Stock price analysis and forecasting based on machine learning

Wenqing Yao[a], Yuting Gu[a], Sanfei Chang[a], Jing Li[b*], Qingbo Zhao[a], Fangli Ge[a]
[a] Zhengzhou Xinda Institute of Advanced Technology, Zhengzhou 450052, China; [b] PLA Strategic Support Force Information Engineering University, Zhengzhou 450052, China

ABSTRACT

With the rapid development of the global economy and the stock market, stock investment has become one of the most commonly used financial management methods, and more and more investors have begun to enter the stock market. At the same time, the high risk of stocks has also become a problem faced by many investors who are new to the market. Traditional stock price research methods mainly use technical indicator analysis or time series analysis, which has the problem of not being able to fit nonlinear data well, while machine learning methods can handle the randomness, chaos and nonlinearity of stock prices well. Therefore, this paper chooses machine learning algorithms as methods for stock price prediction, namely neural networks and support vector machines. After feature engineering processing of historical stock data, relevant machine learning algorithm models are established to predict stock prices, and the analysis and prediction results are presented in a visual way.

Keywords: Stock price prediction, machine learning, neural networks

1. INTRODUCTION

Stock price forecasting and stock data analysis play an important role in mining and early warning of the inherent laws of the stock market. They can potentially and deeply guide and predict the national economy, and to a certain extent have important reference value for the accuracy and strength of macro-control. For listed companies, the stock price forecast will affect the company's future development planning, production scale and capital utilization efficiency. At the same time, for investors, obtaining high returns from the stock market is their ultimate goal of participating in the stock market. Accurately predicting stock price dynamics can bring considerable benefits to investors, which indicates that stock price change forecasts are forward-looking and predictable.

Traditional stock investment analysis methods include technical analysis and fundamental analysis. Technical analysis is a collection of stock trading decision-making methods based on the study of stock market behaviors based on tables or quantitative indicators. Technical analysis was born in the early 20th century, best known as Dow Theory and Gann's Law. Technical analysts typically analyze and evaluate current market conditions based on exchange trading data and various technical indicators they develop. Technical analysis focuses on trading historical data, and mainly analyzes and forecasts stocks based on trading data, with a focus on short- and medium-term investments. Fundamental analysis means that the stock market analysis focuses on basic factors such as macroeconomic conditions, microeconomic conditions, industry conditions, company financial conditions, and operating capabilities. Fundamental investment is mainly value investment, which mainly supports and encourages the future development of a company or enterprise or its actual value return in the future, mainly medium and long-term value return[1].

At present, the methods of predicting stock prices are mainly divided into two categories: one is based on statistical methods, and the other is based on artificial intelligence and machine learning methods. Wu Wei et al. used the BP neural network method in the machine learning method to predict the change trend of the Shanghai stock market composite index in 2001. By adjusting and optimizing various parameters of the network, the prediction accuracy of the final model is as high as 70%. Their research shows the feasibility of applying neural networks to stock price prediction[2]. Li Xiang used BP neural network to predict stock price in 2008. The results show that BP neural network is quite suitable in China's stock market[3]. Peter Zhang made a comparative analysis of the application of ARIMA model and neural network model in time series forecasting in 2003. The results show that the neural network effect is better than the Arim model in predicting results[4]. In 2011, Ouyang Jinliang et al. improved the BP neural network, conducted an empirical study on commodity export forecast, created a method combining the incremental pulse method and the dynamic learning speed correction

[*] 351800040@qq.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250660
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2662176

method, and proposed an improved neural network BP algorithm. It effectively avoids slow convergence and is susceptible to local minimization, thereby improving the optimization of the algorithm[5]. Xiao Yufeng et al. based on the support vector machine on the stock historical data empirical and research, and used the corresponding Gaussian radial kernel function to study the model, and finally achieved a good prediction effect[6]. When Cai Hong and Chen Rongrong studied stock price forecasting in 2011, they combined principal component analysis with neural networks, first using the principal component analysis method to select the features of the initial attributes, and then training the model. It selects the capital stock as the sample stock, and uses six variables such as the daily highest price and transaction volume of the stock in a certain period as the input variables of the model. After the selection of principal component analysis, the input variables are reduced to two, and the first variable The highest price for two days is used as a predictor for the model. Through the analysis of the prediction results, it is found that the running time of the PCA-BPNN model is smaller than that of the pure BPNN model, and the prediction accuracy is higher[7]. In 2016, Zhao Chen et al. used the previous day's closing price, lowest price, stock trading volume, trading volume, highest price, opening price, and stock fluctuation range as input variables and the opening price with the next day's price as the predictor variable, using MATLAB The software builds BP neural network model. The conclusion is that the model is effective in short-term stock price forecasting, but long-term trend forecasting needs to be improved[8]. In 2018, Shang Weiping and Dai Yu used the smooth ARI-Ma-LS-SVM combination model to predict the stock prices of four different industries, which can effectively predict the short-term price trends of these stocks[9]. Han Shanjie and Tan Shizhe used Apple's 240-day average daily opening price in 2018 as the input variable, used a neural network model to predict the company's 240-day average daily stock return, and then used the root mean square error to estimate the prediction performance of the neural network model. The results show that the model has a good prediction effect, but can continue to improve[10]. In 2012, Asadi S et al. used a fusion model of data preprocessing method, LevenBerg-Marquardt (LM) algorithm and neural network algorithm to predict stock market prices. The results show that their method can predict the volatility of stock market prices and achieve good prediction results[11]. In 2010, Yang Xinbin and Huang Xiaojuan established a stock market forecasting system model based on support vector machine by using support vector machine nonlinear extended sample to determine the order of time series model, and using forward floating feature screening method to select features, and carried out simulation experiments on stock prices. The simulation results show that the support vector machine model has higher prediction accuracy than the neural network and CAR model, which proves that the support vector machine is suitable for the prediction of nonlinear problems such as stock market prediction, and has higher accuracy and application value[12]. The research status at home and abroad shows that the application of machine learning algorithms in the stock market is feasible and advantageous.

Due to the instability and complexity of the stock market, stock data is often high-dimensional and nonlinear. In the face of high-dimensional and nonlinear data analysis and processing, machine learning methods have better analysis and processing capabilities than traditional model and analysis methods, and also reduce the cost of building complex models. At present, machine learning has become an important way to predict the stock market. Based on this, two algorithms of neural network and support vector machine are used in this paper. In addition, past stocks will be analyzed and visualized, so that the forecast results can provide reference value for investors and financial researchers.

2. METHOD

2.1 Overall research framework

The goal of this paper is to realize the analysis and prediction of stock prices. In order to predicting stock price, we combined the quantitative and qualitative analysis approach, used the historical data of the stock market, and artificial neural networks and support vector machines to build a prediction model. The overall research framework of this paper is shown in Figure 1. Firstly, data needs to be acquired and stored. The research adopts the method of crawling open source data from the Internet to collect historical transaction data, and stores the obtained data in csv format, which is convenient for data processing, analysis and visualization; secondly, the data needs to be preprocessed, analyzed and visualized. The research deals with outliers, missing values and normalization of the acquired data, and uses line graphs and pie charts to visualize the stock price trend. Finally, a stock price prediction model is constructed. In this study, two prediction models are constructed, one is a prediction model based on Long Short-Term Memory (LSTM, Long Short-Term Memory), and the LSTM is used to capture the time dependence of historical stock price data to predict the stock price at the next moment. The other is based on support vector machine (Support Vector Machine, SVM) for prediction.

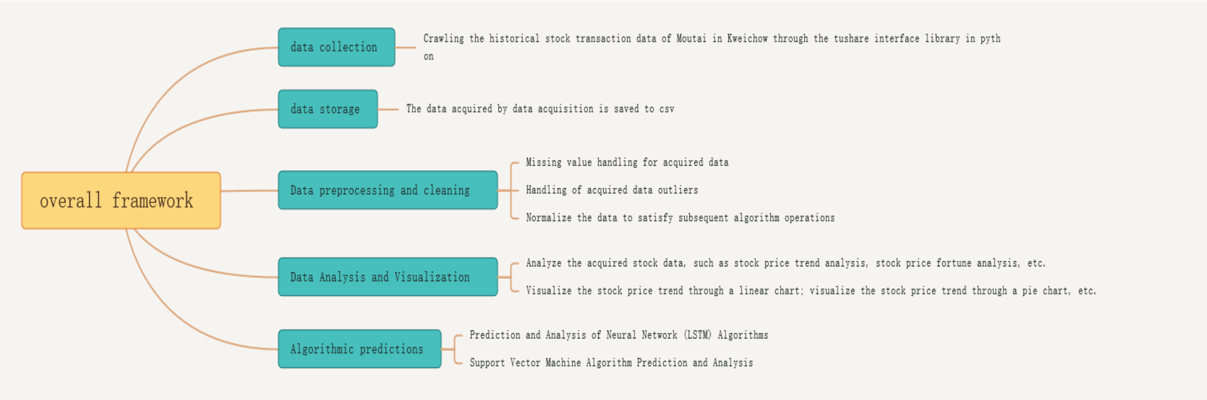
overall framework
data collection
Crawling the historical stock transaction data of Moutai in Kweichow through the tushare interface library in pyth on
data storage
The data acquired by data acquisition is saved to csv
Data preprocessing and cleaning
Missing value handling for acquired data
Handling of acquired data outliers
Normalize the data to satisfy subsequent algorithm operations
Data Analysis and Visualization
Analyze the acquired stock data, such as stock price trend analysis, stock price fortune analysis, etc.
Visualize the stock price trend through a linear chart; visualize the stock price trend through a pie chart, etc.
Algorithmic predictions
Prediction and Analysis of Neural Network (LSTM) Algorithms
Support Vector Machine Algorithm Prediction and Analysis

Figure 1. Overall research framework.

2.2 Algorithm model

2.2.1 Neural Network. This research is mainly based on the long short-term memory network to fit the historical stock data of Kweichow Moutai. LSTM belongs to a kind of neural network, and for LSTM, it solves the problem that recurrent neural networks cannot rely on for a long time. Comparing LSTM and RNN, it is found that three main gates are added: forgetting gate, input gate, and output gate. The LSTM structure is shown in Figure 2.

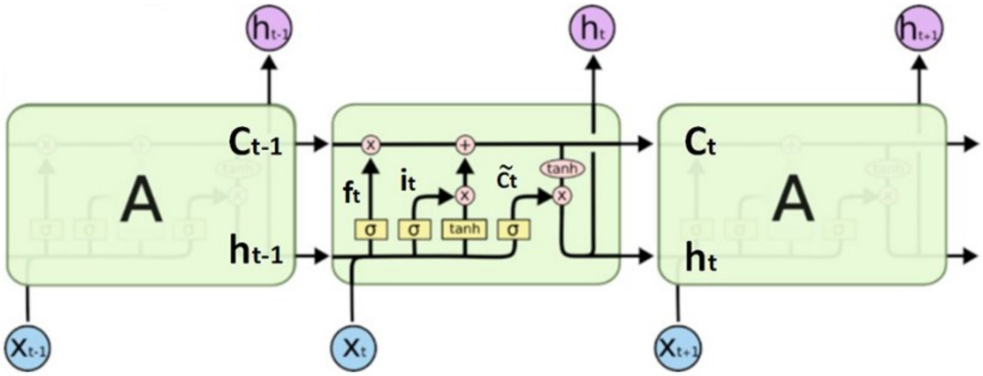

Figure 2. LSTM structure diagram.

In the above figure, X represents the scaled information, + represents the added information, б represents the sigmoid layer, tanh (hyperbolic tangent) represents the tanh layer, h_{t-1} represents the output of the previous LSTM unit, c_{t-1} represents the memory of the previous LSTM unit, X(t) represents the input, c_t represents the latest memory, and h(t) represents the output. The state of each transmission unit is what determines the core of the LSTM network, which is to pass through each horizontal line in the graph. A unit state is equivalent to a conveyor belt, which runs through the entire structure. In this process, only some linear effects are used to ensure the invariance of information transmission. LSTM also has a good performance that can add and remove information transmitted to the cell state, through several structures to manage the transmission of information and called the threshold, the threshold is to selectively allow information to pass. The calculation principle of LSTM is as follows:

$$z_t = \sum_{i=1}^{I} w_{x_i} x_i^t + \sum_{h=1}^{H} w_{h_i} x_i^{t-1} + \sum_{c=1}^{C} w_{s_i} s_c^{t-1} + b_i \tag{1}$$

$$y_t = f z_t \tag{2}$$

where w_{x_i}, w_{h_i} respectively represent the weight distribution of different cellular mechanisms w_{s_i}. w_{x_i} represents the external information variable related to the input gate; w_{h_i} is the input part of the representation $cell$; the w_{s_i} representations $t-1$ moment generally refers to the state. Since the LSTM mode unit is associated with the information sharing of its hidden layer nodes, it can be regarded as a part of the external input; b is a bias vector; f is the activation function $sigmoid$. The mechanism and related parameters of the forgetting and output gates are basically the same as the input, and the final hidden layer unit state value is the input prediction value obtained from the activation function tanh.

2.2.2 Support Vector Machines. SVM is a kind of generalized linear classifier (generalized linear classifier) that performs binary classification on data according to the supervised learning method, and its decision boundary is the maximum margin hyperplane that solves the learning sample. In recent years, it has made breakthroughs in theoretical research and algorithm implementation, and has become a beneficial means to overcome traditional problems such as multi-dimensional index application disasters and experience summarization problems. Although still in the development stage, it has formed the basic framework of its theoretical basis and implementation methods. Support vector machines are the most common methods for solving pattern recognition, discriminant analysis, and regression problems.

SVM is to map the required data to the high-dimensional feature space through nonlinear mapping, and also solve the nonlinear regression data problem for the linear regression problem in this space. However, unlike SVM classification, SVM returns only one point. The optimal hyperplane is not designed to maximize the distance between two sampling points, but to minimize the distance between all sampling points and the optimal overall anomaly of the hyperplane. Compared with the classification problem, the pilots returned by the support vector machine are mainly not located on both sides of the optimal hyperplane boundary, but between the two boundary lines. The principle of SVM is shown in Figure 3.

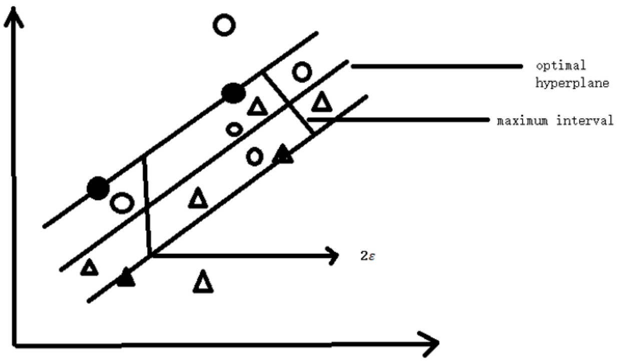

Figure 3. Schematic diagram of the support vector machine algorithm.

In this paper, the Gaussian radial kernel function is selected as the basis for establishing the model. The kernel function must be parameterized y; the Gaussian radial kernel function:

$$K(x_i * x_j) = \exp\left(-\gamma \left||x_i * x_j|\right|^2\right),\ \gamma > 0 \tag{3}$$

This paper uses the SVM method to analyze the rise and fall of the stock market. First, select historical transaction data with great reference value (for example: stock price, trading volume) to form a set of financial time series (t=1.2. L); then, in support vector machine learning, predict the time series group. Processing, at each moment of x, y=I+1 form {x} of the stock market ups and downs. If the trend for i+1 is up, then y=+1, if the trend is down, then y=-1. Finally, the input vector is trained to get the SVM model.

3. EXPERIMENT AND ANALYSIS

3.1 Experimental data source

This paper crawls historical stock education data as experimental data through web crawler technology. Due to the experimental requirements of the paper, the relevant historical stock trading information is crawled, and this data is enough for this paper to experiment with each algorithm and compare its results. Table 1 shows some of the fields required for the data.

Table 1. Examples of data fields.

Name	Type	Description
date	str	date
open	str	opening price
high	float	highest price
low	float	lowest price
close	float	Closing price
pre_close	float	previous close
change	float	ups and downs
pct_chg	float	Quote change
vol	float	volume
amount	float	Turnover

3.2 Data analysis and visualization

Data analysis and visualization is to clearly convey the meaning of the data, to help explain trends and statistics, and to show the data and the deep value contained in the data that cannot be seen in the previous data analysis text reports[13]. This paper conducts simple data analysis and visualization of the acquired historical stock transaction data, so that investors can quickly understand the rise and fall, development status, and stability of the stock through the legend.

(1) Stock price chart analysis and visualization

This paper selects the trading data of the stock since its listing as the basis to draw a trend chart, as shown in Figure 4, from the trend chart, we can analyze the stability of the stock, whether there will be bottomless when it falls, and you can see the stages of the stock's ups and downs, as well as the ups and downs trends.

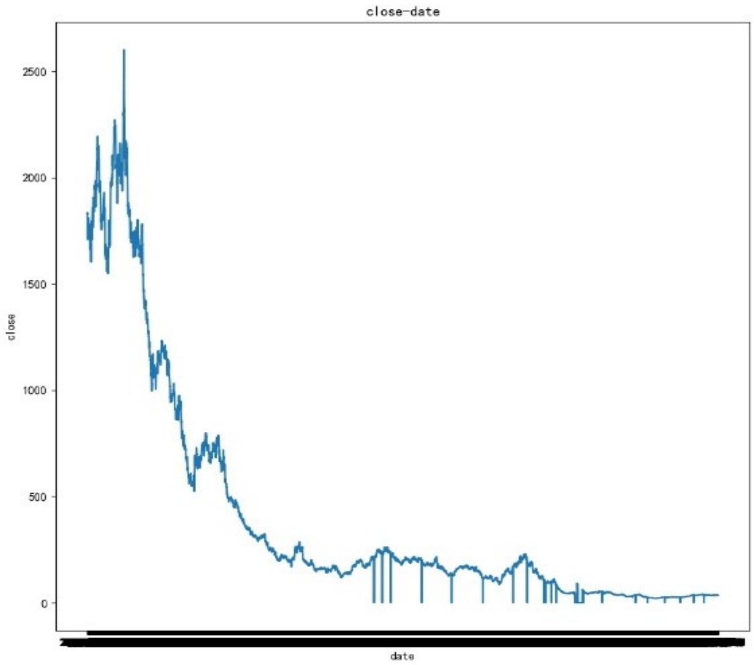

Figure 4. Stock price chart.

(2) Analysis and visualization of the highest point of stock fortune

This article analyzes the maximum daily points of the stock, which allows us to understand that there are more positions in the maximum range to judge whether to invest in this stock. In this paper, the maximum value is divided into four intervals: [0,20] (20,25] (25,30] (30, +∞) as shown in Figure 5. It can be seen from Figure 5 that the stock every day The maximum value of the distribution of the most points is above 30, the least below 20, and the difference between 20-25 and 25-30 points is small, indicating that the stock has a lot of room for daily growth.

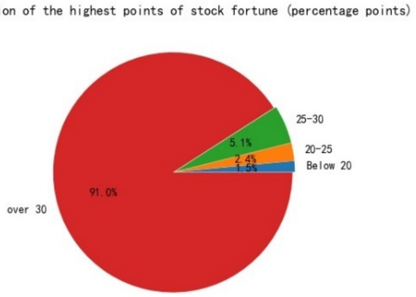

Figure 5. Stock fortune chart.

3.3 Algorithm prediction and analysis

In this paper, the LSTM is used to build a model and the neurons are set to 1024, 512, 66, 1; the number of iterations is set to 1000; the data latitude of X is set to 1, etc. The following methods can be used for the analysis of the prediction results of the neural network LSTM algorithm. The method is to draw and compare the predicted results with the real results, and the prediction effect can be known through intuitive observation. If the predicted curve completely coincides with the real curve or is quite close, the prediction effect is good; otherwise, the prediction model needs to be improved.

The comparison results of prediction evaluation are shown in Figure 6. In Figure 6, the blue solid line is the real data, the red dotted line is the predicted data, the abscissa is the date subscript, and the ordinate is the corresponding stock price. It can be seen from the figure that the prediction results of the LSTM model are generally better.

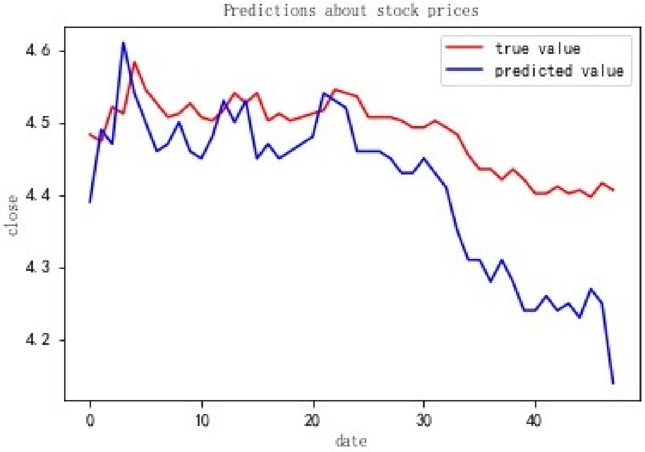

Figure 6. Prediction evaluation comparison chart.

For the prediction result of support vector machine (SVM), this paper predicts the future stock price and displays it in the way of outputting historical data and prediction data, as shown in Figure 7. The solid blue line in Figure 7 represents the historical transaction data, that is, the real data, and the solid orange line represents the predicted data. Comparing the prediction results in Figures 6 and 7, it can be seen that the prediction effect of the LSTM model is significantly higher than that of the SVM. The prediction accuracy of SVM still needs to be improved. Although the prediction accuracy of SVM is much smaller than that of LSTM, it is found that SVM has been greatly simplified in general classification, regression and other problems through the model structure and running time of the two, and the algorithm is simple and has good approximation performance. and generalization performance. It can be seen from Table 2 that the learning time of the prediction model established by SVM is much shorter than that of the neural network. This is mainly due to the use of the conjugate gradient method in the neural network, which greatly prolongs the training time, resulting in an extension of the learning time. From this, it can be seen that the convergence speed of SVM is much faster than that of neural network.

Table 2. Comparison of algorithm training time.

Algorithm	Training time
Neural Network (LSTM)	1088.94s
Support Vector Machines	0.24s

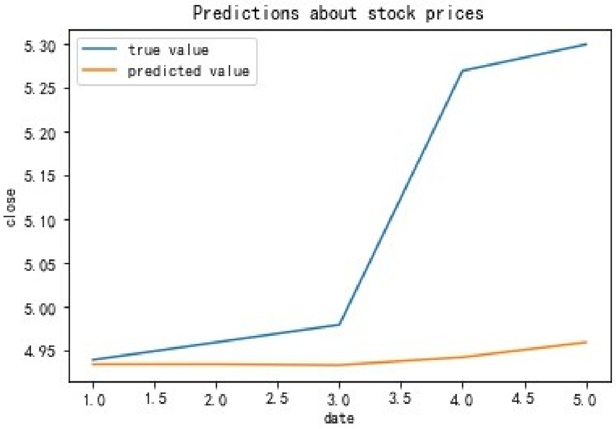

Figure 7. Prediction result graph.

4. CONCLUSION

Neural networks and support vector machines have their own distinct advantages and disadvantages. The same historical data will cause different prediction results between the two depending on the size of the data, the type of data, and the structure of the data. By comparing the forecasting results and forecasting efficiency of the two, this paper analyzes the advantages and disadvantages of the two, and proposes specific applicable scenarios and scopes to provide theoretical support for the stock price forecasting method and even provide a reference for investors. In this paper, it is found through experiments that neural networks are not superior to support vector machines in any case. The accuracy of stock price predictions is affected by multiple factors, such as the size or structure of the dataset. The common point of neural network and support vector machine is that both have certain advantages for processing nonlinear data, and have their own unique theory. For stock data, it is also nonlinear data, so it is relatively appropriate to use neural network and support vector machine as the algorithm for predicting stocks. But at the same time, they also have obvious shortcomings. For example, support vector machines have a solid theoretical foundation and have a certain degree of interpretability, while neural networks usually have problems that are difficult to explain their internal mechanisms. Similarly, neural networks have greater advantages than support vector machines in the face of large data samples. Therefore, the purpose of this design is to explore the advantages and disadvantages of both by using neural network and support vector machine as a model algorithm for predicting stocks.

In addition, at present, this paper only uses a single historical transaction data and a single model to predict the stock price, which has certain limitations. In the follow-up, the advantages of neural network and support vector machine will be combined for models, and a more accurate and more reasonable prediction model will be constructed for prediction.

REFERENCES

[1] Liang, X., “Research on the influencing factors of stock price information content in China’s securities market,” Financial Economy: The Second Half of 2018, (4), 2(2018).

[2] Chen, W., [Using BP Neural Network to Predict the Rise and Fall of Stock Market], Jilin University, (2000).

[3] Li, X., “Stock price prediction based on BP neural network,” Journal of Dalian Maritime University, (z1)3, (2008).

[4] Zhang, G. P., “Time series forecasting using a hybrid ARIMA and neural network model,” Neurocomputing, 50(1), 159-175(2003).

[5] Ouyang, J., “Application of integrative improved BP neural network algorithm in stock price forecast,” Computer & Digital Engineering, (2011).

[6] Xiao, Y., [Stock Price Prediction Method and Empirical Research Based on Support Vector Machine], Central South University, (2010).

[7] Cai, H. and Chen, R. Y., “Stock price prediction based on PCA and BP neural network,” Computer Simulation, 28(3), 365-368(2011).

[8] Sun, X. and Lei, Y., “Research on financial early warning of mining listed companies based on BP neural network model,” Resources Policy, 73(2), 102223(2021).

[9] Cheng, C. P., Chen, Q. and Jiang, Y. S., “Research on stock price prediction based on wavelet decomposition and ARIMA-SVM combined model,” Computer Simulation, (2012).

[10] Han, S., “Talking about world philosophy Design and implementation of deep learning model for stock forecasting based on tensorflow,” Computer Applications and Software, 35(6), 6(2018).

[11] Asadi, S., Hadavandi, E., Mehmanpazir, F., et al., “Hybridization of evolutionary Levenberg-Marquardt neural networks and data pre-processing for stock market prediction,” Knowledge-Based Systems, 35, 245-258(2012).

[12] Yang, X. and Huang, X., “Research on stock price prediction based on support vector machine,” Computer Simulation, (9)4, (2010).

[13] Li, H., “Big data processing technology and exploration,” Information and Computer, 32(15), 2(2020).