Research on Stock Selection Method Based on LSTM Neural Network

Yifan Gao(✉)

University of British Columbia, 17 Foursquare Tips 2329 W Mall, Vancouver, BC V6T1Z4, Canada
gao78@wisc.edu

Abstract. In the field of investment, the selection of good target stocks is one of the keys to the ultimate success of the investment activity. Stock prediction is a study that every investor is trying to do, ordinary investors confirm stock selection for trading by means of technical analysis, and researchers analyze stock data by building mathematical models. Stock data are represented as classical financial time series, and the use of neural networks for stock data prediction is a hot research topic in recent years. In this paper, we analyze the stock investment risk and investment analysis methods based on the actual process of stock investment selection, and analyze the applicability of LSTM in stock investment selection from the perspective of stock selection ability under the large number of stock market investments. The experimental results show that the proposed method has improved the accuracy of stock prediction compared with the single LSTM prediction model, and can predict the stock trend accurately and effectively to a certain extent.

Keywords: LSTM · Neural Network · Stock Prediction

1 Introduction

With the progress of reform and opening up and the rapid development of China's economy, China's stock market has seen unprecedented opportunities for development and has taken an important position in the world stock market. Stock speculation has gradually become a "universal" behavior, and data from the China Clearing and Settlement Center shows that by the end of March 2020, the number of investors in Shanghai and Shenzhen was 150 million [9]. However, China's stock market is mostly retail, irrational investment, influenced by multiple factors, resulting in dramatic fluctuations in stock prices, resulting in a complex pattern of changes within the stock market, poor stability, and difficulty in obtaining effective information [8]. This paper further analyzes the applicability of LSTM neural networks in practical applications by applying LSTM neural networks and traditional BP neural networks to stock investment selection, and provides some support for theoretical research on the technology of artificial intelligence machine learning in the field of stock investment selection [6].

N. Radojević et al. (Eds.): ICAID 2022, AHIS 7, pp. 869–876, 2023.
https://doi.org/10.2991/978-94-6463-010-7_87

2 Stock Investment Risk and Investment Analysis Methods

The price of stock market transactions often changes rapidly, the price is up for profit, the price is down for loss. Therefore, while investors expect to earn high returns, they are bound to bear the corresponding huge risks [4]. At this point, how to avoid risks and reduce them becomes the main topic of concern for investors.

Currently, there are two main types of analysis methods for stock investment in terms of the characteristics and perspectives of research paradigms: basic analysis and technical analysis.

The basic analysis mainly takes the intrinsic value of the enterprise as the main research object, and analyzes the investment value of the listed company from various factors that determine the value of the enterprise and affect the stock price, such as the operation status of the enterprise and the development prospect of the industry. Investors can select companies with investment value to buy and sell stocks based on the basic analysis of listed companies to obtain corresponding profits.

Technical analysis mainly takes the intuitive behavior performance of stock price rise and fall as the main object of study, with the main purpose of predicting stock price fluctuation patterns and trends, starting from the daily technical indicators of stock price changes, and analyzing the fluctuation pattern of stock market prices, so as to predict the trend of price fluctuations, to find the buying and selling points of stocks and gain corresponding profits.

Stock portfolio refers to a method of selecting and matching stocks according to certain rules and principles in order to reduce investment risk, based on the risk level and profitability of various stocks. The purpose is to reduce the risk and maximize the overall return. The theoretical basis is that the stock market is generally not synchronized with the rise and fall of various types of stocks, there are always ups and downs, one after the other. Therefore, when the investment in a stock may not be profitable because of the temporary decline in its price, but also in some other upward trend of the stock to obtain a certain amount of income, so you can achieve the purpose of risk avoidance [5].

3 LSTM Long Short Term Memory Artificial Neural Network

LSTM neural networks are a type of RNN neural network, but they differ from ordinary RNN neural networks in that they add an extra type of neuron called a “memory cell” and a variety of gating structures that can control the memory cell. The LSTM neural network has an additional neuron called “memory cell” and a variety of gating structures that can control the memory cell [2]. This special structure can avoid the redundancy of information, thus effectively alleviating the problems of gradient disappearance and gradient explosion that often occur in recurrent neural networks.

Before the emergence of LSTM, there were two major problems: first, gradient vanishing easily in RNNs; second, time series data that were expanded in step length were generally poorly analyzed in RNN models for experimental data. The proposed modified long short-term memory network improves the structure of the hidden layer to a large extent, thus solving the two traditional problems.

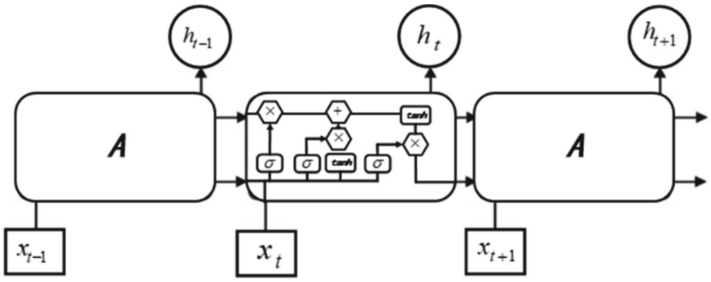

Fig. 1. The three gates of the LSTM

Figure 1 presents a diagram of the memory cell structure of the LSTM model, which introduces several gating structures including forgetting gates, input gates and output gates. Among them, C, called cell state, plays a similar role to h_t in traditional neural networks, while the forgetting gate plays a unique forgetting role in the model, mainly by imposing certain restrictions on the probability and then acting on the hidden state of the previous layer.

According to the principle of the forgetting gate, it is shown that the structure of the model also passes the hidden state within a certain probability limit, so the f_t in the model is set in the range of 0 to 1, and the Sigmoid function, which fits this model, is chosen among many activation functions [3]. The main function of the input gate in the LSTM model is to filter the current input information and to determine the proportion of information in the current cell state, mainly considering the weight of the current information.

According to the principle of the action of the input gate, a certain proportion of the input information is also required, thus *it* is set to a range of values from 0 to 1 in the model. In the environment of the input gate, the Sigmoid function is still used to function, and the formation of a new cell state is based on the new information received, which is then obtained by multiplying x_t and C_t. In the LSTM model, the forgetting gate and the input gate have in common that they both probabilistically qualify the information at different moments, which in turn changes the current cell state G, specifically the process of updating from the original C_{t-1} to the current C.

In the LSTM model structure, the current cell state exists to provide the required information for the output gate, and the extracted information is used to represent, among others, the structural diagram in which o_t takes values ranging from 0 to 1. The Sigmoid, an activation function represented by b, is used to update the cell state in a timely and efficient manner with a certain probability, that is, to initially decide which data are worth leaving behind [1]. The tanh function calculates the input cell states and finally decides which information must be left behind. In the engineering of the formula derivation, C_t then represents the current cell state, which together with the information h_{t-1} contained in the state of the hidden state at the previous moment of t determines the size of the scarf, and the whole process W_c does not have an impact on the current cell state in terms of accounting. In traditional recurrent neural networks, the gradient disappearance is generally highly associated with conservation, but the LSTM model is an improved cornerstone of traditional recurrent neural networks, mainly by adding

several gating organizations to the structure, and when f_t in the forgetting gate is opened, the LSTM model can quickly pass the gradient of C_t to the previous moment of t, i.e., the cell state C_{t-1} at t − 1, in a way that greatly reduces the frequency of the gradient disappearance problem.

4 Design of LSTM Stock Prediction Method Combining Correlation Features

4.1 Predictive Method Design

Feature selection is an extremely important part in the learning process of neural networks. In this paper, we will use LSTM neural network to learn the historical data of stocks, find the nonlinear relationship with the change of target value from the input features, store them in the network in the form of weights and biases, and then verify them using the test set data.

The ability of LSTM network to predict time series lies in its ability to pass the state of the time series with retention during the learning process. Let the time series be $\{X_t\}$, where the target value sequence is $\{Y_t\}$, and use the time window m to predict Yt, which can be represented as follows in the LSTM: $(X_1, X_2, \cdots, X_m) \rightarrow (Y_1)$, $(X_2, X_3, \cdots, X_{m+1}) \rightarrow (Y_2), \ldots$ If a traditional BP neural network is used for prediction, since BP neural networks do not yet have the concept of time series, they can only transform the series into a one-dimensional matrix $(X_1T, X_2T, \cdots, X_mT) \rightarrow (Y_1)$. The LSTM model will separately perform the prediction for each time series $X_1, X_2, \cdots$. The LSTM differs from the traditional BP in that it retains some of the useful memory in the time series during the transfer process, providing more information for the next moment. Models are also divided into single-step prediction and multi-step prediction [7]. Single-step prediction refers to the use of a sliding window to input the desired time series uniformly into the model, output a single prediction value, and then slide the window to the next moment to predict the next prediction value, with no direct connection between two adjacent prediction values; multi-step prediction, on the other hand, is beyond the first prediction value provided by the feature values, and subsequent predictions gradually use the predicted values to replace the true values as input for the next prediction step.

Since multi-step forecasting is prone to bias, this study uses a single-step forecasting method that incorporates into correlation features based on traditional basic data, fundamental data, and technical features to achieve multi-dimensional refinement of input features, and then uses a multi-layer LSTM neural network structure to forecast the closing price.

Prediction with the LSTM model requires the following steps.

1. Perform feature design and design a relevant feature extraction method. In Sect. 3, we focus on how to extract correlation features, in addition to feature design by referring to existing commonly used technical analysis indicators.

2. After the design of the features, the acquisition of the dataset is needed. Since professional databases require fees to be used, and the financial trading terminal of TDX has certain restrictions on information export, while the tushare financial interface provides

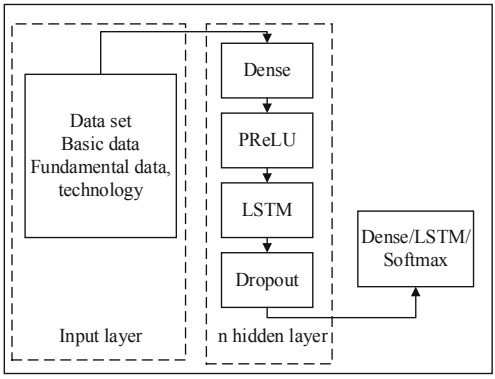

Fig. 2. Network Model Structure Diagram

a free data interface for researchers to use, this study acquires data from tushare.pro and uses the acquired data for cleaning and generating the corresponding features.

3. Before feeding the training features into the neural network, the data set needs to be divided into a training set and a test set, which is normalized due to the quantitative gap that often exists between the features.

4. With the addition of correlation information, a total of four types of input features are used in this study, including basic stock data, fundamental data, technical data and correlation data, which are combined to find the more appropriate features for stock prediction. In this paper, we will first conduct a categorical regression test to test the influence of various data sets on the prediction results and provide help for the selection of feature sets for subsequent regression prediction. In order to verify the effectiveness of the model, BP neural network, SVM and traditional method Linear Regression will be introduced to compare with the model constructed by LSTM.

4.2 Model Structure Design

The model is mainly built using the Dense layer, Prelu layer, LSTM layer and Dropout layer in keras. In the output layer, the closing price regression prediction uses the Dense layer with 1 neuron or LSTM to regress the prediction result, and the output value is a single prediction value; while the up and down classification prediction needs to add softmax activation function to achieve various types of prediction probability output, the number of neurons output with the number of categories of classification. The structure of the model is shown in Fig. 2, where the number of layers of hidden layers needs to be tested and finally obtained.

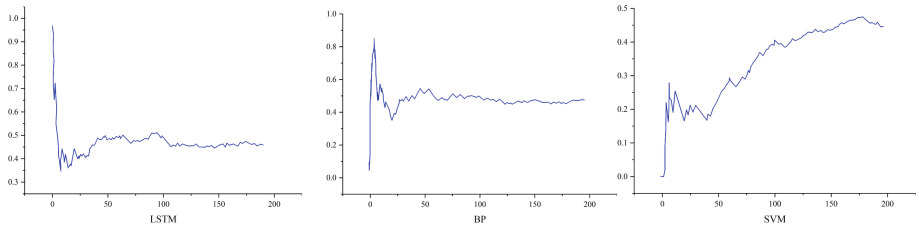

Fig. 3. Comparison of prediction results of multiple models for classification

4.3 Experimental Results and Analysis

In this paper, BP neural network and SVM support vector machine were selected to compare the prediction accuracy with the constructed model, and the prediction accuracy is shown in Fig. 3. The accuracy of the prediction set is shown in Fig. 3. The statistics of the prediction set shows that there are 197 prediction samples with the ratio of 57:61:79. In contrast, the results of the LSTM and BP neural network calculations are scattered across types and do not appear to be completely biased towards the same type.

Forecast is actually an estimate of the future situation, so there is always a certain deviation between the forecast value and the actual value, and the evaluation of the model is often calculated by the forecast deviation, the size of the forecast deviation determines the size of the forecast accuracy, the larger the error value, the lower the model forecast accuracy, and vice versa, the higher the accuracy. Commonly used error evaluation metrics are RMSE, MAE and R^2.

Use the forecast results to make a trend judgment. If tomorrow's forecast price is higher than today's forecast, the model is considered to be predicting a trend up, and the operation of buying at today's closing price and selling at tomorrow's closing price will be carried out. This is used to make a judgment on the model's trend prediction ability and to compare the profitability of the model. In order to simplify the calculation and not to consider the commission, each purchase and sale is done with 100 shares, and this is used as an evaluation indicator for the model.

This study uses BP neural network and Linear Regression as a comparison, using the same data for testing and judging by the above indicators, the degree of fit of each model is shown in Fig. 4. And among the regression prediction, a better prediction model DDL was achieved and was able to predict the general trend of other stocks, and achieved better results relative to BP and Linear Regression under the comparison of several indicators such as RMSE, R^2, error value and self-designed earnings value.

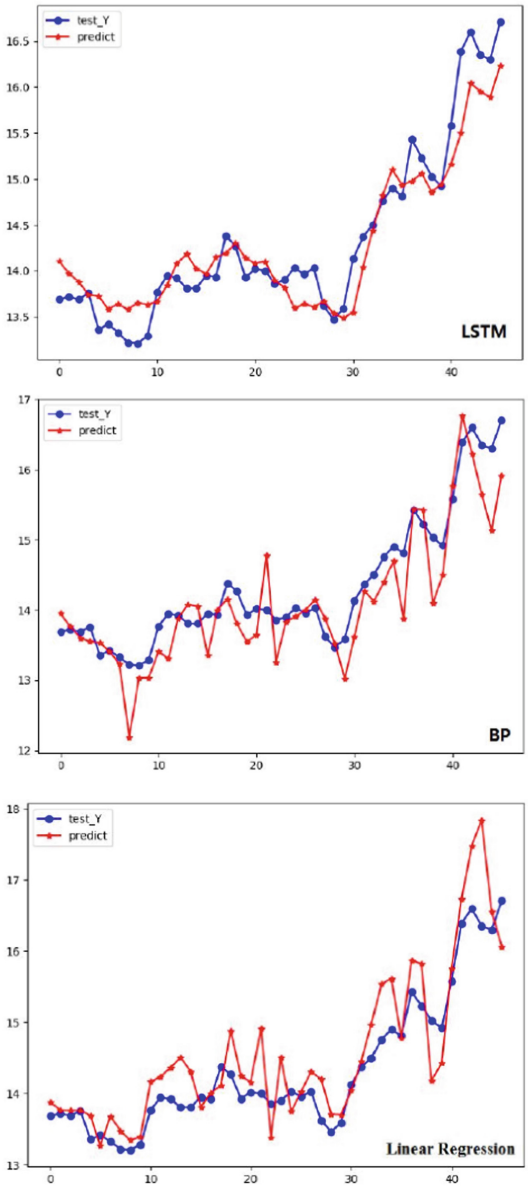

Fig. 4. Comparison of model predictions

5 Conclusion

With the rapid economic growth in China, people's living standards are gradually improving, with surplus assets to invest, the number of people opening accounts in China's stock market is increasing year by year, and many people want to grow their wealth through

stock market investments. And the risks that exist in the stock market discourage people, how to analyze the effective content from the stock market information to assist in trading is what every investor wants to find. In this paper, we find the better deep learning model through the combination of multiple neural network structures and achieve stock prediction by this model combined with stock correlation features.

References

1. Ma CQ, Yang JL, Ren YS et al (2021) Research on price forecasting of CSI 300 index based on H-LSTM model. J Econometrics 1(02):437–451
2. Jiao FX (2021) Research on stock price prediction based on LSTM neural network. Digital Technol Appl 39(03):220–222
3. Li JP (2021) Research on stock trend prediction based on two-layer LSTM model. Technol Innov 07:50–51
4. Fu K, Yin XY, Wang Q (2020) Research on multi-feature stock trend prediction based on LSTM. J Beijing Univ Posts Telecommun (Soc Sci Edn) 22(05):17–27
5. Zhang QQ, Lin TH, Qi XY et al (2020) A review of research on stock prediction based on machine learning. J Hebei Acad Sci 37(04):15–21
6. Zhang YA, Yan BB (2020) A deep learning composite forecasting model for stock market. Comput Sci 47(11):255–267
7. Meng Y, Xu QJ (2021) Stock price prediction based on LSTM neural network-Marshall chain. Time Finance 11:3–6
8. Li ZH, Zhang KL (2020) A comparative study of automobile sales forecasting models based on BP algorithm and LSTM algorithm. J Econ Res 2020(20):84–88+93
9. Bao ZS, Guo JN, Xie Y et al (2020) LSTM-GA based stock price rise and fall prediction model. Comput Sci 47(S1):467–473