Research of Stock Price Prediction Based on PCA-LSTM Model

IOP Conf. Series: Materials Science and Engineering **790** (2020) 012109 doi:10.1088/1757-899X/790/1/012109

Research of Stock Price Prediction Based on PCA-LSTM Model

Yulian Wen, Peiguang Lin* and Xiushan Nie

School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China

*Corresponding author

Abstract. At present, there are some problems in domestic stock market, such as difficulty in extracting effective features and inaccuracy in stock price forecast. This paper proposes a stock price prediction model based on Principal Component Analysis (PCA) and Long Short-Term Memory (LSTM). Firstly, PCA is used to extract the principal components of the technical indicators affecting stock prices, so as to reduce the data correlation and realize data dimensionality reduction. Then, LSTM is used to model and predict the stock price. According to the experimental results of Pingan Bank, compared with the traditional stock price prediction models, the stock price prediction model based on PCA and LSTM can accurately predict the stock price fluctuation trend.
Keywords: Artificial Intelligence; PCA; LSTM; Stock price prediction; Deep learning.

1. Introduction

Stock market is an important part of national economic development. Forecasting stock price movements is important for governments, investors and investment institutions. Therefore, it attracts many scholars to conduct research. However, the price trend of the stock market may be influenced by political factors, macroeconomic factors, legal factors and etc., resulting in great uncertainty and volatility of the stock price, making it a major problem in research.

This paper proposes a deep learning model based on PCA-LSTM to predict stock price fluctuation. Firstly, the model uses PCA to extract the main components from a number of technical indicators that affect stock prices, to achieve data dimensionality reduction, reduce network training time, and improve model performance. Secondly, the stock trading information is based on time series, and LSTM has the potential to learn long observation sequences. So, this paper uses LSTM to predict stock price time series. This paper uses Pingan Bank stock trading information as a data set from January 4, 2016 to December 28, 2018, using Convolutional Neural Network (CNN) model, Multi-Layer Perceptron (MLP) model and Moving Average model as comparative experiments. The experimental results show that the combination of PCA method and LSTM model is better than the comparison models in prediction performance.

2. Related Work

Stock price time series prediction is the prediction behaviour of stock price fluctuation according to the historical data of stock price. The stock market is of great significance in the financial field. Therefore, the research on stock price prediction has attracted the attention of many researchers at home and abroad. Some researchers use the moving average analysis method to study the stock market price trend. For example, Aistis Raudys[1] proposed an optimal stock price smoothing weighting scheme based on the

negative weight moving average. However, the moving average has a certain lag. In recent years, with the extensive application of deep learning technology, many domestic and foreign researchers began to use deep learning to conduct stock prediction research. For example, Volodymyr Turchenko[2] and others proposed to take the stock price of Fiat company as the research object and used MLP neural network to make short-term prediction of stock price. However, the number of parameters required to use the MLP neural network is too large and the scalability is poor. Therefore, Avraam Tsantekidis[3] and others proposed a stock price forecast based on the CNN model. CNN realizes the local connection and weight sharing of neurons, retains important parameters, and reduces a large number of unimportant parameters. Compared with the MLP model, it achieves better learning results. However, CNN also has certain limitations. CNN focuses on spatial mapping and has certain advantages in processing image data. It is not fully applicable to learning time series. The LSTM model is a special type of structure of the RNN model, in which three control units of the forgetting gate, the input gate and the output gate are added. As the information enters the model, the control unit in the model will make judgments on the information, leaving the conforming information and discarding the non-conforming information. Based on this principle, LSTM can solve the problem of long sequence dependence in neural networks. Therefore, this paper proposes a stock price time series prediction method based on PCA-LSTM model.

3. PCA-LSTM Based Stock Price Time Series Prediction Model

This experiment adopts keras as the deep learning framework. LSTM is a deep learning model used to solve the problem of gradient disappearance in long sequences. In order to solve the problem of predicting the daily closing price of Pingan Bank according to the data of the first N days (the forecast range is 1), a deep learning model consisting of two layers of LSTM module is designed, and a dropout layer is set in the middle to avoid over-fitting. Finally, a Dense layer is set to output a specific number. The workflow of the model is shown below.

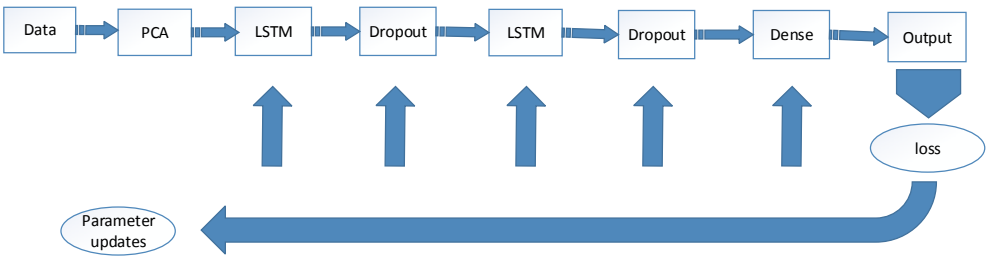

Figure 1. The workflow of model.

3.1. Data Pre-processing

In the research of stock price prediction, there are many factors that affect the stock price, and they may be correlated with each other and have different effects on the result. The PCA method can concentrate these factors affecting the stock price on several main components, so that these main components reflect as much information as possible in the original variables, eliminating data redundancy and achieving data dimensionality reduction.

3.2. Stock Price forecasting

Most data of stock market are time series data, and LSTM neural network has obvious advantages in processing time series information. Based on the RNN, LSTM adds memory cells to each neural unit in the hidden layer, making the memory information in the time series controllable. The information passes through several controllable gates (forgotten gates, input gates, and output gates) when passing between the various elements of the hidden layer. It can control the memory and forgetting degree of the previous information and the current information, so that LSTM has long-term memory function. In addition, LSTM has been successfully applied to many fields, such as image processing[4,5], speech recognition[6,7], handwriting recognition[8] and so on. Therefore, this paper uses LSTM neural network for stock price forecasting. Figure 2 shows the internal structure of LSTM.

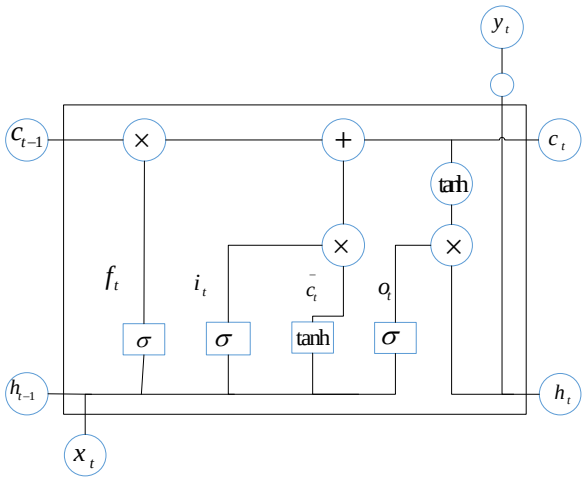

Figure 2. Internal structure of the LSTM.

LSTM has three gates to control the storage state, including the forget gate, the input gate and the output gate.

(1) The Forgotten Gate is used to determine the information that is discarded.

$$f_t = \sigma(w_f \times [h_{t-1}, x_t] + b_f) \tag{1}$$

σ refers to the activation function of Sigmoid, and w refers to the weight, and b refers to the offset. Sigmoid outputs a value between 0 and 1, and f_t determines how much information about the state of the cell can pass at the previous moment. 0 means no information is allowed to pass, and 1 means all information is allowed to pass.

(2) The input gate is used to determine the information that needs to be updated.

$$i_t = \sigma(w_i \times [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\bar{c}_t = \tanh(w_c \times [h_{t-1}, x_t] + b_c) \tag{3}$$

i_t refers to how much information needs to be updated by Sigmoid, 0 means not updated, and 1 means completely updated. $\bar{c}_t$ refers to the output of alternative content to be updated through tanh.

The information to be transmitted is determined by the forgetting gate and the input gate.

(3) The output gate determines the output information.

$$c_t = f_t * c_{t-1} + i_t * \bar{c}_t \tag{4}$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(c_t) \tag{6}$$

The cell state c_t at time t is the sum of the information retained by the cell state at time t-1 and the currently updated information. o_t refers to how much information needs to be output, and 0 means no information is output, and 1 means to output all the information; h_t refers to determines output part.

Dense layer

The Dense layer transforms a vector generated by the LSTM layer into a vector of a specified length.

$$S_d = dense(S_l, n) = (x^{(1)}, x^{(2)}, ..., x^{(n)}) \tag{7}$$

n represents the number of digits output through the Dense layer, n=1, which converts the vector of fixed length of the LSTM layer into a specific number; S_l represents a fixed-length vector generated by the LSTM phase; S_d represents a specific real number of the output.

4. Experiment

4.1. Data Set

The data set adopted in this paper is the stock information of Pingan Bank, including the stock trading information of the three years from January 4, 2016 to December 28, 2018 (731 trading days). The data set is divided into 60% training set, 20% verification set and 20% test set. The model is trained using the training set, and the hyperparameter of the model is adjusted using the validation set, and finally the performance of the model is tested using the test set. The six technical indicators that affect the stock price of Pingan Bank are Close Price, Open Price, High Price, Low Price , Turnover and Trading Volume. The data comes from RESSET financial research platform.

Figure 3. Pingan Bank stock price.

4.2. Data Preprocessing

There may be correlation between various technical indicators that affect the stock price. In this paper, SPSS statistical analysis software is used to conduct principal component analysis on the six technical indicators, including Close Price, Open Price, High Price, Low Price, Turnover and Trading Volume. The analysis result is shown as table 1. According to the following table, the Close Price is the first principal component, its accumulative percentage accounts for 85.296%, and the cumulative percentage of general variance is greater than or equal to 85% to determine the principal component. Therefore, the first principal component is extracted and used to predict the stock price.

Table 1. Results of principal component analysis.

	Initial eigenvalue			Extracting the sum of squared loads		
Ingredient	Total	Percentage of variance	Cumulative percentage	Total	Percentage of variance	Accumulation
1	5.118	85.296	85.296	5.118	85.296	85.296
2	.862	14.368	99.664			
3	.009	.142	99.806			
4	.008	.139	99.945			
5	.002	.041	99.986			
6	.001	.014	100.00			

4.3. Evaluation Indicators of the Model

In this paper, Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) are selected to quantitatively evaluate the performance of PCA-LSTM model and comparison model.

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{x}^{(i)}-x^{(i)})} \tag{8}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\bar{x}^{(i)} - x^{(i)}}{x^{(i)}}\right| \tag{9}$$

$\bar{x}^{(i)}$ refers to the forecast data of the stock price; $x^{(i)}$ refers to the real data of the stock price.
The smaller RMSE and MAPE, the smaller the error between the predicted and real data of the stock price, and the better the performance of model.

4.4. Experimental Results and Analysis

In order to compare the performance of the model, this experiment compares the PCA-LSTM model with CNN, MLP and Moving Average model, and uses Pingan Bank's test set data as the test object (from May 30, 2018 to December 28, 2018). The RMSE and MAPE are used as model evaluation indicators.

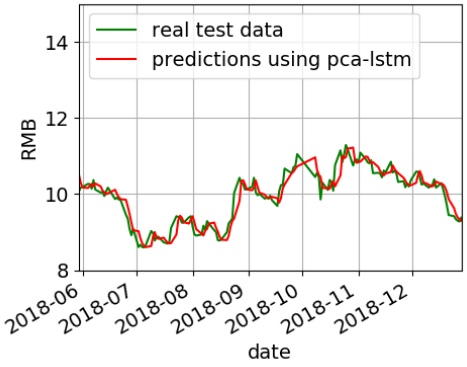

Figure 4. The prediction results of PCA-LSTM model

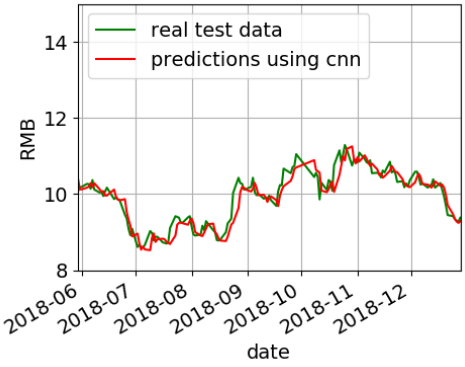

Figure 5. The prediction results of CNN model

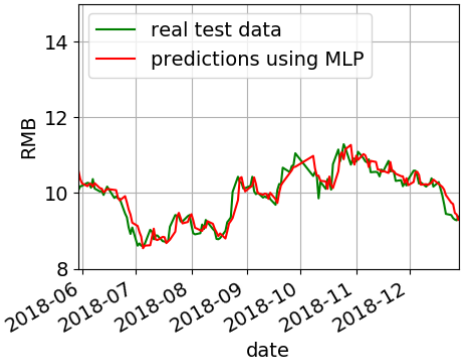

Figure 6. The prediction results of MLP model

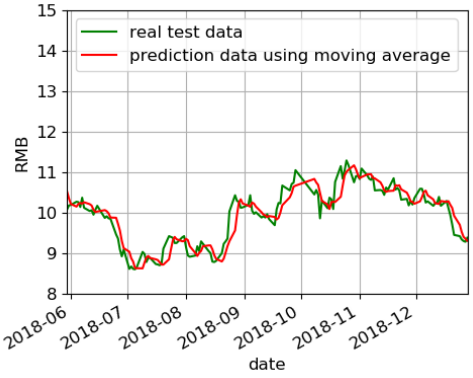

Figure 7. The prediction results of moving average model

Table 2. Comparing the stock price prediction results of models

	RMSE	MAPE
PCA-LSTM model	0.221	1.667%
CNN model	0.230	1.686%
MLP model	0.239	1.838%
Moving average model	0.248	1.943%

From the experimental results of figure 4 to figure 7, it can be seen that the green line represents the real data of the test set, and the red line represents the predicted data of each model. Compared with the other three comparison models, the PCA-LSTM model has a better fitting degree of prediction curve and real

value curve on the test set. As shown in table 2, the RMSE and MAPE of PCA-LSTM model are smaller than those of other three comparison models, proving that its prediction accuracy is higher than that of the other three comparison models.

5. Conclusion

In this paper, the research method of stock price time series prediction based on PCA-LSTM model is proposed. Taking the stock price of Pingan Bank as an example, PCA is used to extract features to obtain important technical indexes that affect stock price, and LSTM is used to predict stock price. The experimental data showed that the RMSE and MAPE of the PCA-LSTM model are 0.221 and 1.667% respectively. Compared with the CNN model, MLP model and Moving Average model, the RMSE and MAPE of the PCA-LSTM model are smaller, which prove that the PCA-LSTM model had better prediction performance. The following work will try to make larger data, further optimize the model and parameters, and improve the performance of the model.

References

[1] Aistis Raudys. Optimal negative weight moving average for stock price series smoothing [C] //CIFEr. London, UK: IEEE, 2014: 239-246

[2] Volodymyr Turchenko, Patrizia Beraldi, Francesco De Simone, et al. Short-term stock price prediction using MLP in moving simulation mode [C] //Proceedings of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems. Prague, Czech Republic: IEEE, 2011, 2: 666-671

[3] Avraam Tsantekidis, Nikolaos Passalis, Anastasios Tefas, et al. Forecasting Stock Prices from the Limit Order Book Using Convolutional Neural Networks [C] // 2017 IEEE 19th Conference on Business Informatics (CBI). Thessaloniki, Greece: IEEE, 2017, 1: 7-12

[4] Ankan Kumar Bhunia, Aishik Konwer, Ayan Kumar Bhunia, et al. Script identification in natural scene image and video frames using an attention based Convolutional-LSTM network [J]. Pattern Recognition, 2018, 85: 172-184

[5] Liu Jun, Wang Gang, Duan Lingyu, et al. Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks [J]. IEEE Transactions on Image Processing , 2017, 27(4): 1586-1599

[6] Yasufumi Moriya; Gareth J. F. Jones. LSTM Language Model Adaptation with Images and Titles for Multimedia Automatic Speech Recognition [C]// 2018 IEEE Spoken Language Technology Workshop (SLT). Athens, Greece, Greece: IEEE, 2018: 219-226

[7] Yi Jiangyan, Wen Zhengqi, Tao Jianhua, et al. CTC Regularized Model Adaptation for Improving LSTM RNN Based Multi-Accent Mandarin Speech Recognition [J]. Journal of Signal Processing Systems, 2018, 90(7): 985-997

[8] Olha Zubarieva, Ivan Deriuge, Vadym Holosko, et al. Space Balancing in Online Handwriting Recognition Postprocessing using Deep Bidirectional LSTM [C]// 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). Niagara Falls, NY, USA: IEEE,2018: 576-581