Research on forecasting Modeling method of Stock Distribution based on Principal component Analysis Neural Network

Xu Jiale, Zhong Weihua, Sun Yuan

School of Economics, Qingdao University, Qingdao 266000, Shandong

Keywords: Stock forecast; Long and short time memory neural network; Regression analysis; Principal component analysis; The time series

Abstract: Due to the influence of many factors, such as economic environment, political policy, market news and so on, it becomes very challenging to predict stock dynamics. The traditional methods of stock volume forecasting usually decompose stock trading volume, and then select appropriate models for different parts to model and forecast, but this method is difficult to grasp the intra-day periodic structure of stock trading volume. Five commonly used prediction methods for predicting stock price changes are studied, and the prediction analysis is carried out by gradually increasing the input dimension of the model. First of all, five optimized prediction models are established-autoregressive average model based on time series (ARMA), grey prediction model (GM (1), BP neural network model (BPNN), support vector regression (SVR) model based on improved grid optimization algorithm, long-term memory neural network model (LSTM), based on Tensorflow to study the model input of single dimension, that is, The closing price of each stock is used as the input of these five models. The principal component method is used to extract the features of multiple indexes of corn index, and then five kinds of neural network models are established by using the extracted principal components, and the opening price is predicted, and finally compared with the ARIMA model. The results show that the PCA-RNN model has achieved better results, is more suitable for the short-term prediction of stock prices, and can provide some reference for decision makers. It is found that the effect of LSTM-based machine learning algorithm is obviously better than other traditional machine learning algorithms. Then, the input dimension of the model is added, that is, 13 indexes that affect the stock price are used as the inputs of the LSTM model to predict the stock price. The mean square error of the model in the training set is 0.1438, which is compared with that of the BP network. The results show that the accuracy and stability of the LSTM network in predicting the intraday trading volume distribution are better than the BP network.

1. Introduction

Due to the economic environment, political policy, market news and other factors, it is very challenging to predict the stock dynamics. In 1900, Louis Bachelier, a French mathematician, first studied the stochastic characteristics of the behaviour of share prices and sparked his interest in studying them. Fan et al. [1] proposed the model ARMA(autoregressive average model) for prediction based on statistical methods in 2003.However, the ARMA model assumes a linear relationship between the lag variables and therefore can linearly approximate the real world complex systems, but fails to predict the evolution of nonlinear and non-stationary processes. With the development of machine learning, nonlinear artificial neural networks similar to human self-learning have good performance and can be used to predict financial time series, including back propagation neural network (BPNN) [2], radial basis function neural network [3], etc. Qiu et al. [4] used genetic algorithm to optimize the artificial neural network model and predict the trend of the Japanese stock market index price. Kumar et al. [5] decompose financial time series into input variables of BPNN through discrete wavelet transform to predict future stock prices. Therefore, in this paper, 13 index data affecting stock price are obtained through web crawler, and 14 features are extracted through news mining. These 27 dimensional features are input into THE LSTM model for dynamic prediction of stocks, so as to obtain better prediction effect.

2. Principal component analysis and LSTM parameter analysis

2.1 Principal component analysis

Principal component analysis is the most commonly used dimensionality reduction method, which converts a group of variables that may be correlated into a group of linearly unrelated variables through orthogonal transformation. the goal of principal component analysis is to replace a large number of related variables with a small number of unrelated variables while retaining the information of initial variables as much as possible. these converted variables are called principal components, and they are linear combinations of observed variables. It is hoped that fewer principal components can be used to approximate the full variable set, and there are two ways to determine the principal components: one is based on eigenvalues, which selects the principal components by selecting indicators whose eigenvalues are greater than 1, and the other is by calculating the cumulative variance contribution rate. when the cumulative contribution rate of variance is greater than or equal to 85%, these components can be extracted to reflect the original data set.

2.2 Principal component analysis

This system uses LSTM combined with TensorFlow deep learning to improve the algorithm of traditional technology analysis .In this study, it is found that when parameters are changed, the performance of the neural network will be significantly affected. Therefore, parameters must be carefully set to ensure the convergence of TensorFlow. The same amount of data is used to study the performance of the commonly used optimizer in the current deep learning training. The results are shown in Table 1.In the empirical study, the Adam optimizer not only shows good performance in terms of speed and entropy loss, but also shows the best performance in the training of prediction model algorithm.

Due to the high computational efficiency of Adam, it is suitable for dealing with the problem of large data sets, and the super parameters seldom need to be adjusted, and the memory is small. Therefore, Adam optimizer is suitable for updating the network weight based on training data iteration. It can be concluded that Adam can effectively solve practical deep learning problems for large models and data sets.

3. Access to information parameter

3.1 Principal component analysis

The data of the Corn Index (CL9) from January 18, 2016 to March 27, 2019, a total of 776 trading days were selected. The data were obtained from CCOM. According to the relevant information of stocks, 8 indexes including opening price, maximum price, minimum price, closing price, trading volume, Ma. MA1, MACD. DIF and OBV. OBV were selected, among which the opening price was used as the stock price forecast index, and other indexes were used as influencing factors for analysis. The change of the stock opening price over time is shown in Figure 1.

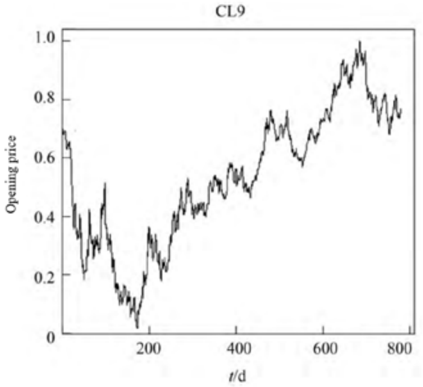

Fig.1 Opening price of corn index

The principal component is extracted as follows:

$$\begin{aligned} C_1 &= 0.434X_1 + 0.436X_2 + 0.435X_3 - 0.213X_4 + 0.434X_5 + 0.135X_6 + 0.423X_7 \\ C_2 &= -0.048X_1 - 0.021X_2 - 0.034X_3 - 0.625X_4 - 0.012X_5 - 0.777X_6 + 0.053X_7 \end{aligned} \tag{1}$$

3.2 RNN neural network modeling

The extracted principal components C1 and C2 were used as the inputs of the neural network for training. In order to select the optimal training model, the RNN neural network was compared with DNN neural network and BP neural network algorithm for analysis. The first 740 data were selected as training samples, and the last 36 data were selected as test samples. LSTM is an improved model of cyclic neural network (RNN), which can solve the problem of gradient disappearance and gradient explosion generated by RNN over long time series. It has been applied to image recognition, speech recognition and other fields. The principle of DNN neural network is to build a single layer of neurons layer by layer, and then fine-tune the Wake-Sleep algorithm to achieve the optimal prediction effect.C1 and C2 are selected as the input of the network and the opening price as the output. After many times of training, it is found that when the number of hidden layers is 10 and the number of nodes in each layer is 20, the prediction effect is the best. The learning process of BP neural network consists of two processes: forward propagation of signal and back propagation of error. In forward propagation, input samples are introduced from the input layer, processed layer by layer through hidden layer, and then transmitted to the output layer. This principle is used for modeling. When the optimal number of iterations is 10000, the output results of each layer reach the optimal level,set up the learning rate and hidden layer nodes and the number of iterations, and through the cross validation method continuously adjust and to select the optimal parameter values, the vector to calculate the optimal 0.1 in the experiments, the number of hidden layer nodes is 1, stabilized after iterative about 500 times. The results are shown in Figure 2.

Will crawl the data is divided into two parts: the part from January 2015 to May 2019 in chronological order of the top 80% of each stock index data, as a training set is used for training model, and 175 data after the training set as a validation set to adjust the LSTM super parameter of the model, such as the number of iterations epochs, LSTM Numbers of hidden layer neurons, etc.);The other part is the indicator data of the latter 20%, which is used as a test set for prediction. The crawled original data is processed with first-order difference to make the sequence stable. Because only on the premise of stationary sequence, can carry on the follow-up study

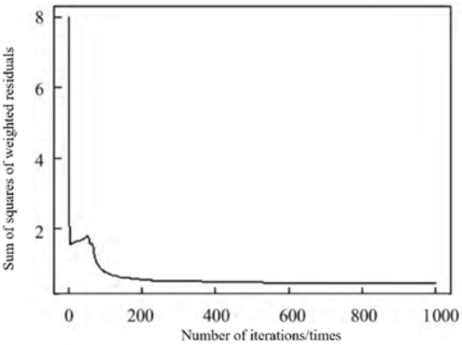

Fig.2 Error image of Elman function

As shown in Table 1, PA-FNN obviously has a better prediction effect. The DNN network is prone to problems such as excessively long training time and overfitting, which require a high degree of parameter selection. However, the traditional BP neural network has many iterations, which leads to a slow convergence rate and a structural choice Diversity, while paying attention to the omission of some hidden neurons.

Tab.1 Square correlation coefficients under three models

	PCA-DNN	PCA-BP	PCA-RNN
R^2	0.8377	0.9133	0.9749

Under the learning of Elman network, pA-RNN neural network model can effectively avoid possible problems such as overfitting and improve the iteration rate of the algorithm by properly modifying the weights.

3.3 ARIMA model

It is found in Figure 3 that the stock data of corn index changes over time and has a certain trend. The original data is not stable, so the data should be stabilized first when ARIMA model is used for modeling. Usually, the non-stationary sequence is processed by difference. Through the stationarity test, it can be found that the data after the primary difference basically tends to be stable, and can be modeled and analyzed according to this sequence. According to the ACF figure, it was found that the sequence gradually decreased to 0 with the increase of order, and the PACF figure found that the sequence gradually decreased to 0 after order 1, so the model was established as ARIMA (1, 1, 0).

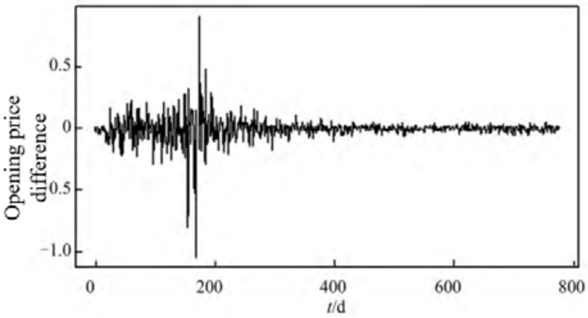

Fig.3 First order difference sequence diagram

In order to judge the prediction effect of the model, it is often evaluated by observing whether the residual error of the model meets the normal distribution. By testing p value, the residual parameter p of the model is 0.1603, which passes the normal test. Meanwhile, R2 = 0.9327 is calculated, indicating that the goodness of fit of the model is relatively high. Then, it is found through experiments that the single-dimensional closing price model is not sensitive to the fluctuation of stock price when the trading volume, free circulation capital stock and other fluctuations are large, so the input dimension of the model is considered to be added. Therefore, 13 indexes affecting stock price are input into LSTM model to establish correlation with stock price and carry out research. However, through the research and analysis of different stocks and different trading days, it is found that the 13-dimensional index can only reflect the impact of stock investors on the stock price volatility, and the stock price will rise when the company has changes in the management structure or breaking news events. The volatility that is not within the control of shareholders. Therefore, in this study, 14 features are extracted by news mining and then input into the LSTM model for further prediction of the stock market. This enables this study to predict the impact of emergencies on the stock market, and help stock investors and companies to prepare for emergencies in advance, so as to reduce economic losses. Since the real-time trend of stocks is also affected by many other external factors, how to analyze different user types and user emotions so as to obtain a more reasonable stock portfolio model is the focus of this research.

4. Conclusion

This paper compares and studies the effects of 5 models and 3 input dimensions on stock market prediction. First of all, in the study of model input with single dimension, it is found that since LSTM is a time-varying memory model based on timing characteristics, different from the traditional BPNN without timing characteristics, the ACCURACY of LSTM based on deep learning is significantly better than that of the traditional machine learning algorithm in dealing with the stock price prediction of continuous time series.

Although the dynamic prediction model in this paper includes the financial crisis and other unexpected situations in the historical series, its ability to study financial risks is still limited, and new theoretical methods need to be continued to be studied to solve it.For example, in future work, mixed model can be used to deal with complex stock price time series with linear and nonlinear variables.

Neural network model has hidden layer units, contains a lot of multi-layer neural network nonlinear transformation, can make it more flexible concisely express complex nonlinear functions, and build sophisticated statistical models, for the prediction of the stock market to be more efficient and accurate, but the depth of the hidden layer neural network training there are still some limitations, it is to continue to explore and improve the problem in the future.

References

[1] Zhang Jie, Qian Weidong. Stock investment strategy based on quantitative analysis [J]. Journal of Hebei Northern University (Natural Science Edition), 2020 (11).

[2] Wang Ting, Xia Yang Yuxin, Chen Tieming. Stock short-term trend prediction based on multi-class characteristic system [J]. Computer Science, 2020 (S2).

[3] Zhang Shengdi. Research on Stock trend tracking Strategy based on reinforcement Learning [D]. Xi'an University of Technology, 2020.

[4] Lin Chunyan, Zhu Donghua. Research on Stock Price Forecast based on Elman Neural Network [J]. Computer applications, 2006 (02).

[5] Zhu Rong. Research on the Evaluation and Control of Enterprise Financial risk [D]. Northeast University of Finance and Economics, 2007