree-based learning models (nonlinear tree-based models). By repeatedly dichotomizing data, Breiman et al. invented classification trees in the 1980s [7]. To generate many classification trees, Breiman combined classification trees into a random forest, i.e., randomly using variables (columns) and data (rows), and then aggregated the classification tree results. As a result of random forest, prediction accuracy increases without significantly increasing the amount of computing power. An optimal predictive algorithm for up to several thousand explanatory variables is considered to be a random forest [8], and its results are more robust to missing data and non-equilibrium data.

In quantitative forecasting, time series refers to a series of data recorded chronologically for phenomena. Scholars have paid increasing attention to time series research in the 21st century due to rapid economic development. An important contribution of Engle and Granger, winners of the 2003 Nobel Prize in Economics, was the establishment of a key concept for describing time-varying volatility, ARCH, and a series of volatility models [9]. Aboom in scholarship on economic time series has been sparked by their contributions to economic time series analysis.

This paper analyzes the military industry sector in China's domestic stock market in the context of the Russian-Ukrainian war based on various data sources. Following a correlation test with a random forest model, the time series-based ARIMA model and neural network model are used to predict the trend of the sector index and finally presented in the form of a chart. Based on this methodology, it is concluded that the military industry sector has benefited from a variety of factors following the Russian-Ukrainian conflict, which led to a rapid rise and eliminated steady progress after a significant shock and achieved certain results. Investors can benefit greatly from the proposed and applied strategies by diversifying and avoiding most of the unsystematic risks and achieving their expected returns.

2 Data and Methodology

2.1 Data

In total, the data in the study includes the closing prices ofthe military industry sector in China from January 6, 2022 to June 8, 2022. The data were obtained from NetEase Finance, and the processing of the data included the calculation of its mean, standard deviation, variance, kurtosis, and skewness. The corresponding calculation results are shown in the following Table 1.

Table 1. Data analysis result

Name	Results
mean	1598.5453
standard deviation	156.4233
kurtosis	−0.0332
skewness	−0.1152
Maximum value	1963.1031
Minimum Value	1184.7762
Median value	1604.0188

2.2 Methodology

2.2.1 Decision Trees

The decision tree algorithm is a method that approaches the discrete function value. It is a symbolic mode of classification. It first processes data, creates readable rules and decisions using inductive algorithms, and then analyzes decision using new data. Essentially, decision making is the process of digesting information through a series of rules [10].

In 1966, E.B. Hunt et al. A decision tree algorithm is proposed. In the late 1970s, J Ross Quinlan proposed the ID3 algorithm for decision trees. Since this time, decision trees have become the mainstream method of computer learning. However, this method ignores the study of the number of leaves, so the C4.5 algorithm is extended on the basis of the ID3 algorithm. The information gain rate of the C4.5 algorithm is as follows [11]:

$$GainRatio(A) = \frac{Gain(A)}{Split(A)} \tag{1}$$

$$\text{Split(A)} = -\sum_{j=1}^{c} \frac{|P_{ij}|}{|E_i|} \times \log_2\left(\frac{|P_{ij}|}{|E_i|}\right)) \tag{2}$$

GainRatio (A) is the previous information gain, and Split (A) represents the split information. In fact, split (A) is to treat each sample as a direct result of equal possibilities. E is the sample set, assuming that E has a training set of class C samples, and the number of samples of each class is P_i [12].

2.2.2 Random Forest

Random forest is relatively new to linear tree-based models. Breiman et al. invented classification trees in the 20th century to repeatedly classify and summarize data in order to reduce the amount of computation. In 2001, Breiman merged random trees into forests, that is, variables (columns) and data (ranked) to form many of the variation trees' classifications, and then summarized the consequence of classification trees.

2.2.3 Time Series ARMA Model

In the early 1970s, American statistician box and British statistician Jenkins jointly proposed a time series analysis method. The ARMA model is the most commonly used in many economic studies now, so as to achieve the purpose of fitting a stationary time series. A stationary time series means that the characteristic of the sequence itself do not change over time. For a set of time series data, the ADF test should be performed first to understand the stationarity of the data. If it is non-stationary, logarithmic or differential processing is performed to make it stationary. Then need to draw the ACF and PACF images of the stationary time series, and determine which alternative models are. Next, the alternative model is estimated and tested, and it is assessed whether the model satisfies the simplicity in terms of SC, AIC and fitting of the alternative model, i.e. H. Minimum information criterion SC, AIC and maximum adjustment. Finally, the optimal ARMA model for modeling and prediction is selected.

An ARMA model consists of an autoregressive model AR(P) model and a moving average model MA(q), so it is called a mixed model. A time series is autoregressive if it can be represented by a linear function consisting of the current value and the current random error, and the formula is:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + u_t \tag{3}$$

If each of the observations in the time series is the average weighted average of the uncertainties in the current period and the previous period, then the average process is a moving process and can be represented by the MA(q) model, and the formula is:

$$y_t = \alpha_0 + u_t + \alpha_1 u_{t-1} + \alpha_2 u_{t-2} + \cdots + \alpha_p y_{t-q} \tag{4}$$

When $p = 0$, ARMA (0, q) is ma (q) model; When $q = 0$, ARMA (P, 0) is (P) model. Since Ma (q) is stable, ARMA (P, q) depends on whether AR (P) is stable. If is a linear function composed of its current value, with the pre-random perturbation and the preceding value, then the time series is the auto-regressive moving average of the series, which can be expressed with the models ARMA (P, q), and the formula is [14]:

$$y_t = c + \beta_1 \gamma_{t-1} + \beta_2 \gamma_{t-2} + \ldots + \beta_p \gamma_{t-p} + u_t + \theta_1 u_{t-1} + \ldots + \theta_2 u_{t-2} + \ldots + \theta_q u_{t-q} \tag{5}$$

Where u_t is a white noise sequence with mean 0 and variance

3 Empirical Result

3.1 The Correlation Detection of Military Sector Index

Before predicting the index, we need to conduct correlation tests for multiple variables to prove that the rise of the stock index is affected by these variables. If the influence exists, it can show that there is a certain law of index change. Here we use the random forest model based on MATLAB to fit the data, and the results are shown below .

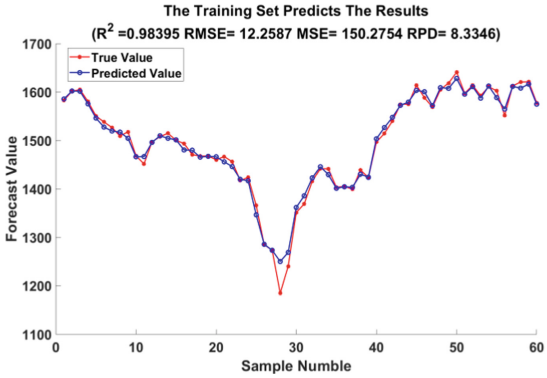

Fig. 2. The predicted result of training set.

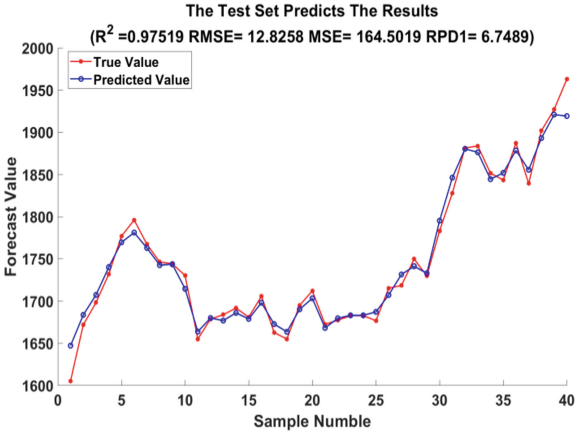

Fig. 3. The result of the test set.

From the Figs. 2 and 3, we can find that the fitting result of the sector index under the random forest model has reached the expected value of the model. The variables used to predict the index here include the trading volume, the range of changes, the total trading amount and the opening price. 4 variables. By observing the fitting results, it is found that the value of R^2 is greater than 0.04, which proves the correlation between the independent variable and the dependent variable, and establishes the necessity of establishing a model.

Because this paper focuses on the change of index with date in the research section, we try to fit the index directly using a date as a variable. However, both the training set and the prediction set are lower than the requirements for building the model (R^2 is lower than 0.4). The reason for this result is that random forests are usually used for multivariate fitting, so we looked for a time series (ARMA) model and a BP neural network model as a forecasting solution.

3.2 Time-Series Model Prediction

3.2.1 ADF Test of ARMA Model

According to the original sequence time series diagram, it can be judged that the time series is a non-stationary time series. Then we use the ADF test to prove our prediction (Fig. 4 and Table 2).

According to the original sequence time series diagram, there are intercept terms and trend terms, so an ADF test with intercept terms and trend terms is performed. The P value is greater than 0.05, and the null hypothesis is accepted, so it can be judged that the original sequence is a non-stationary time series and we can't apply the time-series model on it.

After knowing the sequence is a non-stationary time series, we use the first-order difference to make it stationary. From Fig. 5, it seems stationary.

After checking the results of the unit root test with intercept (1), without intercept and trend term(2), with intercept and trend term(3), all three-unit root tests reject the null hypothesis, the intercept term and trend term of this time series are both not significant,

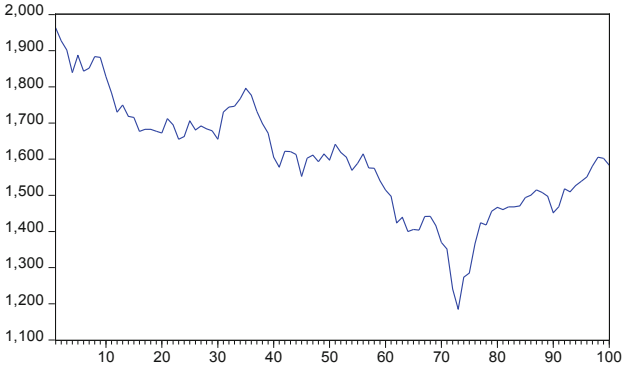

Fig. 4. Time diagram of the original sequence.

Table 2. The result of ADF test

Null Hypothesis: CLOSE has a unit root			
Exogenous: Constant, Linear Trend			
Lag Length:0			
		t-statistic	Prob.
Augmented Dicky-Fuller test statistic		−1.5287	0.8132
Test critical value	1% level	−4.0534	
	5% level	−3.4558	
	10% level	−3.1537	

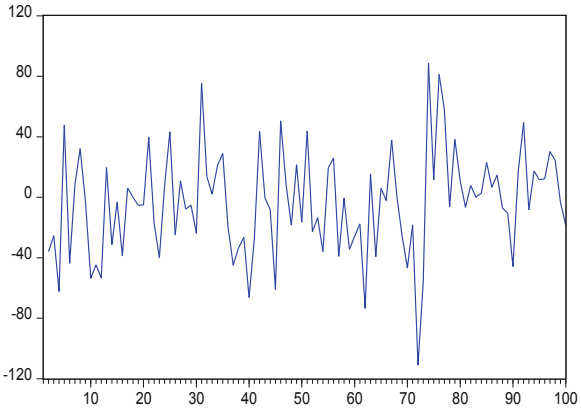

Fig. 5. Time diagram of the first-order difference sequence.

Table 3. The results of the unit root test

Variable	Coefficient	Std. Error	T-Statistic	Prob.
(1) D(CLOSE(-1))	−0.9005	0.1012	−8.8985	0.0000
(2) D(CLOSE(-1))	0.8909	0.1005	−8.8626	0.0000
(3) D(CLOSE(-1))	−0.9295	0.1024	−9.0746	0.0000

did not pass the t-statistic test. So the first difference series is a stationary time series without intercept and trend terms here is a basis for building the ARIMA model (Table 3).

3.2.2 Choose the Effective ARIMA Model

Both the sample autocorrelation map and the partial autocorrelation map of the time series after the first-order difference have no truncated features, and the model cannot be identified by the classical method of smearing and truncating features. According to the modeling experience of stock data, establish ARIMA (1, 1, 1), ARIMA (1, 1, 2), ARIMA (2, 1, 1), ARIMA (2, 1, 2), according to AIC, SC, The HQC information criterion and T-statistic identify the optimal model (Table 4).

From top to bottom are the parameter estimation results of ARIMA(1, 1, 1), ARIMA(1, 1, 2), ARIMA(2, 1, 1) and ARIMA(2, 1, 2). Compared with AIC, which is commonly used in financial analysis, HQC is different from AIC. Although HQC is also used as a parameter to measure the accuracy of the model, it is not asymptotically valid. Only when the value of the dependent variable is the same for all the estimated values to be compared, HQC can be used to compare the estimated models. Therefore, we mainly refer to AIC as the measurement standard here. According to AIC minimum

Table 4. The comparison of ARIMA models

Model	AIC	R-Squared	HQC
ARIMA (1.1.1)	9.9674	0.0046	9.9992
ARIMA (1.1.2)	9.9841	0.0082	10.0265
ARIMA (2.1.1)	9.9851	0.0072	10.0275

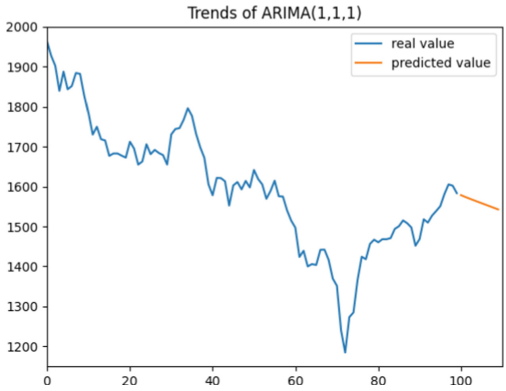

Fig. 6. The prediction of ARIMA (1, 1, 1).

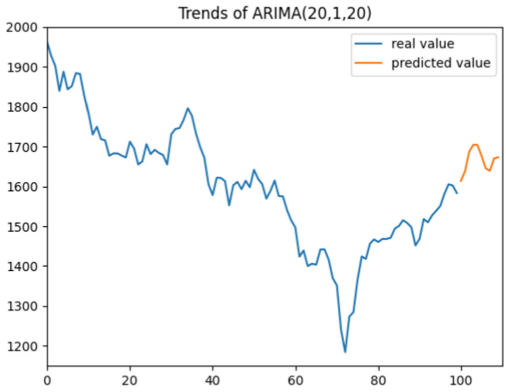

Fig. 7. The prediction of ARIMA (20, 1, 20).

information criterion, ARIMA (1, 1, 1) is the best fitting model. Considering the small number of samples, in order to see the trend as clearly as possible, ARIMA (20, 1, 20) is hereby used as a reference model for prediction.

3.2.3 The Prediction Result

It can be seen from the above figure that if the ARIMA (1, 1, 1) model is used for forecasting, it can be seen that the sector index will drop slightly in the very short term. If ARIMA (20, 1, 20) is used as a reference, it can be seen In the near future, the index of the outbound sector will continue its upward trend with fluctuation (Figs. 6 and 7).

3.2.4 Comparing Two Models and the Reason Choose the ARMA Model

From the above, although MAE is conceptually different from standard deviation (SD) for decision tree and random forest models, they still have certain comparative ability. It can be seen that the standard deviation of ARIMA model is smaller. In fact, the processing and prediction ability of decision tree and random forest model for multidimensional data is significantly higher than that of one-dimensional data in the process of application. Therefore, ARIMA model has the better method in this sample.

3.3 BP Algorithm Prediction

3.3.1 Error Analysis

The increasing trend of valuations is in line with the transaction itself. The non-linear neural network has a powerful adaptability, any non-linear non-complicated relationship can be described, and there are simple learning rules that are convenient for implementing a computer (Fig. 8).

From the figure, we can find that the error of the algorithm with 40 days is in the interval of 0 to 0.14, which proves the reliability of the algorithm.

3.3.2 Prediction Curve

From the graph, it can be concluded that over a 20-day period, the overall sector index has increased by about 17% (Fig. 9).

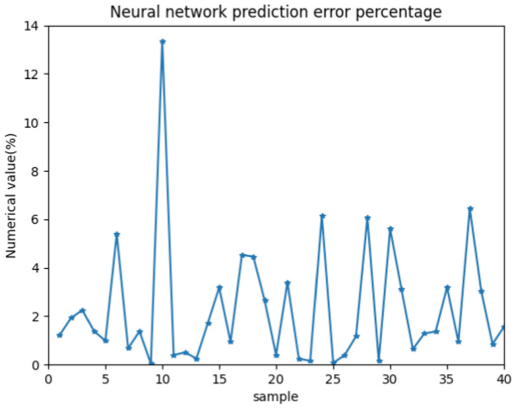

Fig. 8. Error Analysis.

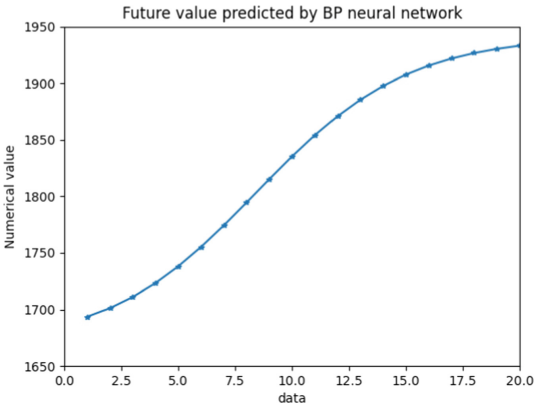

Fig. 9. BP predicted trend.

3.3.3 Shortage of BP

The most serious question is that they can not explain the process and the basis for their reasoning. The user cannot get the required questions and when there is not enough data, the network cannot work either.

4 Conclusion

This article uses the random forest model, time series model, and neural network model to predict the military sector index of China's domestic stock market in the context of the Russian-Ukrainian war. After completing the correlation test through the random forest model, the time series-based ARIMA model and neural network model are used to predict the trend of the plate index and finally present it in the form of a chart. By observing the characteristics of the obtained images, which can conclude that the entire plate is likely to continue the upward trend with possible fluctuations. At the same time, the growth rate is a respectable 5% to 20%. The research value of this article is to help readers clearly and intuitively observe the impact of the war on the military sector of the Chinese stock market since the beginning of the war and to give certain reference opinions to readers who are willing to invest to achieve the goal of stock profit.

Of course, there are also shortcomings in research. Although the time series model has been one of the most basic forecasting tools since its birth, it is undeniable that it has lagged behind the times to a certain extent. At the same time, time, the small number of samples makes it impossible to play the time series. The advantage of the model, this defect also appears in the prediction of the neural network model. Of course, as the war continues, more data will expand the sample, which can alleviate this problem to a certain extent and bring readers more accurate predictions results mainly reflected in the higher degree of fitting between the forecast curve and the actual curve, so the credibility of future exponential changes increases.

References

1. Y. Xu, Z. Liu, C. Su & Stefea Petru, Military industry bubbles: are they crowding out utility investments, Economic Research-Ekonomska Istraživanja, 2022, 35(1), pp. 692–708. DOI: https://doi.org/10.1080/1331677x.2021.1931913
2. O. A. Lamont. Investment plans and stock returns, The Journal of Finance, 2000, 55(6), pp. 2719–2745. DOI: https://doi.org/10.1111/0022-1082.00304
3. S. Patalay, M.R. Bandlamudi, Decision support system for stock portfolio selection using artificial intelligence and machine learning. Ingénierie des Systèmes d'Information, 2021, 26(1), pp. 87–93. DOI: https://doi.org/10.18280/isi.260109
4. M.G. Smith, L. Bull, Genetic programming with a genetic algorithm for feature construction and selection, Genet Programming and Evolvable Machines, 2005, 6(3), pp. 265-281. DOI: https://doi.org/10.1007/s10710-005-2988-7
5. J.R. Quinlan, Induction of decision trees, Mach Learn 1, 1986, pp. 81-106. DOI: https://doi.org/10.1007/BF00116251
6. D. Darwish, Data mining: Concepts, models, methods, and algorithms, International Journal of Computer Science, 2013, 10(4), pp. 103-111.
7. R.A. Olshen, L. Breiman, J. Friedman, & C.J. Stone, Classification and regression trees (1st ed.), Chapman and Hall, 1984. DOI: https://doi.org/10.1201/9781315139470
8. M. Belgiu, & L. Drăguţ, Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114, pp. 24-31. DOI: https://doi.org/10.1016/j.isprsjprs.2016.01.011
9. R. Engle, Co-Integration and Error-Correction: Representation, Estimation, and Testing. In C. Granger (Author) & E. Ghysels, N. Swanson, & M. Watson (Eds.), Essays in Econometrics: Collected Papers of Clive W. J. Granger (Econometric Society Monographs, 2001, pp.145–172. DOI:https://doi.org/10.1017/CBO9780511753978.009
10. J. Quinlan, Programs for machine learning, Morgan Kauffman, 1993, pp. 23–30.
11. Y. Liu, N. Xie, Improved ID3 algorithm. 2010 3rd International Conference on Computer Science and Information Technology, 2010, pp. 465–468. DOI: DOI: https://doi.org/10.1109/ICCSIT.2010.5564765
12. S. Mohapatra, Data mining principles and algorithms. Beijing: Tsinghua University Press, 2005, pp. 117-121.
13. Z. Xiong, Research on MB exchange rate forecasting model based on combining ARMA with neural networks, The Journal of Quantitative & Technical Economics, 2011, 28(6), pp. 64-76. (in Chinese)
14. W.K. Ding, An empirical analysis of stock price prediction based on ARMA model, Rediantoushi, 2019, p. 151. (in Chinese)