Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230084

Stock prediction using LSTM model

Junyan Xiao

Huaer zizhu academy No.155Tanjiatang Road, minghang, Shanghai, China.

xiaojunyan318619@qq.com

Abstract. With the development of the times, investors are increasingly in demand for stock price forecasting. However, stock price fluctuations are full of uncertainty, making traditional machine learning algorithms more erroneous in long-term forecasting. Based on the LSTM model, this paper uses Tushare to obtain the historical price of stocks, and the optimal structure and best training parameters of the LSTM model in stock price prediction are determined experimentally. The prediction accuracy of the LSTM model was evaluated by MAE, and the best result was 69.15, which achieved accurate prediction of stock prices. Compared with the traditional SVR model and the ARMA model, the prediction results of LSTM are more in line with the actual value, and the prediction accuracy of the algorithm is higher.

Keywords: stock prediction, data mining, LSTM.

1. Introduction

With the passage of time, the level of science and technology is constantly improving, and more and more investors have extremely high demand for stock forecasting, and now the use of big data analysis technology for stock market prediction has become research-worthy. Nowadays, although the return on investing in stocks is high, there are investment risks, stock price fluctuations are full of uncertainties and the causes of investment risks are unclear, and people are also very concerned about stock price trends[1]. So now there are many well-known experts looking for a way to accurately predict stock prices, and they are doing research through a series of neural network models and helping to predict subsequent stock prices.

Traditional machine information models are widely used in stock forecasting. Tian Xiang et al. [2] use SVR (vector regression method) to construct a short-term forecast model of stock index, and propose a nonlinear time series forecasting model. Feng Pan et al. [3] used the ARMA (Autoregressive Moving Average Model) to predict the opening price of stocks, which confirmed the effect of the ARMA model on financial time series data. Wang Shuai et al. [4] verified the effectiveness of the support vector machine (SVM) model in the trend prediction of the CSI 300 Index.

However, traditional models have large errors in long-term stock predictions, and many scholars apply neural network models to stock price predictions. Liu [5] The accuracy predicted using the B-P neural network and the gray G ARCH model is significantly better than that of the G ARCH model. Deng[6] proposed the DAE-BP model to reduce the D AE dimensionality of stocks first, and then use the B P neural network to predict the stock price, and achieved good prediction results.

DOI: 10.54254/2755-2721/8/20230084

However, due to its too simple structure, traditional BP neural networks have problems such as overfitting, convergence to local extremes, gradient disappearance or gradient explosion. In 1997, Hochreiter and Schmidhuber [7] proposed the LSTM model to solve the problem of gradient explosion in traditional RNN models. Its excellent performance on time series and selection memory are very suitable for highly random stock price prediction problems.

This paper uses the ability of long short-term memory neural network (LSTM) to more effectively obtain the time series characteristics of stock historical trading data, establishes the LSTM neural network prediction model, and compares the prediction results with SVR and ARMA models, highlighting the unique advantages of LSTM prediction model in capturing the time series structure of data[8-10].

2. Methods

The LSTM is a specific recurrent neural network (RNN) consisting of a memory cell and three logical control units: Input Gate, Output gate, and Forget Gate, as shown in Figure 1.

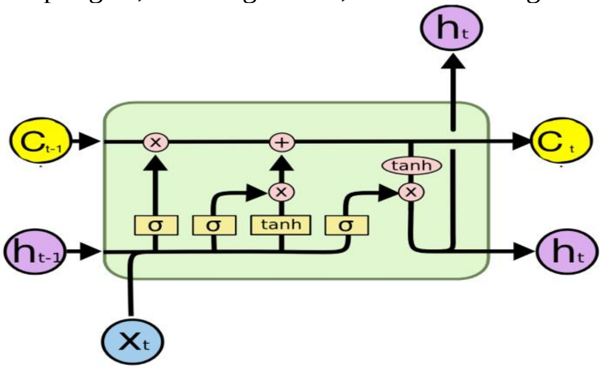

Figure 1. LSTM structure diagram.

Where, the input gate is used to remember the present information; The forgetting door is responsible for controlling what information is discarded, and the output door determines the final output information. The input of each gate is the hidden state of the previous moment and the input of the current moment, and the formula for each gate is defined as follows:$h_{t-1}x_t$

Input Gate:

$$I_t = \sigma(x_t W_{xI} + h_{t-1} W_{hI} + b_I) \tag{1}$$

Forget Gate:

$$F_t = \sigma(x_t W_{xF} + h_{t-1} W_{hF} + b_F) \tag{2}$$

Output Gate：

$$O_t = \sigma(x_t W_{xO} + h_{t-1} W_{hO} + b_O) \tag{3}$$

where is the sigmoid activation function, and is the argument for each gate.$\sigma W_x W_h b$

Memory cells update the cell state based on the calculation results of the input gate and the forgotten gate. It first counts candidate memory cells:

$$\tilde{C}_t = tanh(x_t W_{xC} + h_{t-1} W_{hC} + b_O) \tag{4}$$

Calculate the current cell state using the following equation:

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t \tag{5}$$

where $\odot$ is point-by-point multiplication.

In addition, the output gate can be used to calculate the hidden state of the current memory cell:

$$h_t = O_t \odot tanh(C_t) \tag{6}$$

Using the above formula, LSTM can delete or add information to the cell state, and through the fine mobilization of each gate structure, effective hidden information can be selectively learned from complex time series data.

DOI: 10.54254/2755-2721/8/20230084

The magnitude of a stock's decline is caused by the high factor, including the opening and closing prices. Previous stock price prediction studies only trained and tested prices as series data through models. In order to consider more factors of stock prices, the LSTM model is used to predict stock prices, which has good time series data, and this paper focuses on time series trend prediction in combination with LSTM technology to make the prediction results of stock prices more accurate.

3. Results

3.1. Experimental environment

This paper uses the Pytorch open-source machine learning library to build the LSTM model. Compared with other frameworks, pytorch has the advantages of simple syntax and high computational efficiency. CUDA 10.3 deep learning framework is used to accelerate model operation. The hardware environment for all experiments is NVIDIA GEForce RTX 3060 graphics card, Intel i7-11800 CPU, and 32G memory, which ensures the efficiency of data reading and model operation.

3.2. Dataset

This article uses Tushare to obtain the historical price of stocks and process the data using the Pandas toolkit. A company's historical transaction records from 1993 to 2017, totaling 5,000 days, were used as experimental datasets. At the same time, at a ratio of 7:3, the dataset is divided into a training set and a test set to verify the generalization of the model. The model takes characteristics such as closing price, opening price, highest price, lowest price, trading volume, transaction value and other characteristics as the model input, and the highest price of the next day is the output. Considering that different data needs to be used as model input features, and the magnitude of each feature data is different, it is easy to reduce the model training speed and affect the performance of the trained model. Therefore, Equation 7 is used to normalize the characteristic data into [-1,1].

$$x' = \frac{x - \bar{x}}{\sigma} \tag{7}$$

where and are the mean and standard deviation of the data, respectively.$\bar{x}\sigma$

3.3. Model training and result analysis

In order to evaluate the performance of the trained model, this paper takes mean squared error (MSE) as the loss function of the training process, and takes the mean absolute error (MAE) as the final evaluation index of the model, and the formula is as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{8}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{9}$$

where is the stock forecast value on day i, which is the corresponding sample value$\hat{y}_i y_i$

At the same time, Adam optimization algorithm is used to train the model [8], which has the advantages of high computing efficiency and small memory occupation. Table 1 shows the parameter settings for this algorithm.

Table 1. Model training parameters.

Parameters	Value
Max step	10000
Dropout keep probability	0.9
The initial learning rate	0.01
The momentum value	0.9

In addition, in order to improve the training efficiency, a large learning rate is used at the beginning, and the learning rate is adjusted to 0.5 times for every 5 epochs. At the same time, in order to prevent the

DOI: 10.54254/2755-2721/8/20230084

model from overfitting, an early stop mechanism is added during the training process, and when the model training for 5 consecutive epochs does not improve the prediction effect of the validation set, the model is considered to have converged, that is, the training is stopped.

The impact of model structure on prediction performance. Firstly, the model structure, including the number of LSTM layers and the number of neurons in each layer, is tested on the model performance. The price of the following day is predicted using the stock characteristics of the previous 30 days, and the results are shown in Table 2.

Table 2. Influence of model structure on prediction accuracy.

Number of Neurous	1 LSTM layer		2 LSTM layer		3 LSTM layer	
	MAE	Training steps	MAE	Training steps	MEA	Training steps
8	5365.12	10000	3145.26	10000	1286.41	10000
16	2568.37	10000	1347.61	10000	668.15	10000
32	1179.39	7562	321.45	8132	352.21	6563
64	526.24	6452	236.26	7014	96.49	5254
128	305.87	5145	69.15	5238	136.77	4931
160	257.77	3186	107.25	4275	205.19	3396
192	276.11	1862	283.61	2315	292.83	2564
256	173.28	1459	415.38	1569	464.17	1678
512	396.25	1170	475.21	1260	562.44	1325

It can be seen from the experimental results that the model achieves the best effect when using 2 layers of LSTM, each layer with 128 neurons each. SIMILAR RESULTS HAVE BEEN ACHIEVED IN STOCK DATA FROM OTHER COMPANIES, SUCH AS APPLE, AAON, AXDX, AND OTHERS. The consideration is that when the number of parameters is too low, the model fitting ability is poor, and it is difficult to obtain better results within the number of iteration steps. With the increase of the number of layers and neurons, the model fitting ability improves and the MAE gradually increases. However, when there are too many parameters, it is easy to overfit, and the model converges quickly, but the MAE on the test set is low.

The impact of training data on prediction performance. The impact of the training data on model performance was then tested. In general, the model training process is susceptible to batch size; At the same time, the length of the historical data used is an important factor that affects the prediction effect of time series data. Therefore, this paper uses the optimal network structure in the previous section to focus on testing the impact of these two data on the final performance. Set the batch size to 4, 8, 16, 64, 128, 256, and investigate the impact of using the data of the first 1, 5, 10, 20, 30, 60, and 90 days to predict the prediction performance, and the results are shown in Table 3.

Table 3. Influence of training data on prediction accuracy.

Batch size	Timestep						
	1	5	10	20	30	60	90
4	183.24	198.33	202.10	215.33	266.25	282.13	202.31
8	169.51	211.26	175.86	196.75	251.31	241.34	219.42
16	217.72	136.42	155.37	184.42	172.44	175.38	183.52
64	206.98	119.26	102.47	98.35	69.15	80.56	77.25
128	274.16	96.32	92.55	86.30	76.28	72.33	76.51
256	331.56	88.60	79.83	127.26	92.46	86.25	89.38

DOI: 10.54254/2755-2721/8/20230084

It can be seen that when the batch size of the training process is 64, the model is best obtained when the data from the previous 30 days is used for prediction. Since the larger the batch size, the easier it is for the model to learn the feature distribution of the sample data; However, too large batchsize will make the model learn too much redundant information, resulting in poor performance. For the length of historical data, to a certain extent, the longer it is, the more conducive to improving the prediction accuracy; However, on the one hand, too long historical data will lead to reduced computational efficiency, on the other hand, the model may degrade performance due to too much redundant historical data.

3.4. Comparative experiments

Finally, the prediction results of the LSTM model are compared with traditional machine learning models such as SVR and ARMA, and the results are shown in the figure 2, 3 and 4.

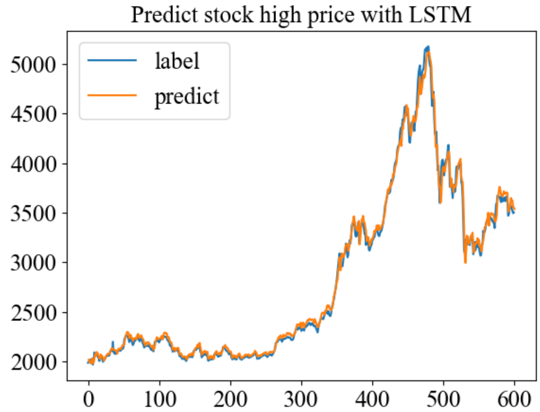

Figure 2. LSTM stock forecast results.

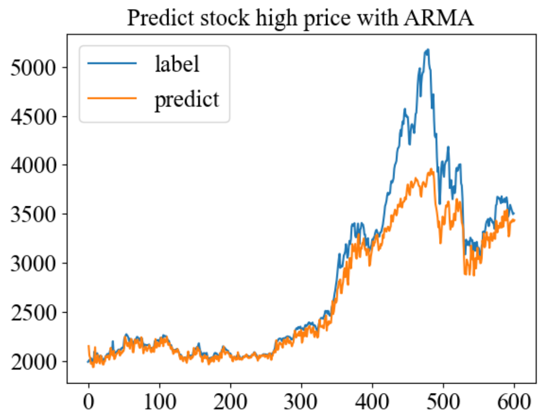

Figure 3. ARMA stock forecast results.

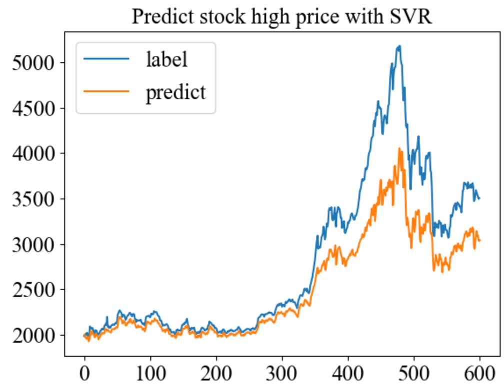

Figure 4. SVR stock forecast results.

It can be seen that traditional machine learning models are difficult to cope with drastic changes in stock prices; The prediction errors of the ARMA model and SVR model are 2 96.34 and 317.23, respectively, which are significantly higher than those of the LSTM model, indicating that LSTM is more suitable for stock prediction tasks.

4. Conclusion

Due to the many factors affected by stock prices and the characteristics of large fluctuations, some traditional forecasting methods are not highly accurate. Based on the deep learning algorithm, this paper obtains stock data based on Tushare, and applies the LSTM neural network model to stock price prediction. The optimal model structure was determined experimentally to use two layers of LSTMs with 128 neurons each; When using batch size 64 and timestep 30 at the same time, the best prediction accuracy can be obtained. Compared with traditional machine learning algorithms ARMA and SVR models, LSTM has a higher prediction accuracy.

References

[1] Pahwa, N., Khalfay, N., Soni, V., & Vora, D. (2017). Stock prediction using machine learning a review paper. International Journal of Computer Applications, 163(5), 36-43.
[2] TIAN Xiang,DENG Feiqi. Application of Accurate Online Support Vector Regression in Stock Index Forecasting[J]. Computer Engineering,2005,31(22):18-20
[3] FENG Pan,CAO Xianbing. An Empirical Study on Stock Price Analysis and Prediction Based on ARMA Model[J]. Mathematics in Practice and Theory, 2011(22): 84-90
[4] Wang Shuai，Shang Wei． Forecasting directionof china security index 300 movement with least squares support vector machine[J]．Procedia Computer Science， 2014， 31:869 － 874．
[5] Liu, E. . (2021). Comparison of stock price prediction ability based on GARCH and BP_ANN. 2021 2nd International Conference on Computing and Data Science (CDS).
[6] Deng, X. , Liang, W. , & Huang, N. . (2019). Stock prediction research based on dae-bp neural network. Computer Engineering and Applications.
[7] Hochreiter, S.; Schmidhuber, J. Long short-term memory[J]. Neural Comput.1997, 9, 1735–1780.
[8] Kingma, D. P., & Ba, J. (2015, January). Adam: A Method for Stochastic Optimization. In ICLR (Poster).
[9] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).
[10] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).