Prediction of Shanghai Composite Index Based on Macroeconomic Indicators and Artificial Intelligence Method

Heng Lyu[1,2], Muqing Zhu[1], Hao Lin[3], Hanzhen Huang[1], Huiying Fang[1], and Zili Chen[1](✉)

[1] Guangzhou Huali College, Guangzhou, Guangdong, China
f549640907@qq.com
[2] King Mongkut's University of Technology Thonbur, Bang Mod, Thung Khru, Bangkok, Thailand
[3] Hong Kong Founder Securities, Zhuhai, Guangdong, China

Abstract. The stock market can be defined as a market that, on the one hand, facilitates companies that need financing and, on the other hand, provides opportunities for investors who need to invest. By predicting the rise and fall of stock indices, it can bring guidance to individuals and companies when to enter the financial market, and it can also provide theoretical implications for government economic policy making. However, the stock market is a complex system full of various information, it is not only affected by past information, but also by current political, economic and psychological factors, so it is difficult to accurately predict the rise and fall of the stock index. At present, the stock index rise and fall prediction methods are mainly applied technical analysis method and measurement time series analysis method, which applied technical method is used by more groups, because it almost does not need too much analysis but according to personal investment habits and experience, subjective color. The econometric time series method is a method that is effective only when used in an ideal situation, which requires the input of the independent variable indicators and the target variable is preferably linear, if it is a non-linear situation, the results will have no reference significance. In this paper, we combine the main capital flow model with support vector machine as a tool to construct a stock index up/down prediction scheme.

Keywords: Macroeconomics · Artificial intelligence · SSE Composite Index

1 Introduction

This paper focuses on the prediction of Shanghai stock market using macroeconomic indicators and machine learning algorithms. With the continuous development of China's financial industry, the stock market has gradually become one of the focuses of investors. As one of the largest cities in China, Shanghai's stock market is representative and has certain connections with other regions of the country, so it is important to analyze it and

K. Hemachandran et al. (Eds.): ICAID 2023, AHIS 9, pp. 185–194, 2023.
https://doi.org/10.2991/978-94-6463-222-4_18

propose a corresponding forecasting model. In this context, this thesis will start from the following aspects: firstly, introduce the relevant literature on stock price forecasting at home and abroad; then explain the mechanism of macroeconomic factors on stock price fluctuation and the commonly used economic models; then detail the machine learning algorithm used, Support Vector Machine (SVM), and its optimization method; finally, the macroeconomic data and machine learning model are combined to forecast the SSE Composite Index in the future period, hoping to achieve the purpose of improving the efficiency of investment decisions and reducing risks.

2 Macro-economy and Stock Market Forecast

2.1 Macro-economy

Macro-economy refers to various economic activities of a country or region in a certain period, including gross domestic product (GDP), consumer price index (CPI) and so on. These data reflect the overall economic situation and inflation of the country or region. For investors, macroeconomic information is one of the most important reference factors. For example, when we decide to buy a certain stock or fund, we need to look at the current macroeconomic situation to judge whether it is appropriate. In addition, the government will also issue macroeconomic reports to guide public expectations and formulate corresponding policies. Therefore, it is of great importance to accurately forecast macroeconomic trends [1]. In China, macroeconomics mainly consists of annual economic operation data published by the National Bureau of Statistics. Among them, the most core indicator is the GDP growth rate. From historical experience, China's GDP growth rate shows obvious cyclical fluctuations and is generally divided into three stages: the high growth stage, the medium-low growth stage and the new normal. In addition, there are some other macroeconomic indicators such as year-on-year growth rate of industrial value added, year-on-year growth rate of total retail sales of consumer goods, etc. In addition to the macroeconomic indicators mentioned above, artificial intelligence technology is also widely used in the field of macroeconomic analysis. Take machine learning algorithm as an example, it can train a large amount of historical data, summarize the patterns and build models from them. This approach can be used not only to predict future economic trends, but also to help policy makers make more informed decisions. For example, in the area of monetary policy, AI technology can simulate the effects of different monetary policy tools and thus optimize the central bank's decisions [2].

2.2 Stock Market Forecast

1) Traditional Forecasting Methods

When forecasting stock prices, the commonly used methods include technical analysis, fundamental analysis and evolutionary game models. Among them, technical analysis is a method to determine future trends by observing stock price charts. This method relies mainly on past stock price data as a reference basis, which is presented graphically to

help investors make decisions. However, it has limitations because it cannot take into account other factors that affect the movement of stock prices. Fundamental analysis is a method of predicting the rise or fall of a stock price by studying the company's financial statements, industry environment and macroeconomic policies to find the intrinsic value pattern. This method usually takes a long time to get accurate results. Evolutionary game modeling, on the other hand, treats the stock market as a complex dynamic system and uses the returns of different strategies to build a mathematical model and eventually arrive at the optimal strategy. Although these methods have their own advantages and disadvantages, they share the common disadvantage that it is difficult to fully grasp the market changes and slow to react to unexpected events or major news [3].

2) Measurement Methodology

In this paper, the ARIMA model (shown in Fig. 1) in time series analysis is used for empirical study. This model is a classical time series forecasting method, which was proposed by Box and Jenkins in 1978. The basic idea is to use past information to infer future trends by building an appropriate mathematical model of historical data. Specifically, it is to transform the non-stationary time series into a stationary time series and then use techniques such as differencing or sliding average to extract the seasonal cycle change pattern to predict the future trend. In addition, this paper also selects some indicators reflecting the condition of the stock market as prior indicators, including stock price index (SH), consumer price index (CPI), and industrial value added growth rate (INDEX). These indicators are obtained from the official website of National Bureau of Statistics or Wind database [4]. Among them, SH represents the total market capitalization of companies listed in Shanghai Stock Exchange; CPI represents the change of consumer price level; and INDEX represents the year-on-year growth rate of GDP.

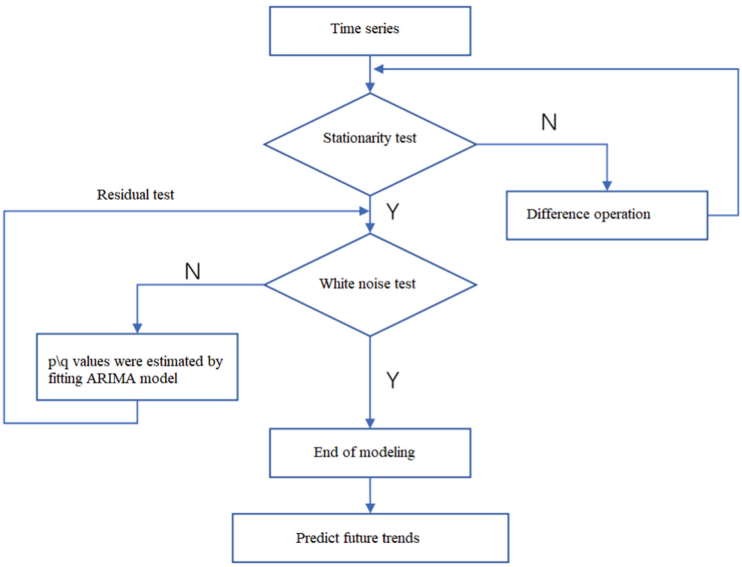

Fig. 1. ARIMA model

3) Artificial Intelligence Approach
With the continuous development of computer technology, Artificial Intelligence (AI) has been widely used in the field of finance.AI refers to a computational model that simulates the process of human intelligence activities through algorithmic programs for the purpose of achieving specific task goals. Currently, common AI methods include Support Vector Machine (SVM), Decision Tree, Neural Network, and so on. These methods have the advantages of high efficiency, accuracy and interpretability, and are widely used in stock price forecasting, futures price forecasting and exchange rate forecasting. In this paper, the SSE Composite Index is selected as the research object and the SVM method is used for forecasting analysis [5]. At the same time, because the SSE Composite Index has a certain degree of volatility, the Random Forest algorithm is introduced to optimize it. In addition, considering that the selection of data samples has a large impact on the results, a logistic regression model is used to screen out the macroeconomic indicators with a high correlation with the SSE Composite Index [6].

3 Neural Network and Support Vector Machine Research

3.1 Introduction to Neural Networks

BP neural network is known as error backpropagation neural network. It is a multilayer feedforward network trained according to the error backpropagation algorithm, whose basic feature is to minimize the loss function by continuously adjusting the weights and biases, so that the output results are closer to the desired value. In this paper we use MATLAB software for simulation experiments and optimization of the model. It is mainly used to determine whether there is a long-term stable equilibrium relationship or causality between two time series, specifically, when one variable is not a Granger cause of the other, the two variables are considered uncorrelated. Conversely, if one variable is the Granger cause of another variable, then it can be said that there is a stable relationship between the two variables over time. Granger causality tests are often used to study the association between financial data such as stock price fluctuations as well as exchange rates. The autoregressive conditional heteroskedasticity model is referred to as the ARCH model. The model was first proposed by Siegel (Shiller) and has since been extended to the more general case of including all factors that may affect stock price changes [7]. In this case, the ARCH term represents the stochastic volatility shock, which is used to describe the asymmetric nature of stock market returns. The ARCH-LM test is used to detect significant differences in the sum of squares of residuals at different lag orders using this model.

A neural network is a combination of a large number of simple processing units called neuron models, and a typical neuron model structure is shown in Fig. 2.

In the neuron shown above, xj is the input variable of the model, wij is the connection power of the model, ui is the output of the neuron after linear combination of the input variables, θi is the threshold value of the model, and the threshold value is the deviation when expressed in bi, and the output of ui after deviation adjustment is vi. f(.) is the excitation function of the model and yi is the final output of neuron model i. The

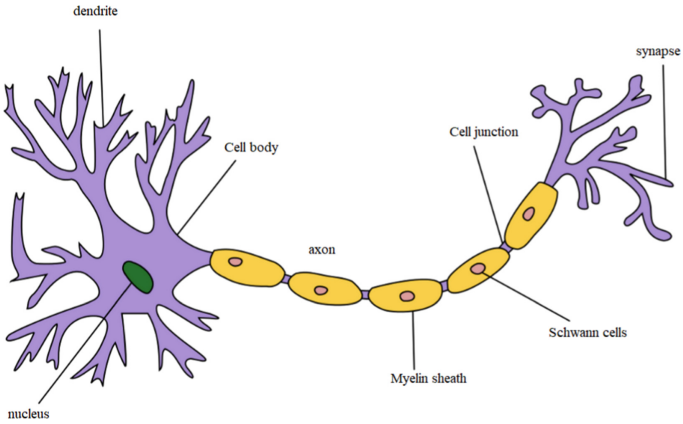

Fig. 2. Neuron model structure

mathematical expressions of the process are shown in Eqs. (1), (2) and (3):

$$u_i = \sum_j w_{ij} x_j \tag{1}$$

$$v_i = u_i + b_i \tag{2}$$

$$y_i = f\left(\sum_j w_{ij} x_j + b_i\right) \tag{3}$$

Different functions can be used as excitation functions for neuron models, but the most commonly used and most basic functions are mainly of three types, threshold functions, segmentation functions and Sigmoid functions (and S-shaped functions).

(1) Valence function

Research on artificial intelligence algorithms began in the early 1960s, when countries around the world were actively engaged in scientific research on computer simulation technology and intelligent computing and other related fields. It was only in the late 1980s that China began to undertake system engineering and theoretical research in this area [8]. It can be used for parameter estimation, fuzzy reasoning and neural network algorithms to build up a rule base and use it to obtain a predictive decision model. The valence function has the functional form shown in Eq. (4).

$$f(x) = \begin{cases} 0, & v < 0 \\ 1, & v \geq 0 \end{cases} \tag{4}$$

(2) Segmentation function

The segmentation function is a special nonlinear programming method which can solve complex problems with global and stability. In this paper, we use the optimal segmentation method of SIR approximation. In the solution process, continuous sequences as

well as discrete sequences are introduced to perform predictive analysis of the combined index: the continuous period is decomposed into a number of adjacent periods and the correlation coefficient between them is calculated by using the SIR approximation, and then a new value is generated according to the actual value between each segment compared with the given value, which is the branching fixed base function, so that a new sequence is obtained, and this value is the set of aggregated points [9]. Then the branching basis function using the SIR approximation method to solve its value and the ratio of the given value that the combination of the index. The form of the branching function is shown in Eq. (5).

$$f(x) = \begin{cases} 0, & v \geq 1 \\ v, & 1 > v > -1 \\ -1, & v \leq -1 \end{cases} \tag{5}$$

(3) Sigmoid function

In order to make the composite index reflect the research results faster and better, we simplified the established model. First, the Sigmoid function is introduced in the construction of the pooling layer. Its Sigmoid function is defined as shown in Eq. (6).

$$f(v) = \frac{1}{1 + \exp(-av)} \tag{6}$$

3.2 Learning of Neural Networks

In a neural network, we need to minimize the loss function by continuously adjusting the weight parameters. This process is called Back Propagation algorithm (BP) or Error-Backpropagation (BP) for short. Specifically, when we feed training data into a neural network, each neuron outputs a result, and the difference between these results and the actual value is the error signal. Next, we update the network using an optimization algorithm such as gradient descent in order to make the error as small as possible. Stochastic Gradient Descent (SGD) is usually used as an optimizer in the training process of neural networks. SGD is a commonly used optimization algorithm whose basic idea is to use the current gradient to calculate in which direction more steps should be moved in the next step. Since SGD has the advantages of being efficient and easy to implement, it is widely used in the training of various deep learning models. In conclusion, the training of neural networks is very important, and only after sufficient training can accurate and reliable prediction results be obtained [10]. In summary, the process of learning of neural networks is shown in Fig. 3.

3.3 Support Vector Machine

Linearly separable support vector machine (SVM) is a binary classification model (as shown in Fig. 4), whose basic idea is to map data into a high-dimensional space for linearly indistinguishable problems. In this process, the low-dimensional input is mapped to the high-dimensional output by a kernel function, thus making the originally linearly

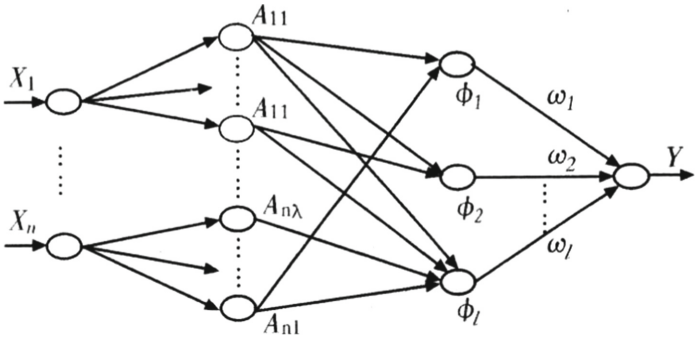

Fig. 3. The learning process of neural network

indistinguishable samples linearly separable. Common kernel functions include linear kernel, polynomial kernel, and radial basis kernel. Among them, radial basis kernel functions are widely used in machine learning because of their good generalization ability and fast convergence speed. For nonlinear separable support vector machines, the commonly used algorithms include Linear Regression Tree (LRT), Gradient Boosting (GB), Support Vector Machine (SVM) and so on. All these algorithms can be regarded as improved algorithms developed on the basis of support vector machines, and they solve some problems of traditional support vector machines from different perspectives respectively, and have achieved certain results.

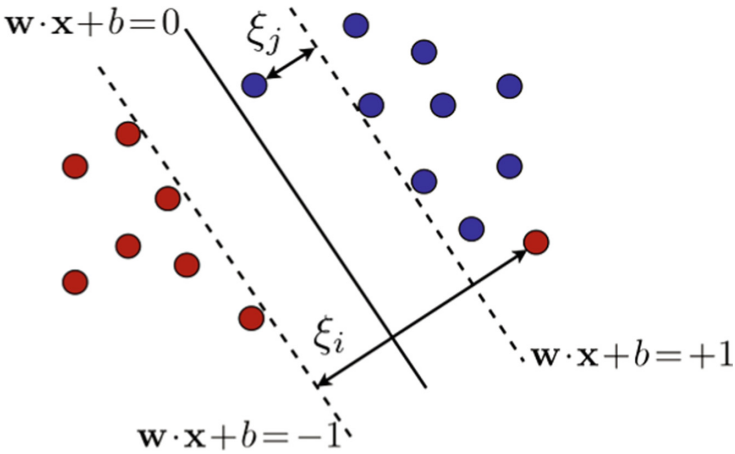

Fig. 4. Support vector machine

4 An Empirical Study of Machine Learning Methods in SSE Index Forecasting

4.1 Introduction to the Empirical Environment

This section introduces the process of data mining and analysis using the Python programming language in combination with common deep learning frameworks such as Scikit-learn and TensorFlow. To verify the validity of the proposed model, we selected historical data for the three-year period from 2018 to 2020 as the training set and predicted the stock price trend for the coming year. By fitting the historical data, the coefficients of each independent variable and their corresponding t-values were obtained, so that the degree of contribution of each independent variable to the dependent variable could be calculated. In addition, we constructed a classifier using a support vector machine algorithm to determine which category of up or down range the current stock price belongs to. Ultimately, we combined the two models to produce predictions for the SSE Composite Index for the coming year.

4.2 SSE Index Prediction Based on BP Neural Network

When using BP neural network for SSE index prediction, first, we need to determine the number of nodes in the input, hidden and output layers. In this paper, 10 features are selected as input layer nodes, 3 hidden layer nodes and 1 output layer node. Then, we use forward propagation algorithm to train the model and get the final results. Specifically, we set the training set as 876 sets of data and the test set as 294 sets of data. In addition, we also compared the prediction effect of BP neural network with the traditional linear regression model and found that the difference between them was not significant. Therefore, we believe that BP neural network is a feasible forecasting method that can be used for short-term forecasting of the SSE index. In summary, this section focuses on how to use BP neural network to forecast the SSE index. Through experimental verification, we conclude that BP neural network is an effective forecasting method for short-term forecasting of the SSE index [11].

4.3 SSE Index Prediction Based on Support Vector Machine

When using SVM model for SSE index prediction, firstly, we divide the data set into three parts: training set, validation set and test set. Then, the optimal parameters are determined by the grid search method and the cross-validation method is used to evaluate the model performance. Finally, the obtained model is used to predict the test set, and the results show that the model has high accuracy and generalization ability. Specifically, we found that the model performed best when the penalty factor C took a value of 20; and the type of kernel function affected the final prediction accuracy, so we chose RBF as the kernel function. In addition, due to the large variability among different industries, we also introduced the industry characteristic factor to further improve the prediction effect.

Comparison of BP Neural Network Results and Support Vector Machine Error Results

Table 1. The BP neural network results and the support vector machine error results are shown.

	Training samples directly output results MSE	MSE after reverse normalization of training samples	Verify the MSE after the reverse normalization of the sample	Verify the MSE after the reverse normalization of the sample
BP neural network	0.0037	86905	26019	26019
Support vector machine	0.0014	32560	24313	24313

The BP neural network results and the support vector machine error results are shown in Table 1.

From the comparison of the error results of the two models, the training sample error of the improved BP neural network output is larger than that of the support vector machine output, and from the validation sample error, both of them do not have a large error in fitting the samples, and the error of the support vector machine is slightly better than that of the BP neural network. The smaller validation sample error indicates that both models have good generalization ability [12].

5 Conclusion

In this paper, we focus on forecasting the Shanghai stock market using macroeconomic indicators and machine learning models. In the empirical analysis, we use multiple data sources and combine different algorithms to build the forecasting models. By comparing the experimental results, we find that selecting the appropriate feature set and optimizing the parameter settings can significantly improve the forecasting accuracy. At the same time, we also found that traditional time series analysis methods are not suitable for prediction tasks in all cases, while deep learning methods such as neural networks exhibit better performance. Therefore, combining the two can be further explored in the future to obtain more accurate prediction results. In addition, the data set involved in this study is limited and only some macroeconomic variables are considered, and we can try to introduce more relevant factors to expand the training set in the future.

References

1. Carrera Andrea. An Introduction to Macroeconomics. A Heterodox Approach to Economic Analysis[J]. Review of Political Economy, 2023(1):35-38.
2. Wang Jian, Shao Wei, Ma Chenmin, et al. Co-movements between Shanghai Composite Index and some fund sectors in China[J]. Physica A: Statistical Mechanics and its Applications, 2021:573.
3. Zaid Muayad Basheer, Omer Abdulrahman Jadaan, Hayder Dhahir Mohammed. Determinants of Macroeconomics in Jordanian Islamic Banks[J]. Journal of Business and Management Studies, 2022(4):41-44.

4. Ding F. The Empirical Analysis of Shanghai Composite Index based on GARCH Model[C]//Institute of Management Science and Industrial Engineering. Proceedings of 2019 9th International Conference on Social Science and Education Research (SSER 2019).Francis Academic Press, 2019:168-174.
5. Da Silva Sergio, Matsushita Raul. Editorial: Granularity in Econophysics and Macroeconomics[J]. Frontiers in Physics, 2022:77-80.
6. Wu Haijian, Li Qianqian. Empirical Study on Shanghai Composite Index Forecast Based on ARIMA Model[J]. Journal of World Economic Research, 2018(6):60.
7. Achdou Yves, Han Jiequn, Lasry Jean-Michel, et al. Income and Wealth Distribution in Macroeconomics: A Continuous-Time Approach[J]. The Review of Economic Studies, 2022(1):89-92.
8. Iwamoto Yasushi, Miyakawa Daisuke, Ohtake Fumio. Introduction to the special issue "SIR Model and Macroeconomics of COVID-19".[J]. Japanese economic review (Oxford, England), 2021(4):72.
9. Xiu Yan, Chen Xinye. Study on Prediction of the Shanghai Composite Index Based on EMD and NARX Neural Network[J]. Frontiers in Artificial Intelligence and Applications, 2017:296.
10. Rivot Sylvie. The Friedman-Lucas Transition in Macroeconomics: A Structuralist Approach[J]. The European Journal of the History of Economic Thought, 2021(4):28.
11. Ying-hong Dong, Hao Chen, Wei-ning Qian,et al. Micro-blog social moods and Chinese stock market: the influence of emotional valence and arousal on Shanghai Composite Index volume[J]. Int. J. of Embedded Systems, 2015(2):76-78.
12. Yong-Qiong Zhu, Ye-Ming Cai, Fan Zhang, "Motion Capture Data Denoising Based on LSTNet Autoencoder," Journal of Internet Technology, vol. 23, no. 1, pp. 11-20, Jan. 2022.