

PAPER • OPEN ACCESS

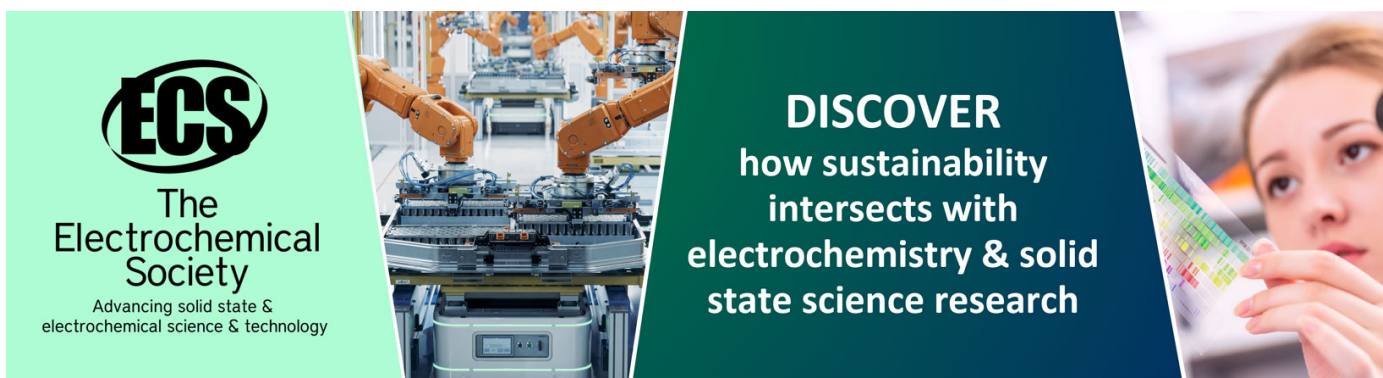
Research on Stock Prediction Model Based on Deep Learning

To cite this article: Haiying Zhang *et al* 2020 *J. Phys.: Conf. Ser.* **1549** 022124

View the [article online](#) for updates and enhancements.

You may also like

- [Offline and online LSTM networks for respiratory motion prediction in MR-guided radiotherapy](#)
Elia Lombardo, Moritz Rabe, Yuqing Xiong et al.
- [State of Health Prediction of Lithium-Ion Batteries Using Combined Machine Learning Model Based on Nonlinear Constraint Optimization](#)
Yawen Liang, Shunli Wang, Yongcun Fan et al.
- [Performance comparison of solar radiation forecasting between WRF and LSTM in Gifu, Japan](#)
Jose Manuel Soares de Araujo



ECS
The
Electrochemical
Society
Advancing solid state &
electrochemical science & technology

DISCOVER
how sustainability
intersects with
electrochemistry & solid
state science research

Research on Stock Prediction Model Based on Deep Learning

Haiying Zhang¹, Qiaomei Liang¹, Siheng Li¹, Rongqi Wang^{2,*}, Qingqiang Wu¹

¹School of Informatics Xiamen University Xiamen, China

²Xiamen Rural Commercial Bank Xiamen, China

*Corresponding author e-mail: wangrongqi@fjnx.com.cn

Abstract. In order to obtain stable returns, this paper aims to establish a quantitative model with higher prediction accuracy. According to the time series characteristics of financial data, the prediction model with financial time series data is constructed by using the time-memory sequence model LSTM, which is applied to the representative SSE 50 series stocks. And based on this, the LSTM model with Encoder-Decoder mode and a hybridized framework of LSTM with CNN are built to improve the original model. Feature extraction is performed on the input data by using CNN, and then as an input to the LSTM, the extracted features are used for sequence prediction with the LSTM model.

1. Introduction

Although stocks are highly real-time and unstable, the large amount of historical data, as an objective reaction of the stock market, necessarily indicates the future trend of the stock market to a certain extent. Experts, scholars and investment institutions aim to obtain a certain degree of prediction on the stock market, and get higher benefits by analyzing these data.

To this end, numerous experts and scholars at home and abroad have proposed a variety of methods to predict the changing trend of the stock market and build a stock forecasting model. Statistical and econometric models such as multiple regression and exponential smoothing [1]. However, the extremely complex nonlinear characteristics of the stock market make the limitations of the traditional statistical model in dealing with nonlinear problems, and the prediction effect is not satisfactory.

In recent years, the strong ability of machine learning in dealing with nonlinear problems has made it stand out in financial forecasting and has been widely used in stock market forecasting. However, traditional machine learning models, such as SVM, KNN, etc., are prone to fall into local optimum, and the shortcomings of the time characteristics of financial data are gradually exposed.

Deep learning is a further improvement in traditional machine learning [2-7]. Based on machine learning, it combines low-level features to form more abstract high-level attribute representation categories and features to discover distributed feature representations of data. The neural network, which simulates the mechanism of the human brain to interpret the data, also makes it a powerful advantage in dealing with the effects of multiple complex factors, highly unstable, and random nonlinear problems. It overcomes the shortcomings of traditional machine learning.

But at the same time, the classical BP neural network prediction model does not necessarily apply to sequence prediction problems with temporal characteristics, and suffers from problems such as gradient disappearance. The Long Short Term Memory Network (LSTM) further solves these problems, so this study constructs a stock prediction model based on LSTM.



2. Methodology

2.1. LSTM

The Long and Short Time Memory Neural Network (LSTM) was proposed by Hochreiter and Schmidhuber in 1997. LSTM is a special RNN model designed to avoid long-term dependency problems. Its specific structure is shown in Figure 1.

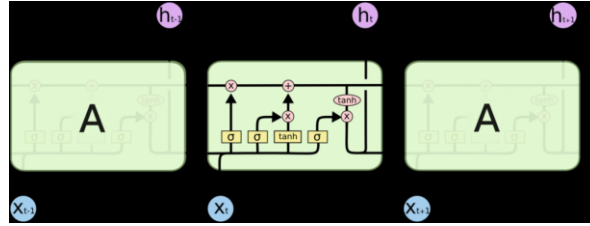


Figure 1. Main structure of LSTM

As can be seen from the figure, in addition to the tanh layer, the hidden layer of the LSTM has three special "gate" structures through which the LSTM forgets or enhances the information. Specifically, the "gate" structure is a combination of sigmoid and bitwise multiplication. It can be understood that when the door is closed, that is, when the sigmoid function outputs 0, the information cannot pass; otherwise, when the door is opened, the sigmoid function outputs 1, the information is passed. This unit includes four elements: input gate (i), forget gate (f), output gate (o), and memory cell (c). The specific calculation formula for LSTM is as follows:

$$i_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_i) \quad (1)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

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

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

The specific structure of the three gates is shown in Figure 2. The cell in the middle of the figure refers to the state of the neuron, that is, the memory information, and the state information is recorded and saved by a parameter.

The Input Gate and the Output Gate are used to receive parameters and output parameters, respectively. The Forget Gate is the key to the LSTM memory function. It can choose whether the correction of the parameters is forgotten. If you choose to forget, the status information transmitted by the previous neuron is completely forgotten.

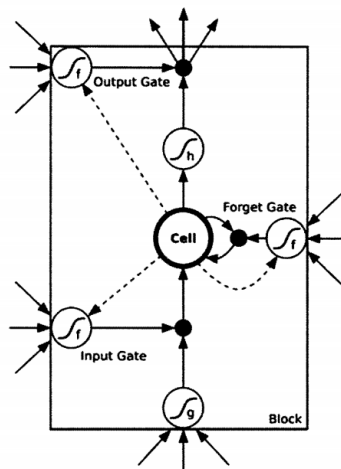


Figure 2. Specific structure of LSTM unit

Between the same layer of the LSTM neural network, the previous unit transmits the state information of its neurons to the next unit, and each unit also produces an output. In some prediction problems, it is for the next unit. The prediction information is usually compared and corrected with the correct information, and the output at the last time step in the final unit is the prediction result we need.

2.2. Encoder-Decoder LSTM

Sequence prediction problems are often the next value in the predicted sequence data or the class label of the output input sequence. This is typically constructed as an input time step to an output time step (one-to-one) or a sequence prediction problem with multiple input time steps to one output-to-one type(many-to-one).

There is a more complex sequence prediction problem, where the input data is a sequence and the output prediction data is also a sequence. Such a problem is called a sequence-to-sequence prediction problem, also known as the seq2seq problem. It is a problem in the form of many-to-many. One common method for solving seq2seq is the Encoder-Decoder model, which is the encoding-decoding model. Due to the advantages of the LSTM model in the prediction of sequence data, the encoder and decoder here are composed of LSTM, so it is called Encoder-Decoder LSTM model. This model consists of two components: the encoder LSTM is used to read the data of the input sequence and encode it to obtain one code vector, and the other LSTM is used as the decoder. The previous step is coded as input, and the output is predicted. sequence. The structure of the Encoder-Decoder model is shown in Figure 3.

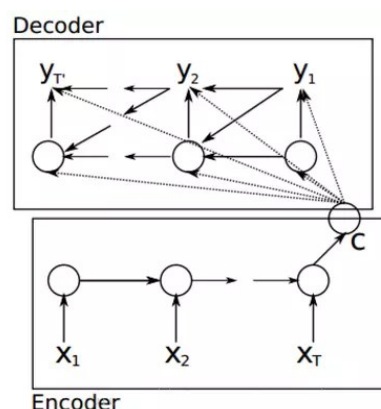


Figure 3. The structure of the Encoder-Decoder model

The financial time series prediction problem to be carried out in this experiment seems to be not too consistent with the Encoder-Decoder model mentioned above, but the experiment does not expect the sequence directly output by the decoder as our prediction result, but hopes to mine more hidden information in the financial time series data and its time characteristics. Then re-enter the sequence data after the encoding-decoding operation into the LSTM model for stock prediction, in anticipation of achieving more prediction than simply using the LSTM model.

2.3. LRCN

Inputs with spatial structure (such as images) cannot be accurately modeled using the classical LSTM model, and convolutional neural networks can solve this problem better, so they are widely used in the image field. Some researchers have compared financial stock data with images, apply convolutional neural networks to stock price forecasts and achieve good results. However, CNN lacks the same processing as the traditional BP neural network for the temporal feature of sequence data, and LSTM is good at it. Therefore, by combining CNN with LSTM model, better results are obtained from the prediction of the data.

Long-term Recurrent Convolutional Networks (LRCN) consists of a convolutional (CNN) layer and a long-short-time memory (LSTM) layer, where the CNN layer is used for feature extraction of input data and then input into LSTM. Its model structure is shown in Figure 4.

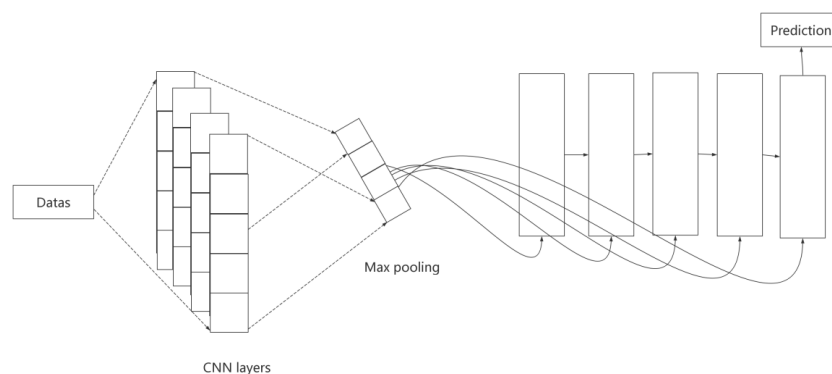


Figure 4. Main model structure diagram of LRCN

In this experiment, the CNN-LSTM model is applied to the prediction of financial time series data, and it is expected that it can fully exploit the characteristics of financial time series data to obtain better prediction ability than the LSTM model.

3. Model

The whole process of stock prediction model based on LSTM neural network is composed of five modules: data selection, data preprocessing, model building, evaluation index and model testing. The overall process is shown in Figure 5.

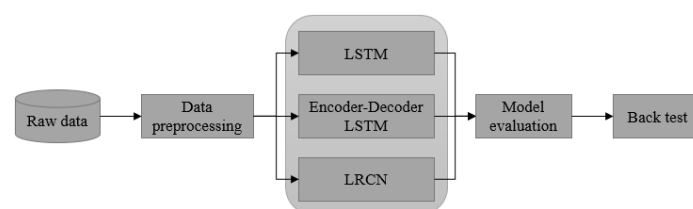


Figure 5. Overall flow chart

Three models are used in this experiment: LSTM, Encoder-Decoder LSTM and LRCN, and their respective structures are shown in Figure 6.

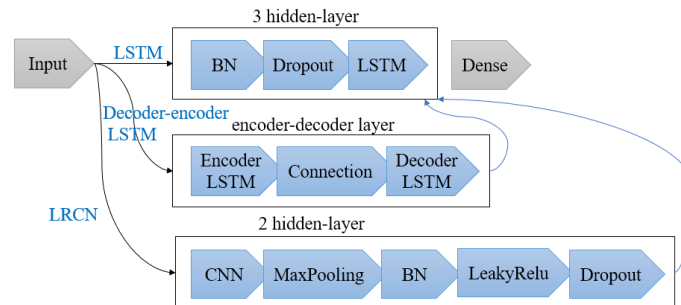


Figure 6. Model structure

The LSTM model uses a three-layer stacked structure with 32 neurons per layer and a BatchNormalization layer and a Dropout layer after each LSTM layer (the Dropout layer has a parameter value of 0.5) to prevent over-fitting of the model and enhance the generalization ability of the model. Finally, the result of the classification is output through a fully connected layer and softmax as the activation function. At the same time, the model's optimizer uses the "adam" optimizer and incorporates learning rate attenuation techniques to help the model achieve better results.

The Encoder-Decoder LSTM model can be divided into two parts. The first part is the classic Encoder-Decoder form, in which the layer and the decoding layer are composed of LSTM. The connection layer is used between the coding layer and the decoding layer. The second part of the LSTM prediction layer is the same as the classical LSTM prediction model, and its parameter configuration is basically the same as above. However, the LSTM layer here uses only one layer for prediction. Experiments show that if the number of layers in the LSTM layer is increased, the over-fitting phenomenon will appear and become serious. Therefore, the LSTM here uses only one layer for prediction.

The LRCN model is composed of two CNN layers and one LSTM layer. The input data first passes through the CNN layer to achieve the feature extraction result, and then the extracted effective feature sequence is input into the LSTM layer for sequence prediction. Similarly, the BatchNormalization layer and the Dropout layer are added between the layers (the parameter value of the Dropout layer is set to 0.5) to prevent the model from over-fitting, and the generalization ability of the model is enhanced. The LeakyRelu layer is also added after the CNN layer. Finally, the result of the classification is output through a fully connected layer and softmax as the activation function. At the same time, the model's optimizer uses the "adam" optimizer and incorporates learning rate attenuation techniques to help the model achieve better results.

4. Experiment Results and analysis

4.1. Empirical analysis

The three stock forecasting models proposed in this paper: LSTM, Encoder-Decoder LSTM and LRCN neural network stock forecasting model are compared to verify the rationality and effectiveness of LSTM series neural network model in the field of stock market prediction.

The selected data pool of this experiment is the SSE 50 series stocks, verifying the overall effect of the model on the stock market's overall stock, and ensuring the generalization of the model.

After the data preprocessing is completed, the training data is input into the models for training. After the training, the trained model is used to predict the test data. The performance of the two models on all 50 stocks of SSE 50 (the data are average) is shown in Table 1:

Table 1. The forecast Results of SSE 50 series stocks

Evaluation index	<i>LSTM</i>	<i>Encoder-Decoder LSTM</i>	<i>LRCN</i>
Accuracy	0.5425	0.5691	0.5773
Loss value	0.9616	0.9415	0.9401
Precision	[0.4079 0.6871 0.3851]	[0.4754 0.6290 0.4993]	[0.5493 0.5841 0.5201]
Recall	[0.3395 0.7655 0.2941]	[0.4031 0.7216 0.3878]	[0.2662 0.8557 0.2539]
F1 score	[0.3705 0.7241 0.3335]	[0.4362 0.6721 0.4365]	[0.3586 0.6942 0.3412]

The data in the table are in the form [a, b, c], where a represents the model's prediction index for the first category (ie, the falling stocks), b represents the model's prediction index for the second category (ie, the shock stocks), and c represents The model is a predictor for the third category (ie, rising stocks).

It can be seen from the above evaluation index data that the Encoder-Decoder LSTM model and the LRCN model are superior to the classical LSTM model in terms of accuracy and loss values. In terms of precision, the precision of the shock stocks is still the highest, and it is obvious that LRCN is stronger than the other two models in terms of the precision of the forecasting of the falling and rising stocks. In terms of recall rate and F1 score, the three models are still weaker for the falling and rising stocks, and LRCN is weaker than other models.

Overall, the Encoder-Decoder LSTM model and the LRCN model are superior to the classic LSTM model, but the LSTM model also performs well, demonstrating the effectiveness and potential of the LSTM series model in stock market forecasting.

4.2. Back test

Considering the input data of the backtest, the output of the stock forecasting model is the change of the closing price of the stock over the previous day, that is, the rise, fall, and shock. These predictions will serve as triggers for trading signals in the backtesting process, corresponding to long, short and no transactions in the transaction.

First, set the initial conditions for the backtest, the initial amount of funds is set to 1000000 RMB, the proportion of the handling fee is set to 0.00018, the proportion of stamp duty is set to 0.001, and the maximum number of buy or sell lots per transaction is set to 200 lots. Subsequently, according to the daily forecast of the model, the corresponding stock trading operation is carried out.

The overall average backtest results of the 50 stocks of the SSE 50 series are shown in the following table. The following data are average values, not cumulative values.

Table 2. The backtest Results of SSE 50 series stocks

Indicator	<i>LSTM</i>	<i>Encoder-Decoder LSTM</i>	<i>LRCN</i>
Average rate of return	23.52%	32.14%	11.86%
Average strategy rate of return	0.34%	0.44%	0.56%
Average winning rate	56.28%	56.86%	59.51%

In terms of total revenue, the Encoder-Decoder LSTM model has obvious advantages, the LSTM model is second, and the LRCN model has the least benefit. In terms of strategic rate of return, the LRCN model performed better. In general, all three models can obtain considerable returns, but the ability of the LRCN model to obtain revenue is significantly weaker.

5. Conclusion

In this experiment, the daily stock data of SSE 50 series stocks are trained by three models to predict the closing prices. According to the research predictive accuracy and other research indicators and the real backtesting performance indicators, it is empirically proved that the Encoder-Decoder LSTM model has stronger capabilities and better effects than the classic LSTM model in the stock market forecasting research. Although the LRCN model has a higher accuracy rate, the actual backtesting effect is the worst. It also shows that the LSTM model has a more aggressive trading strategy in the backtesting, but it can obtain relatively high returns in a good time period. The LRCN model is relatively stable and has the ability to reduce risk, but at the same time the benefits are low. The experiment shows that the LSTM series model can indeed obtain considerable benefits, and validates the validity, rationality and potential of the LSTM series model in the field of forecasting research in the stock market.

References

- [1] Engle, Robert F. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica* 50.4 (1982): 987-1007.
- [2] Chong, Eunsuk, C. Han, and F. C. Park. "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies." *Expert Systems with Applications* 83 (2017): 187-205.
- [3] Xiao, Ding, et al. "Deep learning for event-driven stock prediction." *International Conference on Artificial Intelligence* 2015.
- [4] Hu, Ziniu, et al. "Listening to Chaotic Whispers: A Deep Learning Framework for News-oriented Stock Trend Prediction." *Papers* (2018).
- [5] Singh, Ritika, and S. Srivastava. "Stock prediction using deep learning." *Multimedia Tools & Applications* (2016): 1-16.
- [6] Tashiro, Daigo, et al. "Encoding of high-frequency order information and prediction of short-term stock price by deep learning." *Quantitative Finance* (2019).
- [7] Naik, Nagaraj, and B. R. Mohan. "Stock Price Movements Classification Using Machine and Deep Learning Techniques-The Case Study of Indian Stock Market." *International Conference on Engineering Applications of Neural Networks* Springer, Cham, 2019.