Regression analysis and prediction using LSTM model and machine learning methods

Fangbi Tan*

Zhongnan University of Economics and Law Wuhan, China

*Corresponding author: fangbitan@zuel.edu.cn

Abstract. In this paper, the LSTM model in deep learning is applied to regression analysis, and the LSTM model is used to solve the problems of nonlinearity and data interdependence in regression analysis, so as to improve the traditional regression analysis model. Through the actual modeling application experiment, on the one hand, the prediction accuracy of different model parameters is compared and analyzed, on the other hand, the effectiveness and practicability of LSTM model in multiple regression analysis and prediction are confirmed.

Keywords: LSTM, regression analysis, time series prediction.

1. Introduction

Regression analysis is one of the important application fields of machine learning methods. Regression analysis method can be used to establish the dependence relationship between variables based on the existing data and describe it with mathematical model [1]. In generally, if there is obvious linear correlation between variables, a simple linear regression model can be constructed and solved by using the least square method. However, the relationship between objective things and information in the real world often has nonlinear characteristics. For the data with local nonlinear characteristics, the local weighted linear regression method can be used for fitting. For the nonlinear data with strong volatility, such as stock trading information, the traditional linear regression method is often unable to well analyze its change characteristics and rules [2]. On the other hand, the deep learning model represented by various kinds of deep artificial neural network has better nonlinear approximation ability, and can often achieve better results in dealing with nonlinear regression analysis problems such as stock trading data prediction [3]. Among them, the LSTM (Long Short-Term Memory) model [4] can realize selective memory for series data, it has obvious advantages in feature recognition and learning of time-series data, so it has been widely used [5] [6]. However, at present, LSTM model is mostly used to solve the discriminative and generative problems in the natural language processing process, and is seldom used in regression analysis and prediction. This paper takes the regression prediction analysis of China Baoan stock trading data as the application object. Considering the strong volatility of stock trading data, the traditional statistical regression methods and models cannot deal with the nonlinear volatility of stocks well, resulting in low accuracy [7] [8]. On the other hand, the stock trading data itself belongs to the time series data, which is dependent on each other. In view of this, this paper tries to make use of the advantages of LSTM model in nonlinear fitting and time series data processing, and uses the stock historical trading information to establish the regression analysis and prediction method of trading data

based on LSTM model, so as to use the trading data of one trading day to predict the highest price information of the next trading day.

2. LSTM model

LSTM model is a traditional RNN (recurrent neural network) model [4]. RNN ensures the continuous transmission of data through internal multi loop, and constantly adjust the weight in the way of Backpropagation. When propagated to the Activation Function, the slope will become extremely large or extremely small, and exploding gradient or vanishing gradient will appear. This is the gradient exploding problem, or the gradient vanishing problem. In addition, the traditional cyclic neural network still has the problem of long-term dependence in long sentence processing. LSTM uses three gates structure to effectively solve the above problems, and can remember valuable information for a long time. Therefore, LSTM has been widely used in time series data learning, especially in natural language processing.

3. Regression analysis method based on LSTM model

This paper will use Python language to write and implement the regression analysis method of stock trading data based on LSTM model to complete the prediction analysis of stock trading data. The specific model construction steps are as follows:

3.1. Stock trading data acquisition.

This paper selects the stock trading data of China Baoan as the research object of regression analysis. China Baoan outstanding shares on June 25, 1991 in the Shenzhen stock exchange listed trading code 000009. This paper obtained the trading data of the stock from January 2, 1992 to December 13, 2020 by compiling a web crawler in Python language, totaling 6848 trading days. Each data contains 8 columns, among which the first 7 columns are characteristic information, including the opening price, the highest price, the lowest price, the closing price, Chg (%), the trading volume (board lot) and the transaction amount (ten thousand RMB) information of the stock on the trading day, and the eighth column is label data, that is, the highest price information of the stock on the next trading day. Data is stored and managed in Excel files, as shown in Fig.1.

	A	B	C	D	E	F	G	H
1	11.9	12	11.8	11.85	0.42	6608	783	11.85
2	11.8	11.85	11.7	11.75	-0.84	7355	864	11.85
3	11.7	11.85	11.65	11.75	0	1081	127	12
4	11.75	12	11.7	12	2.13	15216	1826	12
5	11.9	12	11.8	11.95	-0.42	8830	1055	12.25
6	12.2	12.25	12.1	12.2	2.09	1806	220	12.2
7	12.2	12.2	11.9	12.2	0	2590	316	12
8	12	12	11.7	11.95	-2.05	3106	371	11.8
9	11.8	11.8	11.5	11.55	-3.35	3988	461	11.65
10	11.65	11.65	11.1	11.15	-3.46	3183	355	11.6
11	11.2	11.6	11	11.55	3.59	3304	382	11.85
12	11.6	11.85	11.6	11.65	0.87	3542	413	11.65

Figure 1. Stock trading data

3.2. Data Standardization.

Because the input characteristic parameters of the model involve the information of stock trading price, fluctuation range, trading volume and turnover amount, and their value ranges are quite different. In order to avoid the influence of different dimensions and orders of magnitude on regression prediction analysis, it is necessary to normalize the data and unify the range of input characteristic parameters and prediction output values to 0 to 1, The data standardization methods used in this paper are as follows [10]:

$$y = \frac{x - mean}{max - min}$$

Where, x and y are the original value and standardized result of the variable in a certain book respectively. mean, max and min represent the mean, maximum and minimum value of the variable in all samples respectively.

3.3. Data Segmentation.

The data set obtained in this paper contains 6848 data samples. All the sample data are divided into training set and test set, and the ratio is 95:5, that is, 6505 training data and 343 test data.

3.4. Model Construction and Parameter Setting.

Based on Baidu's PaddlePaddle framework and Python language, this paper builds a deep learning model to complete the regression analysis and prediction of stock trading data. The parameter settings of the model are shown in Table 1. Considering that both input and output are discrete values in regression prediction analysis, the data dimensions of input layer and output layer are 1 dimension. The dimension of LSTM hidden layer is 4 times that of input layer. The output data of LSTM will go through the Max pooling operation of the pooling layer, and the Tanh function will be used for nonlinear processing. Finally, the output result is obtained by using the dense layer with output dimension of 1. In the training process, MSE loss function commonly used in regression model is used and optimized by Adam optimizer. The learning rate is set to 0.001.

Table 1. Model Parameters

parameter	value
Dim of Input Layer	1
Dim of Output Layer	1
hidden-dim for LSTM	4
Pooling	Max pooling
Activation Function	Tanh
Loss Function	MSE
Optimizer	Adam
Learning rate	0.001

In order to verify the influence of different parameter selection on the model training and prediction effect, this paper will compare and test the different values of other model parameters, including epoch (300 and 500), batch size (64, 128, 256 and 512). In addition, in order to optimize the fitting effect of the model on nonlinear data, this paper also attempts to add a dense layer with dimension 128 after the pooling layer. The comparison analysis is carried out on the fully connected layer with the activation function of Relu.

4. Results and discussion

Considering that the data set involved in this paper is not large, in order to ensure the convergence effect of training. First, the model training rounds are set to 300 rounds, and the batch size of each training is set to 128. The loss curves of the training process are shown in Fig.2(a). It is noted that when the model is trained to about 200 rounds, the loss value tends to be stable and the model converges. On the other hand, the application effect of the model in the test set composed of 343 pieces of data is shown in Fig.3 (a) and Table II. In general, the model has achieved good results in the test set, but there are still large errors in individual test sample data. The overall mean square error loss of the test data set is 0.001086, which is greater than the average loss of the last 10 batches in the training process, indicating that the model has a certain degree of over fitting problem.

Next, the training rounds of the model are increased to 500 rounds, and other parameters remain unchanged. The loss curve of the training process is shown in Fig.2 (b). It is noted that the model loss has converged before 200 rounds, and the average loss value and test loss of the last 10 batches are larger than that of 300 rounds. Obviously, more training rounds cannot improve the actual application effect of the model. Therefore, in the subsequent experiments, the model would always remain unchanged at 300 rounds. Next, the network structure of the model is changed, and a FC full connection layer is added after the maximum pooling layer. The number of nodes is 128, and the activation function is the Relu function. The training loss curve of the model is shown in Fig.2 (c). The convergence speed of model training can be accelerated obviously after adding the full connection layer, and it has been converged within 50 rounds. In addition, it can be seen from the results in Table II that the training accuracy of the model is not significantly improved after adding the full connection layer (the average loss of the last 10 batches is 0.000238), but the application effect in the test set is better. In Fig.3 (c), it can be seen that the fitting effect of the model on the test data set is excellent after adding the full connection layer, most of the data are accurately predicted. The reason may be that the application of the full connection layer and its activation function further improves the recognition ability of the model to the nonlinear characteristics in the stock exchange information, thus improving the overall prediction effect of the model.

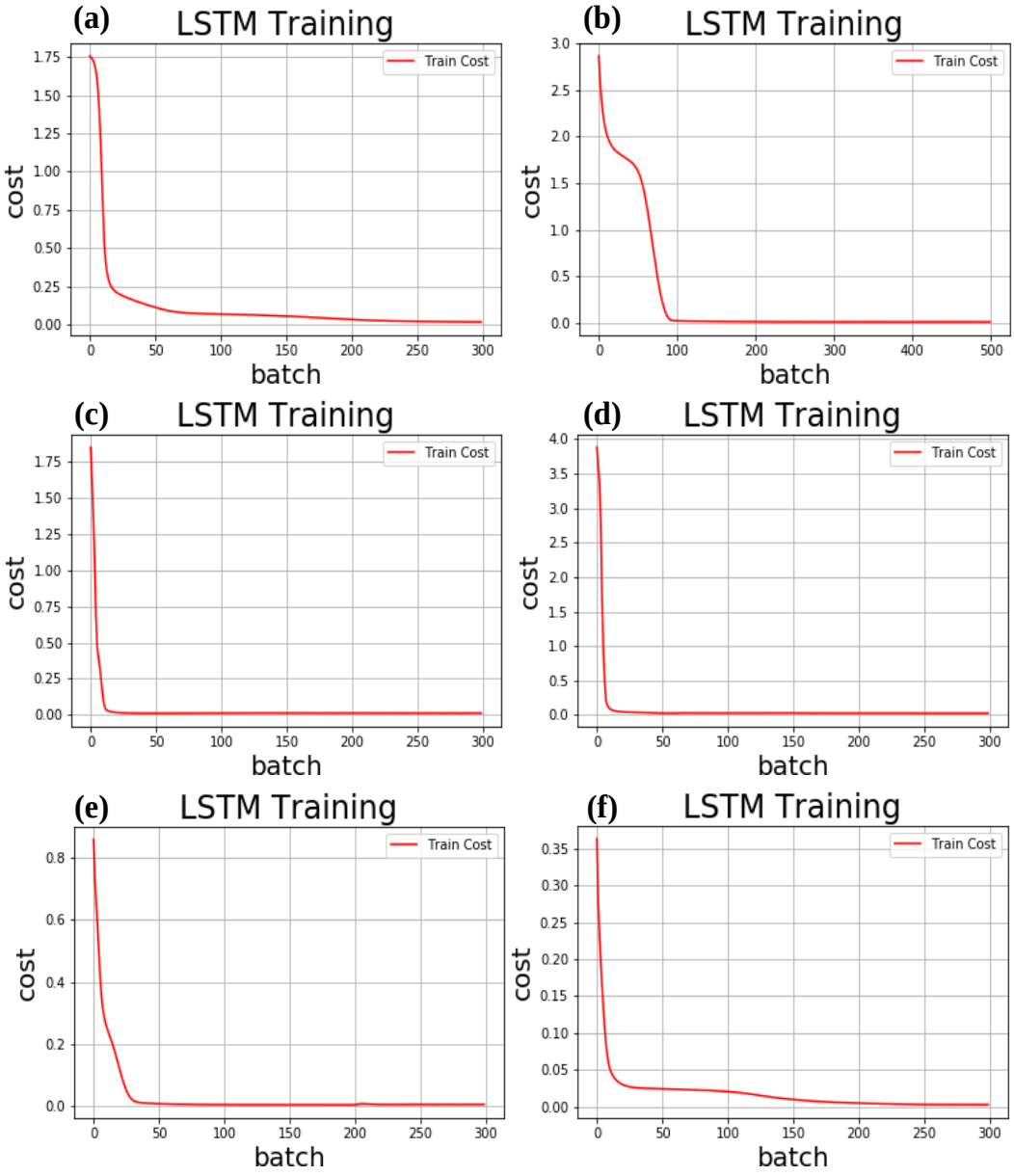

Figure 2. Loss curves obtained in training

Table 2. Performance in MSE for training and testing

Model	Performance in MSE	
	Training: last 10 batches	Testing
Epoch=300	0.000248	0.001086
Epoch=500	0.000329	0.001138
fc	0.000238	0.000788
fc batch_size=64	0.000074	0.001416
fc batch_size=256	0.000302	0.000853
fc batch_size=512	0.000169	0.000752

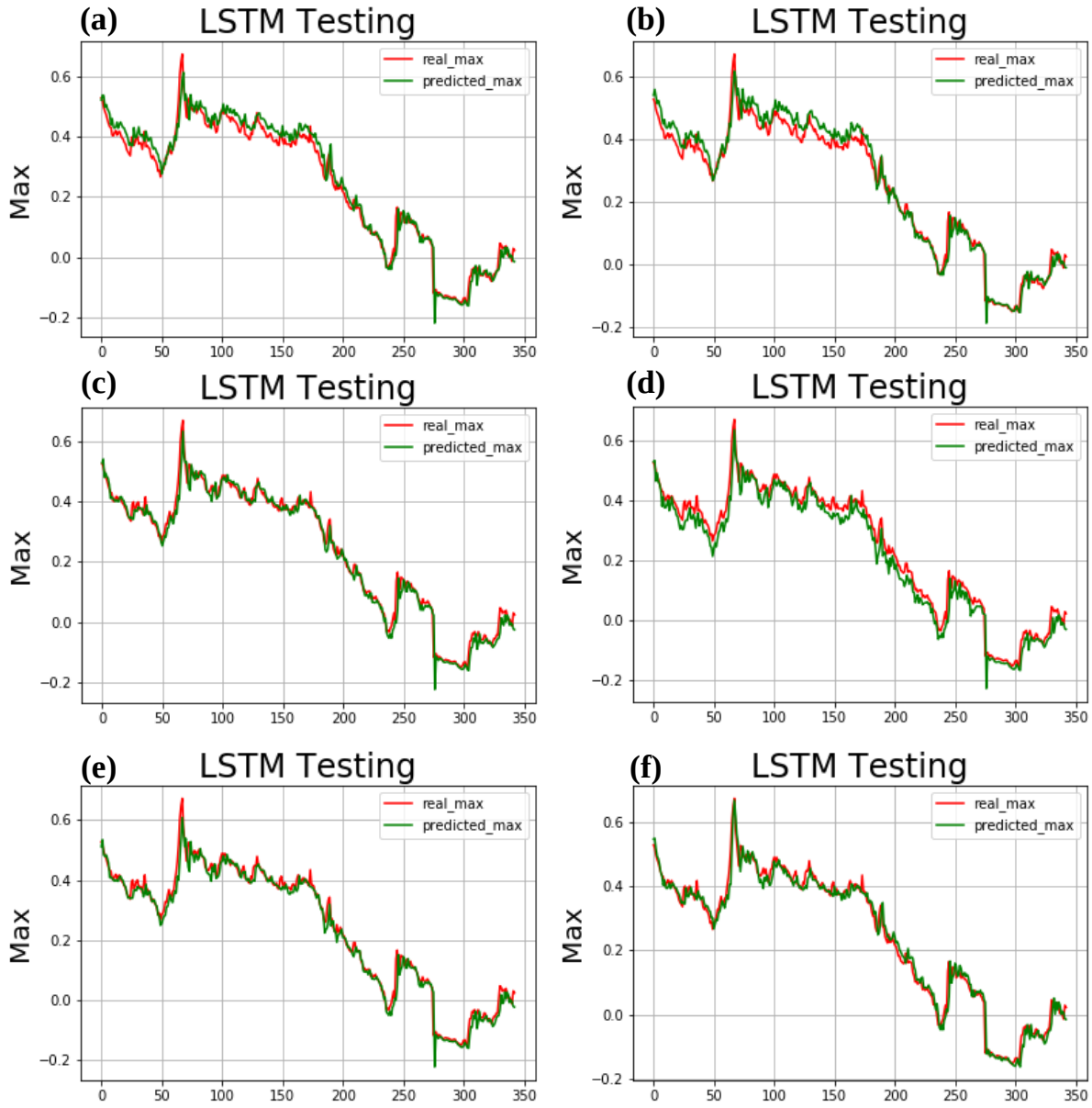

Figure 3. Model testing results

Finally, this paper analyzes the influence of training batch size on model training and test results. When the training batch size changes to 64, 256 and 512, the change curves of model training loss are shown in Fig. 2 (d) (e) and (f), respectively. It is noted that the size of training batch has a certain degree of influence on the convergence speed of the model. The smaller the size of training batch is, the faster the convergence speed of the model is. According to the results shown in Table II, the change of model training batch size also has a significant effect on the training performance. When the training batch size is reduced to 64, the model training accuracy is significantly improved (the average loss of the last 10 batches is as low as 0.000074), but the test accuracy is the worst under all parameters (as shown in Table 2 and figure 3 (d)), It can be seen that the model has a serious over fitting problem in this case. On the other hand, when the model training batch size increases, the training accuracy changes to a certain extent, but the test accuracy changes relatively little. Overall, with the inclusion of a full connectivity layer, the model performs well on the test data set for the larger training batches (for the scenarios with batch size>=128), as shown in Fig. 3 (c), (e), and (f).

5. Conclusion

In this paper, deep learning theory is applied to construct regression analysis and prediction method based on LSTM model, and regression prediction is made on the stock trading data of Baoan in China. The results show that LSTM model gives full play to its advantages in the analysis of time series data, and can well deal with the nonlinear fluctuation characteristics of stock trading data, and the overall application effect of the model is good. In addition, the experimental results of this paper also show that after adding the full connection layer to the LSTM model, it can better fit the nonlinear change time series data, and the simulated training batch size will also have a certain impact on the application effect of the model. The application results of this paper further expand the application range of the deep neural network model represented by LSTM model, and also provide a new solution for complex multivariate nonlinear regression analysis and prediction.

Acknowledgments

The authors would like to appreciate the supports for this study from the National Natural Science Foundation of Hubei Province (no. 2020CFB134) and the Fundamental Research Funds for the Central Universities, Zhongnan University of Economics and Law (no. 2722020PY041).

References

[1] HTS Saraçli. “Joinpoint Regression Analysis and an Application on Istanbul Stock-Exchange,” Alphanumeric Journal, vol. 2, no. 2, pp. 70 - 75, 2014.
[2] Chiang T C, Li J, Lin T. “Empirical investigation of herding behavior in Chinese stock markets: Evidence from quantile regression analysis,” Global Finance Journal, vol. 21, no.1, pp. 111-124, 2010.
[3] Purey P, Patidar A. “Stock Market Close Price Prediction Using Neural Network and Regression Analysis,” International Journal of Computer Sciences and Engineering, vol. 6, no. 8, pp. 266-271 ,2018.
[4] Hochreiter S, Schmidhuber J. “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735 – 1780, 1997.
[5] Jing N, Wu Z , Wang H. “A Hybrid Model Integrating Deep Learning with Investor Sentiment Analysis for Stock Price Prediction,” Expert Systems with Applications, vol. 178, no. 3, pp. 115019, 2021.
[6] Jin Z, Yang Y, Liu Y. “Stock closing price prediction based on sentiment analysis and LSTM,” Neural Computing and Applications, vol. 32, no. 3, 2020.
[7] Yu P, Yan X . “Stock price prediction based on deep neural networks,” Neural Computing and Applications, vol. 32, no. 6, pp. 1609 - 1628, 2020.
[8] Garcia-Fuentes P A, Ferreira G, Harrison R W, et al. “Consumer Confidence in the Food System, Media Coverage and Stock Prices of Food Companies: A Regression Analysis,” Agricultural and Applied Economics Association, 2010 Annual Meeting, Denver, Colorado, July 25-27, 2010.
[9] Liang X, Ge Z, Sun L, et al. “LSTM with Wavelet Transform Based Data Preprocessing for Stock Price Prediction,” Mathematical Problems in Engineering, pp. 1-8, 2019.
[10] Zhuge Q, Xu L, Zhang G. “LSTM neural network with emotional analysis for prediction of stock price,” Engineering Letters, vol. 25, no. 2, pp. 167 - 175, 2017.