2019 19th International Conference on Control, Automation and Systems (ICCAS 2019)
Oct. 15~18, 2019; ICC Jeju, Jeju, Korea

Stock prediction based on random forest and LSTM neural network

Yilin Ma[1] , Ruizhu Han[2*] and Xiaoling Fu[3]

[1] School of Economics and Management, Southeast University,
Nanjing, 211189, China (ylmaseu@163.com)
[2] School of Economics and Management, Southeast University,
Nanjing, 211189, China (daodao-777@163.com) * Corresponding author
[3] School of Economics and Management, Southeast University,
Nanjing, 211189, China (fufei1980@163.com)

Abstract: The data in the stock market are intricate. Principal Component Analysis (PCA) based on LSTM neural network can remove noise and improve the accuracy of stock prediction. A stock prediction model based on random forest and LSTM neural network is proposed to further improve the performance of stock prediction. Based on the data of Shanghai Composite Index from 2013 to 2017, this model and PCA + LSTM neural network model are simulated and compared. The experimental results show that this model is more suitable for stock prediction than PCA + LSTM model. In addition, the returns of trading strategies based on the above two models are higher than the benchmark buy-and-hold strategy, and the trading strategies based on the proposed model perform best.

Keywords: principal component analysis, LSTM neural network, random forest, stock prediction.

1. INTRODUCTION

Stock market is essentially a dynamic, non-linear, chaotic and noisy system, so the prediction of stock price and its fluctuation has always been a very difficult task in time series prediction. In fact, the fluctuation of stock price is influenced by many factors, such as political events, economic situation, corporate policy, bank interest rate, investor sentiment and so on.

In recent years, due to the advantages of machine learning technology in data mining, many researchers use machine learning models such as neural networks, random forests and support vector machines to predict stock prices and trend [1-6].

On the one hand, LSTM neural network, as a special recurrent neural network, is designed to solve the problem of gradient disappearance in the training of recurrent neural network, so as to learn the long-term dependence of time series data, and then make the neural network more effective in mining the correlation of time series data, and more accurately predict stock prices. However, in the research of LSTM neural network used in stock price forecasting [7-8], most of them only train LSTM neural network with untreated market sequence data such as opening price, closing price, maximum price and lowest price and technical index, which makes the training model absorb a lot of noise and can not effectively analyze and screen the input characteristics, thus leading to the unsatisfying forecasting performance. On the other hand, due to the frequent changes of a large number of noise and important features in the stock market, stock prediction becomes complex and the prediction effect is not good. Principal Component Analysis (PCA) is a classical method for de-noising and extracting main features. It preprocesses the original data and improves the prediction effect by dimensional reduction. However, classical feature extraction methods such as principal component analysis (PCA) can generate satisfying results.

Unlike the statistical method of principal component analysis, random forest uses multiple decision trees to train and predict samples. It has the ability of feature analysis, that is, the trained model can measure the importance of each input feature. Random Forest (RF) can be introduced as feature extraction to screen important input features, which can further improve the performance of stock price prediction. Therefore, a new stock price forecasting model based on random forest and LSTM neural network is proposed in this paper. The random forest model is used to analyze the importance of input features, and then some important input features selected are combined with LSTM neural network for forecasting.

As a new financial market, the fluctuation of stock price in China's stock market attracts the attention of many researchers and investors. Therefore, this paper predicts the stock trend of Shanghai Composite Index, an important index in China's stock market, and then guides investment practice.

2. MODEL

The prediction model based on random forest and LSTM neural network consists of two parts: random forest and LSTM neural network. Specifically, firstly, the random forest model is used to fit the training data, then the importance score of each input feature based on the training data is obtained, then the more important input features are screened out, and the training data set is reconstructed. Secondly, the new

training data are imported into the LSTM neural network in time sequence for training, and then the validation data are used to monitor the training process to prevent the training from over-fitting. Finally, the test data are imported into the trained LSTM neural network for prediction. Next, we introduce the random forest model and LSTM neural network model respectively.

2.1 Random forest

Random Forest (RF), as a classical ensemble model, is a classifier composed of multiple decision trees, and its output category is determined by the number of individual outputs. This special structure makes it possible to reduce the defects of individual classifiers and synthesize the advantages of each classifier, thus producing better results than a single model.

The construction of random forest consists of two aspects: random selection of data and random selection of characteristics. Specifically, the random selection of data refers to the use of Bagging sampling from the original data, and the random selection of features refers to the random selection of some features as training feature sets when each node of the decision tree splits. Random forest model is very easy to implement, because only two main parameters need to be considered: the number of decision trees and the number of features randomly selected by each node of the decision tree. According to the research of Breiman [9], this paper sets the number of decision trees to 500, and the number of features selected randomly by each node of the decision tree is $\log_2 d \approx 7$, which $d = 180$ is the total number of input features.

2.2 LSTM neural network

LSTM was proposed by Hochreiter and Schmidhuber [10] in 1997. Its design was originally designed to enable recurrent neural networks to learn long-term dependencies in time series data. Specifically, the traditional recurrent neural network can learn the data correlation of long time interval theoretically, but in practical application, the recurrent neural network can only effectively learn the data correlation of short time interval, but can not perform well about the data correlation of long time interval. The LSTM neural network based on the delicate design of its hidden layer nodes can effectively solve the above problems.

LSTM has four components: forget gate, input gate, output gate and cell state. LSTM controls information discarded or added through three gates, namely, forget gate, input gate and output gate, so as to realize forgetting or memory function. The forget gate is a sigmoid function that controls the forgetting degree of the cell state of the previous cell through the output h_{t-1} of previous cell and the input x_t of this cell. The input gate combines a tanh function to control the input information. Specifically, the tanh function generates a new input information $\overline{C_t}$, while the input gate generates i_t through a function similar to the forget gate to control the information of the input cell state. The output gate controls the output of the current cell state through variable o_t and tanh functions. Cell status C_t is determined by the forget gate f_t and the input gate i_t.

$$f_t = sigmoid(W_f x_t + U_f h_{t-1} + b_f) \quad (1)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \overline{C_t} \quad (2)$$

$$i_t = sigmoid(W_i x_i + U_i h_{t-1} + b_i) \quad (3)$$

$$\overline{C_t} = \tanh(W_C x_t + U_C h_{t-1} + b_C) \quad (4)$$

$$o_t = sigmoid(W_o x_t + U_o h_{t-1} + b_o) \quad (5)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (6)$$

LSTM neural network through the above elaborate design, so that it can effectively mine the correlation between the data with long intervals in time series data, so that it can more effectively predict future stock prices and fluctuations. Because the single hidden layer neural network can approximate any non-linear function theoretically, this paper uses the single hidden layer LSTM neural network to carry on the experiment, and the number of hidden layer nodes is set to 300.

3. Experiments

3.1 Data Preprocessing and Dimension Reduction

The Shanghai Composite Index is an important index in China's stock market. This paper uses the stock price of the Shanghai Composite Index (SCI) as the experimental data. The input characteristics are 35 technical indicators. See Table 1. Most of these indicators have been used in the literature [11-12] and achieved good results. Therefore, this paper adopts the above indicators.

Table 1 Technical indicators.

Open price (x_o)	Close price (x)
High price (x_h)	Low price (x_l)
BIAS (5,10)	DIFF (12,26)
DEA (5,10)	RSI (6,12)
Momentum (6,12)	Williams index (5,10)
Oscillator (6,12)	Price rate of change
Moving average (5,10,20)	Stochastic (%K, %D)

Psychological line	True range of price m ovement
$(x(t)-x(t-1))/x(t-1)$	
$(x(t)-x_o(t))/x_o(t)$	
$(x(t)-x_l(t))/(x_h(t)-x_l(t))$	
$(x(t)-MA20(t))/MA20(t)$	
$(MA5(t)-MA5(t-1))/MA5(t-1)$	
$(MA20(t)-MA20(t-1))/MA20(t-1)$	
$(x(t)-bollinger_{upper})/bollinger_{upper}$	
$(x(t)-bollinger_{lower})/bollinger_{lower}$	
$(x(t)-\min(x(t-1),x(t-2),...,x(t-10)))/\min(x(t-1),x(t-2),...,x(t-10))$	
$(x(t)-\max(x(t-1),x(t-2),...,x(x-10)))/\max(x(t-1),x(t-2),...,x(x-10))$	

The experimental data were collected from January 4, 2013 to November 30, 2017. (The original data were derived from Wind Information software and the technical indicators were calculated from the original data). We deleted some days of missing or abnormal data, and the final data contained 1194 days' data. For each technical index, there is a big difference between its fluctuation range, so the input technical index is standardized. The process is as follows:

$$x_i = \frac{x_i - \overline{x_i}}{\sigma_i} \tag{7}$$

where $\overline{x_i}$ and σ_i represent the mean and variance respectively. Then, we use random forests to select features for standardized data.

3.2 Experimental process

The training process of LSTM model is in the form of window marking, that is, using the data of the past 30 days to predict the direction of today's trend, and using 1 and 0 to express the moving direction of stock prices respectively. This paper divides the training data into three parts: the first 70% as training data, the middle 15% as validation data and the last 15% as test data. The function of validation data is to prevent the over-fitting problem of neural network training. Next, we will introduce the experimental process in detail.

Firstly, we use the random forest model to select features, that is, we use the above 35 technical indicators as input features to predict next day's stock trend. The training data is the data contained in the training set and validation set of the LSTM model mentioned above. Secondly, the more important features are selected according to the importance of the features in the model. Finally, a new training data set is constructed based on the selected input features. The new data set is classified according to the above data classification methods, and then trained and predicted by LSTM neural network.

In order to better compare the advantages of the proposed model, we use PCA + LSTM and LSTM models as benchmarks. Keras deep learning package and Scikit-learn machine learning package are used to realize LSTM neural network and random forest respectively.

4. EXPERIMENTAL RESULTS ANALYSIS AND TRANSACTION SIMULATION

The importance of input features derived from random forests is shown in Fig. 1. As shown in Fig. 1, there are many features whose feature importance score is less than 0.03. Therefore, we choose 0.03 as the threshold to select the input features whose score is greater than 0.03. Finally, six input features are screened out, as shown in Table 2.

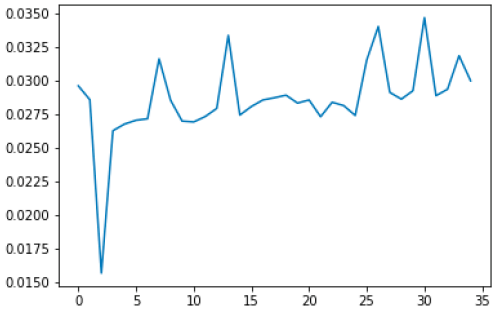

Fig. 1 Feature importance.

Table 2 Selected technical indicators.

BIAS (5)	True range of price m ovement
$(x(t)-x(t-1))/x(t-1)$	
$(x(t)-x_o(t))/x_o(t)$	
$(x(t)-x_l(t))/(x_h(t)-x_l(t))$	
$(x(t)-bollinger_{upper})/bollinger_{upper}$	

The performance of the model is measured by Accuracy and F-measure, which are defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{13}$$

$$F-measure = \frac{2\cdot P\cdot R}{P+R} \tag{14}$$

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN} \tag{15}$$

P, R, TP, TN, FP and FN denote precision, recall, true positive, true negative, false positive and false negative respectively.

Table 3 Experimental results of accuracy.

	Accuracy	Standard deviation
RF+LSTM	61.18%	$2.083*10^{-2}$
PCA+LSTM	60.49%	$2.047*10^{-2}$
LSTM	48.26%	$2.565*10^{-2}$

Table 4 Experimental results of F-measure.

	F-measure	Standard deviation
RF+LSTM	0.7589	$9.852*10^{-4}$
PCA+LSTM	0.7435	$2.958*10^{-2}$
LSTM	0.4965	$6.713*10^{-2}$

Table 3 Experimental results of rate of return.

	Rate of return	Standard deviation	Information ratio
RF+LSTM	5.90%	$1.494*10^{-3}$	25.74
PCA+LSTM	6.08%	$8.954*10^{-3}$	4.495
LSTM	4.12%	$1.078*10^{-2}$	1.912
Buy-and-hold	2.06%		

The experimental results show that the performance of RF + LSTM model is better than that of PCA + LSTM model, and both are significantly better than LSTM model. So the following is only a comparison of the two models. In terms of accuracy index, the average prediction accuracy of RF + LSTM model is higher than that of PCA + LSTM model, but the fluctuation of prediction accuracy is slightly higher than that of random forest model. For F-measure index, the prediction accuracy of RF+LSTM model is higher than that of PCA+LSTM model, and its volatility is significantly reduced. The experimental results are shown in tables 3 and 4.

To further illustrate the advantages of the proposed model, we conducted a simple transaction simulation to study whether high accuracy and F-measure mean high returns. Specifically, when the model predicts the next day's stock price rise, traders choose to buy or hold, and the profits are approximated by the next day's change amount. On the contrary, they sell and their profits are recorded as zero. For simplicity, this paper assumes zero transaction costs, dividends and taxes, and prohibits leveraged trading and short selling. In contrast, we simulated the buy-and-hold strategy as a criterion to measure the predictive model. The transaction simulation period of the model is the model test period. At the same time, we use the rate of return, the standard deviation of rate of return and the information ratio (i.e., excess return / the standard deviation of excess return) to describe the ability of the model. The experimental results are shown in Table 5.

The results show that: 1. The return of trading strategies based on the above three models is greater than that of buying and holding strategies, which shows that the above three models are suitable for forecasting stock prices and guiding practice. 2. The information ratio of the transaction strategy of the RF+LSTM model proposed in this paper is significantly higher than that of the other two strategies, which shows that the model can effectively mine the potentially important information in the input data and obtain stable high returns.

5. CONCLUSION

This paper presents a stock price forecasting model based on random forest and LSTM neural network, and compares the model with PCA+LSTM neural network and LSTM neural network. Through the analysis of input features importance by random forest model, we can find out the more important input features at this stage. Experiments show that the feature extraction method is more effective than the classical PCA algorithm and can visually display the filtered features. Although LSTM neural network can regularly delete invalid information and retain important information when importing time series data because of its exquisite design, the experimental results show that the simple use of raw data is not good for training, so it needs to combine more efficient feature extraction methods. Therefore, the model proposed in this paper provides a good direction for future research.

6. ACKNOWLEDGE

This research was supported by National Natural Science Foundation of China (No. 71390335).

REFERENCES

[1] E. Guresen, G. Kayakutlu, T. U. Daim, “Using artificial neural network models in stock market index prediction,” *Expert Systems with Applications*, , Vol. 38, No. 8, pp. 10389-10397,2011.

[2] M. Qiu, Y. Song, “Predicting the direction of stock market index movement using an optimized artificial neural network model,” *Plos one*, Vol. 11, No. 5, pp.e0155133, 2016.

[3] Y. Kara, M. A. Boyacioglu, “Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange,” *Expert Systems with Applications*, Vol. 38, No. 5, pp.5311-5319, 2011.

[4] K. Zbikowski, “Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy,” *Expert Systems with Applications,* Vol. 42, No. 4, pp. 1797-1805, 2015,

[5] J. Patel, S. Shah, P. Thakkar, et al, “Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques,” *Expert Systems With Applications*, Vol. 42, No. 1, pp. 259-268,2015.

[6] M. Ballings, D. V. D. Poel, N. Hespeels, et al, "Evaluating multiple classifiers for stock price direction prediction," *Expert Systems with Applications*, Vol. 42, No. 20, pp. 7046-7056, 2015.
[7] L. D. Persio, O. Honchar, "Recurrent neural networks approach to the financial forecast of Google assets," *International journal of Mathematics and Computers in simulation,* Vol. 11, pp. 7-13, 2017,
[8] R. Xiong, E. P. Nichols, Y. Shen, "Deep Learning stock volatility with Google Domestic Trends," arXiv:1512.04916.
[9] L. Breiman, "Random forests," *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001.
[10] S. Hochreiter, J. Schmidhuber, "Long-short term memory," *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780, 1997.
[11] R. Singh, S. Srivastava, "Stock prediction using deep learning," *Multimed Tools And Application,* Vol. 76, No. 18, pp. 18569–18584, 2017.
[12] Z. Guo, H. Wang, J. Yang, et al, "A stock market forecasting model combining two-directional two-dimensional principal component analysis and radial basis function neural network," *Plos One*, Vol. 10, No. 4, pp. e0122385, 2015.