International Journal of Economics and Finance; Vol. 10, No. 8; 2018
ISSN 1916-971X E-ISSN 1916-9728
Published by Canadian Center of Science and Education

Multi Factor Stock Selection Model Based on LSTM

Ru Zhang[1], Chenyu Huang[2], Weijian Zhang[1] & Shaozhen Chen[1]

[1] Finance Department of International Bussiness School, Jinan University, Zhuhai, Guangdong Province, China

[2] Financial Management Department of International Bussiness School, Jinan University, Zhuhai, Guangdong Province, China

Correspondence: Shaozhen Chen, Finance Department of International Business School, Jinan University, Qianshan Road 206#, Zhuhai City, Guangdong Province, Post No. 519070, China. E-mail: 1813012994@qq.com

Received: May 21, 2018 Accepted: June 15, 2018 Online Published: June 28, 2018

doi:10.5539/ijef.v10n8p36 URL: https://doi.org/10.5539/ijef.v10n8p36

Abstract

This paper takes CSI- 300 stock as the research object, and uses the LSTM model with memory characteristics and the traditional multi factor analysis to build an improved multi factor stock selection model. In back testing experiments, we use the trained LSTM model to forecast the stock returns and make a portfolio classification to construct the investment strategy. The result shows that the multi factor stock selection model based on LSTM has good profit forecasting ability and profitability.

Keywords: LSTM, quantitative investment, multi-factor selection model

1. Introduction

Quantitative investment has the advantages of objective rationality, accuracy, controllability, efficiency and sensitivity (Liang & Yongping, 2018), which have attracted the attention of the financial industry and academia. Among them, multi factor stock selection model (Malkiel & Fama, 1970; Asmess, 1997; Chen & Zhang, 1998; Mohanram, 2005) is widely used as a classic model of stock investment. In recent years, the rapid development of machine learning algorithms have provided new ideas for the research of quantitative investment, such as the use of support vector machine (Lifang et al., 2006; Yanfeng & Feng, 2006) and neural network algorithm (Hongxing & Zhaojun, 2002; Kun, Yong, & Wei, 2009; Bo, 2010; Wei, Weiqiang, & Bo, 2001) to predict stock prices. As an improved recurrent neural network (Xuejun & Win, 2016; Xiong, Nichols, & Shen, 2015; Ruiqi, 2015; Zhang, 2001), LSTM is more suitable for prediction of stock market time series than shallow neural network because it stores historical information through a cyclic feedback structure.

To sum up, this paper takes the CSI-300 stock as the research object, and uses the LSTM model with memory characteristics to combine the traditional multi factor analysis to build an improved multi factor stock selection model. In the test, the stock return rate of the trained LSTM model is predicted and classified, hoping to build a stock investment strategy with high yield and high accuracy, which will provide new ideas for the cross research in the field of neural network science and quantitative investment.

2. Theoretical Model

2.1 RNN Model

The biggest difference between RNN model and traditional neural networks is that it is tied to time. That is to say, it contains a network of cycles. The results of next time are affected not only by the input of the next time, but also by the output of the previous time, which means information has a lasting impact.

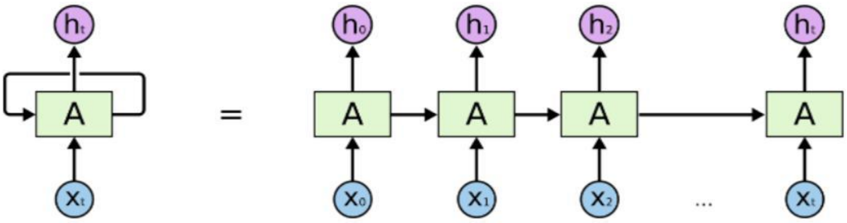

Figure 1. Recurrent neural network and its expansion form

2.2 LSTM Model

The LSTM is a well-designed RNN network, although both the LSTM and the original RNN contain three layers: the input layer, the hidden layer, and the output layer. However, the LSTM and the original RNN have a large difference in the design of the hidden layer, mainly because the LSTM has a special cell structure in the hidden layer. We can comparison the following two charts to better illustrate it.

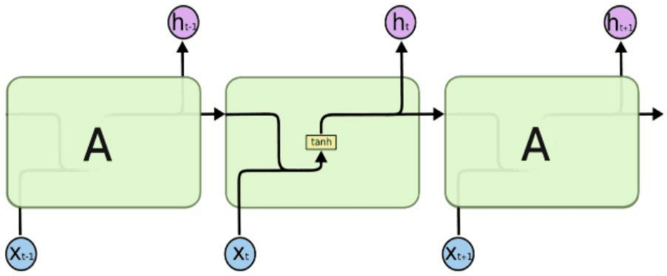

Figure 2. Design of the hidden layer of the original RNN

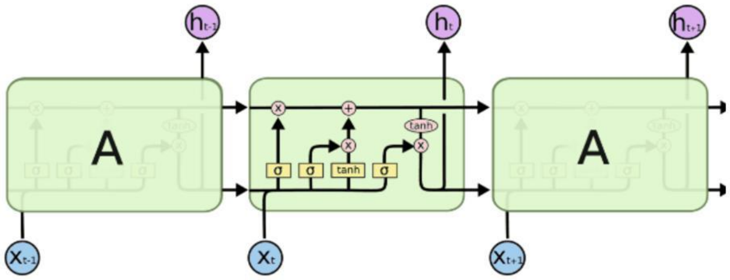

Figure 3. The hidden layer design of LSTM

Each black line transmits a single vector, from the output of one node to the input of other nodes. The pink circles represent the operations of pointwise, such as the sum of the vectors, the product, etc. The yellow matrixes are the learned neural network layer. The combined lines represent the connection of the vectors, and the separate lines indicate that the contents were copied and then distributed to different locations.

It can also be seen from Figure 3 that LSTM is to change a simple type of activation into several parts of the linear combination of storage cell to activate which means each time you can control the output information of the next step. For example, whether to include the previous information, how many problems are involved, and so on.

Each storage unit consists of three major components, input gates, output gates, and forget gates.

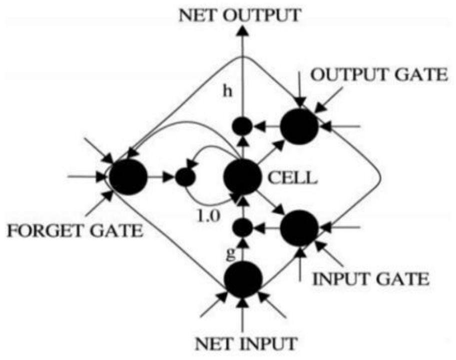

Figure 4. The unit structure of LSTM

Input gate: to control the current input and the previous output, and the information of the new cell enters here.

Forget gate: to transmit more effectively, information should be filtered to decide which information can be forgotten.

Output gate: update the information in the new cell state.

2.3 Multi Factor Stock Selection Model Based on LSTM

The data structure of the multi-factor model processing is standard panel data, including three dimensions: stocks, time and factors; the corresponding strains are the returns of $T+1$ period.

When applied to the LSTM network structure, there is some differences from the traditional multi factor model: the rate of return in the $T+1$ period is still a training label, the factor corresponds to the feature of the sample and the stock corresponds to a sample, but the time dimension is a cyclic process in LSTM: the factor data in the past $T-n$ period should be included in the forecast of yield of $T+1$.

3. Empirical Analysis

3.1 Parameter Setting

(1). Back testing time: From May 1, 2007 to April 30, 2016, the number of monthly data training samples under this time interval exceeds 18W (each stock represents one sample at the end of each month)

(2). Strategy Time: May 1, 2016 - April 30, 2017

(3). LSTM time length (steps): 24 months, that is, each training sample contains factor data of the past 24 months and input them into neural network from the first month, and circulate the return value and the next month factor into the neural network simultaneously, and so on, until the forecast value of twenty-fourth months is obtained.

(4). Number of factors: Due to training in the neural network, we do not evaluate the validity of the factors at the beginning of the period, nor do we combine the factors and input all them into models. (Excluding some of the factors that are highly correlated and belong to the same category, this process can reduce the possibility of model training overfitting). The final 48 small factors are selected and belong to 10 common style factors.

(5). Number of classifications: In order to verify the accuracy of forecasting and exclude some of the noise in the sample, we classify the sample yield types into three categories: rising (monthly yield is greater than 3%) and falling (monthly yield is less than -3%), Neutral (monthly rate of return is between -3% and 3%)

(6). Batch size: this parameter belongs to the system parameter of LSTM, which is the parameter used to calculate the gradient in the algorithm. That is to say, every training, the batch size samples in the total training sample are randomly selected as the training sample.

(7). Number of hidden neurons: This parameter also belongs to the system parameter of LSTM. It is the number of “nerves” that the input sample and hidden layer cells are connected to. It is limited by the performance of the computer and can only be set to three digits and 2 hidden layers.

(8). Learning rate: The LSTM system parameter is the speed at which the gradient falls during training. If it is too high, the gradient will easily disappear. If it is too low, the training will be too slow.

(9). Cross-checking ratio: To prevent overfitting of the model, 90% of the 18W samples were selected as the training set to train the model parameters, while the remaining 10% did not participate in the training and only tested as a test set. If the accuracy of the training set and the test set increase at the same time, the overfitting of the model may be too small.

3.2 Model Training

(1). Data preprocessing: according to the multi factor process, the cross section factor is treated with kicking off extreme value and standardization. At the same time, in order to eliminate the effect of the industry, the section single factor is used to return the industry matrix, and the residual is taken as the factor data for the final input.

(2) In-sample training: after 100 iterations, the result of training convergence has been observed.

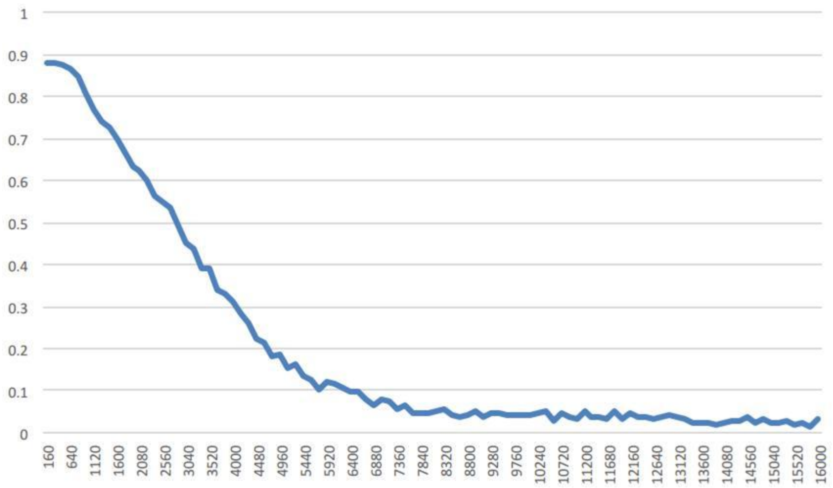

Figure 5. The loss rate of LSTM

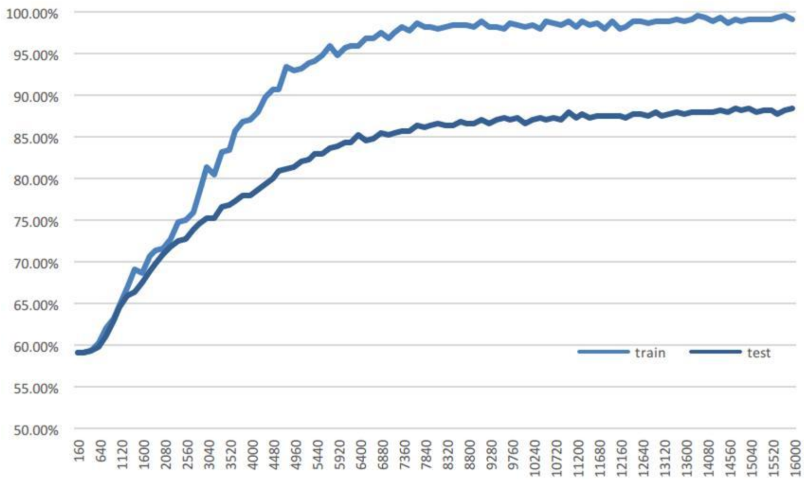

Figure 6. LSTM cross-check accuracy

3.3 Out-of-Sample Testing

Through the final result of the training, we enter the out-of-sample data 2016-2017 and get the model's estimate of stock returns over the next 12 months. The accuracy rate is shown in Figure 8:

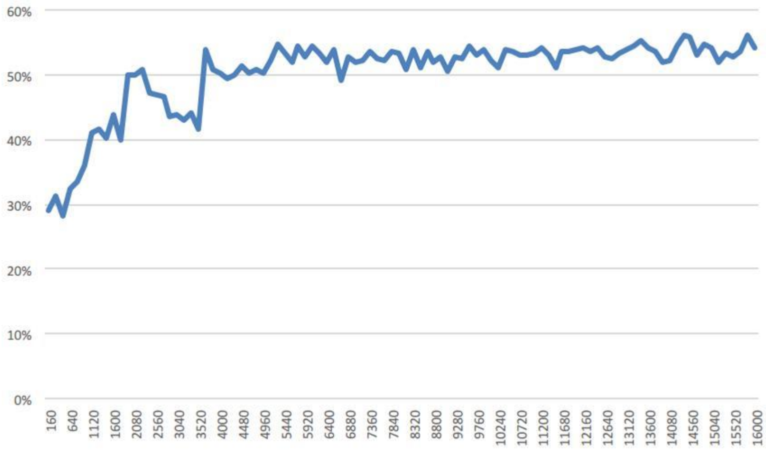

Figure 7. Accuracy of stock selection outside the LSTM sample

The final convergence level of the accuracy rate outside the sample is only higher than 50%, but it is necessary to distinguish the true prediction degree that this level can reflect. In order to intuitively test the stock selection effect outside the LSTM model sample, we choose the prediction result of each month that the model provide as the stock selection standard.

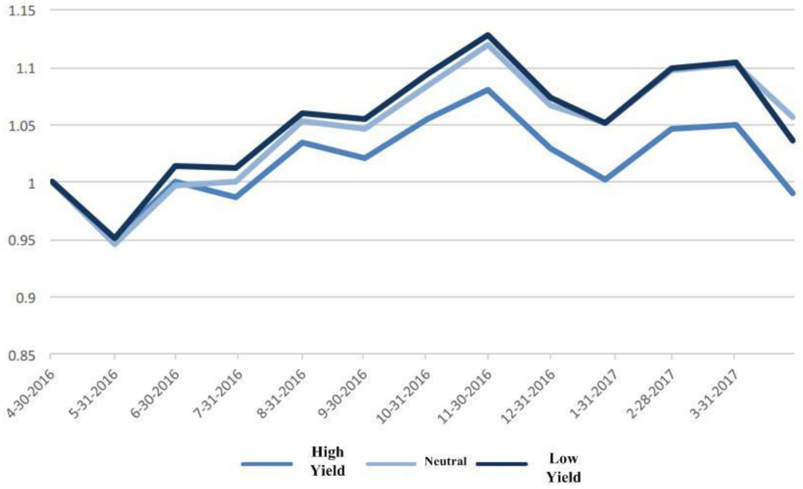

Figure 8. A-share forecast portfolio net value

It can be seen that in the most recent year, the model has a higher winning percentage for high and low returns, but it is less effective for forecasting the median neutrality.

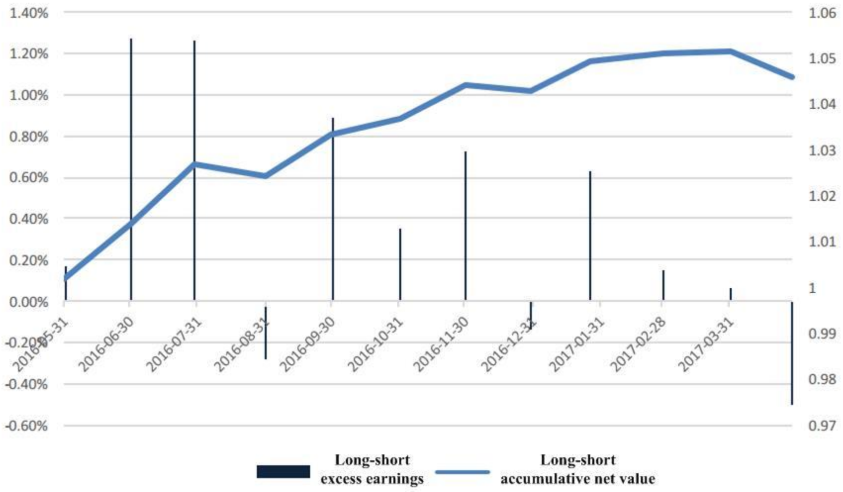

Figure 9. The cumulative net value of all A-share long and short portfolios

Over the past 12 months, the excess yield has been 75%. From the net cumulative value of short and long term, the excess earnings over the past 12 months were 4.5%.

In order to further verify the accuracy of the model for stock forecasting, we change the stock selection criterion from the model output to the activation value before the model final prediction. Because we classify the predicted target into three categories (high, medium and low), the neural network chooses the category with the largest activation value as the prediction category. Therefore, the activation value actually reflects the prediction probability of the model for the future stock returns.

Based on this, we reconstruct three types of stock portfolios. Each period selects 30% of the stock with the largest activation value as the corresponding combination.

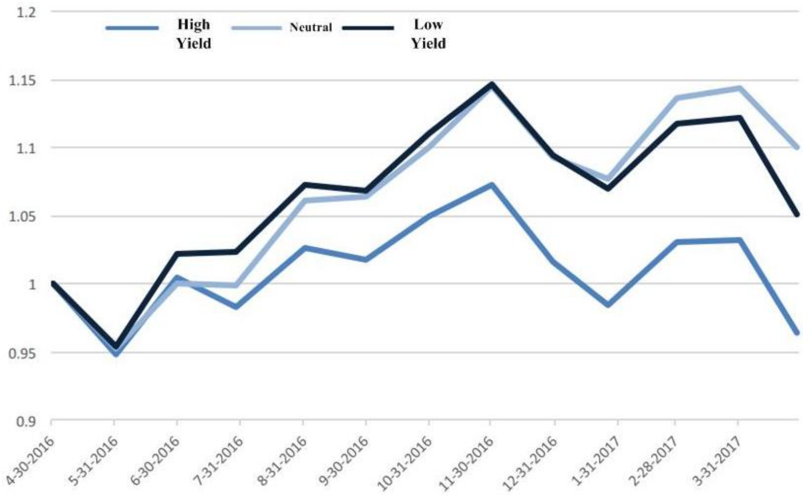

Figure 10. 30% space combination net value

It can be seen that the prediction effect of the model on the neutral earnings still has not improved, but the forecast effect of long-short return is more accurate than that of the full A-share.

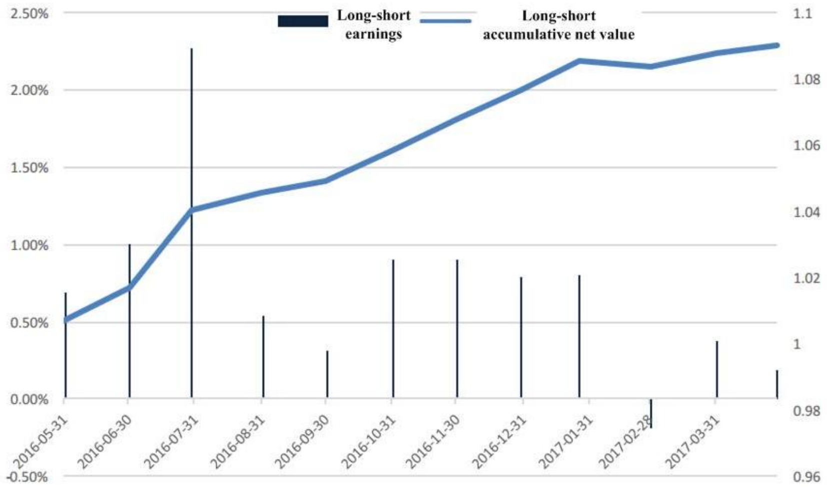

Figure 11. 30% cumulative net value of long-short portfolio

The excess return of Over-the-counter was 9%, while the monthly winning percentage of these 12 months exceeded 90%. Through the backtesting of out-of-sample data, we found that through the LSTM learning, the prediction of stock returns is actually more accurate. At the same time, the prediction probability of the model for different types of returns can further reflect the probability of stocks' rising and falling.

4. Conclusion

The development of multi-factor models tends to be mature, and the alpha yield of factors has declined. If it is a core issue to maintain the benefits of multi-factor models in the quantitative field, we believe that the directions of expansion includes new factor mining, stock pool differentiation, and exploration of non-linear factor features. Machine learning is an effective solution to nonlinear problems. Specifically to the LSTM involved in this article, it is through the extension of the time dimension and the expansion of the space depth to spread the current factor space into the space of higher dimension, and find the effective path in it to realize the prediction of the factor model.

After strictly distinguishing the training set, the test set, and the data set, we can get the convergent result with higher accuracy through training, and get significant excess returns in the data back test. The accuracy of cross test is close to 90%, and the winning rate of short term out of the sample is more than 90% in the recent 12 months. The surprising point of these results is that by using the basic LSTM structure, such high accuracy and significant level can be got before the optimization of parameters, and further improvement and optimization of the model can be expected. At the same time, these results are within expectation which means their powerful data processing capabilities will be exposed in the field of investment when we no longer use machine learning and neural networks as complex "black boxes".

References

Asmess, C. S. (1997). The Interaction of Value and Momentum Strategies. *Finance Analysis Journal,* (3), 29-36. https://doi.org/10.2469/faj.v53.n2.2069

Bo, S. (2010). *Research on stock price forecasting based on BP neural network.* Hunan University, 2010.

Chen, N., & Zhang, F. (1998). Risk and Return of Value Stocks. *Journal of Business, 71*(4), 501-535. https://doi.org/10.1086/209755

Hongxing, Y., & Zhaojun, S. (2002). Research on Wavelet Neural Network Method in Stock Market Forecasting. *Journal of Industrial Engineering and Engineering Management, 16*(2), 32-37.

Kun, Z., Yong, Y., & Wei, L. (2009). Stock price model based on combination of wavelet and neural network. *Computer Engineering and Design, 30*(23), 5496-5498.

Liang, O., & Yongping, T. (2018). Application of Quantitative Investment in Futures Market. *Inner Mongolia Coal Economics,* (2), 81-82. https://doi.org/10.13487/j.cnki.imce.011412

Lifang, P., Zhiqing, M., Hua, J., et al. (2006). Application of Support Vector Machine Based on Time Series in Stock Forecasting. *Computing Technology and Automation, 25*(3), 88-91.

Malkiel, B. G., & Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance, 25*(2), 383-417. https://doi.org/10.2307/2325486

Mohanram, P. S. (2005). Separating Winners from Losers among LowBook-to-Market Stocks using Financial Statement Analysis. *Review of Accounting Studies, 10*(2-3), 133-170. https://doi.org/10.1007/s11142-005-1526-4

Ruiqi, S. (2015). Study on the forecast model of the price trend of US stock index based on LSTM neural network. *Capital University of Economics and Business*.

Wei, W., Weiqiang, C., & Bo, L. (2001). Using BP Neural Network to Forecast Stock Market Ups and Downs. *Journal of Dalian University of Technology, 41*(1), 9-15.

Xiong, R., Nichols, E. P., & Shen, Y. (2015). Deep Learning Stock Volatility with Google Domestic Trends.

Xuejun, J., & Win, C. (2016). The Impact of American Expansionary Monetary Policy on China's Inflation——An Analysis Based on In-depth Long-and Short-term Memory Neural Network. *Shanghai Finance,* (3), 80-83. https://doi.org/10.13910/j.cnki.shjr.2016.03.014

Yanfeng, W., & Feng, G. (2006). Stock Market Prediction Based on Support Vector Machines. *Computer Simulation, 23*(11), 256-258.

Zhang, G. P. (2001). An investigation of neural networks for linear time-series forecasting. *Computers & Operations Research, 28*(12), 1183-1202. https://doi.org/10.1016/S0305-0548(00)00033-2

Copyrights