

Research on trading strategy based on improved LSTM network model

Sitong Peng^{1,*,#}, Zijue Zhang^{1,#}, Zhifeng Lin^{2,#}

¹ Business college, ZhuJiang College, South China Agricultural University, Guangzhou, Guangdong, China

² School of Information Engineering, ZhuJiang College, South China Agricultural University, Guangzhou, Guangdong, China

*Corresponding author: 771933821@qq.com

These authors contributed equally.

Abstract. The economic foundation determines the superstructure of society. Sustainable development of society is inseparable from economic prosperity, and investment has an irreplaceable role as one of the "troika" to promote social economy. Based on this, the modeling in this paper aims to study the trading strategy problem. In this paper, firstly, outliers and missing values are processed by Laiida's rule for the attached data, and then a multi-layer LSTM weighted model is proposed to learn the influencing factors of input time span on output time span by using multi-layer LSTM network, and the model is trained to achieve higher accuracy relative to the LSTM model through the embedded method: superposition of the influence generation cycle sequence. Attention mechanisms are introduced and short and long term forecasts are continuously made. The fluctuations of MACD, RSI indicators, Apriori and other parameters are quantified; then, a risk measurement index is constructed, where the greater the cyclical price fluctuations, the greater the risk. The results show that timely buying and selling maximizes returns, and that a \$1,000 investment ends up with a \$900,000 return. The model is run again by increasing or decreasing the percentage of transaction costs, setting 0.5%, 1%, 1.5% and 2% points, to test the magnitude of sensitivity (trading decision results) at different transaction costs. The results show that the model is more robust, i.e., the model constructed in this paper has better results under different transaction costs.

Keywords: Support vector machine; stochastic forest algorithm; sensitivity analysis.

1. Introduction

The increasingly mature capital market has not only become an important channel for financing large, medium and small enterprises in China, but also a convenient and fast investment platform for the majority of residents and institutional investors in China. However, how to screen out financial instruments with investment value from the market with hundreds of stocks, bonds and funds distributed and obtain excess returns from them has always plagued the majority of investors, and it is also a hot issue in the study of financial academia [1]. At present, there are three main schools of securities price analysis theory widely used in the investment industry: fundamental analysis theory, technical analysis theory and quantitative investment theory that has been hot in recent years, the entry points and methods used in the three theoretical research are different, each has its own advantages and disadvantages, in general, the performance effect of quantitative investment theory in specific practice is better than the first two [2].

Whether it is fundamental analysis or technical analysis, it is related to the subjective experience of investors, and it is easy to be dominated by the greed and fear of human nature, and it is often unable to achieve due returns in investment decisions. Since the 1970s, with the continuous innovation of financial theory and the rapid development of computer technology, a more rational and higher-yield investment strategy has begun to circulate, which is today's third-party securities price analysis method-quantitative investment [3]. Quantitative investment is based on massive data, through mathematical models to achieve the selection of high-quality stocks, in practice has repeatedly surpassed the market performance, for example, the famous medal fund based on quantitative investment from 1988 to 2010 average annual return of 33%, while the same period is

good at fundamental analysis theory to invest in the average annual return of 22%. Quantitative investment is highly sought after in the investment community for its strict discipline, systematicness, decentralization and probability, and this paper realizes quantitative stock selection by establishing a support vector machine and a random forest model, proving the effectiveness of quantitative stock selection in practice [4].

The main research tasks of this paper are: Develop a model that gives the best daily trading strategy based only on the price data of the day. Using your model and strategy, how much is the initial \$1,000 investment worth on September 10, 2021?

Provide evidence that your model provides the best strategy. Determine the sensitivity of the strategy to transaction costs. How do transaction costs affect strategy and results?

2. Model construction and improvement

2.1 Problem analysis

Problem requires the development of a model that gives the best daily trading strategy based only on the price data of the day. Using your model and strategy, how much is the initial \$1,000 investment worth on September 10, 2021?

Firstly, the raida method is used to process the outliers and missing values for the attachment data, and then, a multilayer LSTM weighted model is proposed, and the multilayer LSTM network is used to learn the influencing factor of the input time span on the output time span, through the embedded method: influencing the superposition of the generation cycle sequence, training the model, and achieving higher accuracy relative to the LSTM model. The Attention mechanism was introduced and short-term and long-term forecasts were made continuously. The fluctuations of parameters such as MACD, RSI indicator, Apriori, etc [5]. are quantified; then, a risk measurement index is constructed, and the larger the cyclical price fluctuation, the greater the risk. The results show that timely buying and selling will maximize the returns, and the investment of \$1,000 will eventually have a return of \$900,000 [6].

2.2 Model building

2.2.1 Feature reconstruction of GBDT models

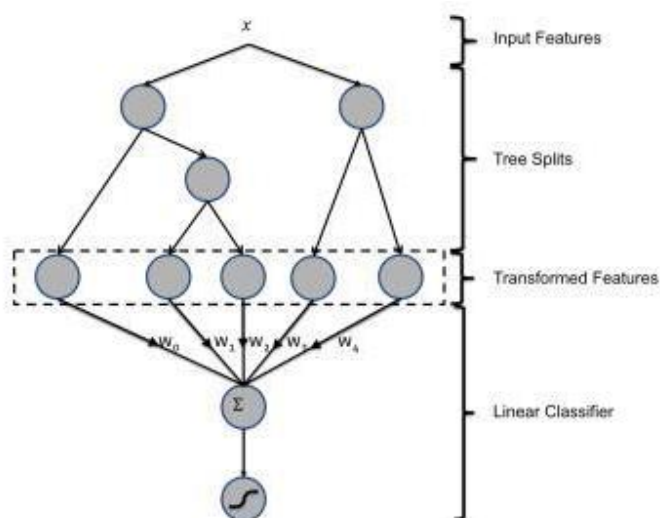


Figure 1. GBDT model

To understand the feature reconstruction of GBDT, Figure 1 shows a simple GBDT and FFP merging model with only three subtrees, $t=3$. Suppose there is a sample x at this time; entering the GBDT model, the sample is divided into two leaf nodes: the second leaf node of the first subtree, the

second leaf node of the second subtree, and the first leaf node of the third subtree, as shown in the red node in the Figure 1 [7].

Factor analysis techniques, logistic regression and Pareto principle were used in data analysis process. First, Kaiser-Meyer-Olkin (KMO) test and Bartlett test will be performed. According to Williams, Onsman and Brown (2010), the KMO test and Bartlett' test should be used to assess if the data is appropriate for factor analysis.

The process of forming the GBDT model, that is, the process of establishing a subset, is actually a process of continuous combination of original individual features, and the combination of features is usually superior to individual features. New features of GBDT reconstructions usually have strong expressive abilities.

If the number of candidate factors for a multifactor model is n , and the feature names are encoded in natural numbers, then the original feature set of the dataset is, the original feature vector of the dataset input sample.

$$x_i = \{x_1, x_2, \dots, x_N\} \quad (1)$$

$$y_i = \{y_1, y_2, \dots, y_N\} \quad (2)$$

GBDT then maps the input vector to a t -dimensional vector called, the feature reconstruction process, i.e.: $x_i \in R^N \omega_i \in R^T$

$$GBDT : x_i \rightarrow \omega_i \quad (3)$$

Among them are: $x_i \in R^N$

$$\omega_i = \{\omega_1, \omega_2, \dots, \omega_T\}, \omega_k \in L_K \quad (4)$$

2.2.2 FFM model input

The FFM model input is shown in Figure 2:

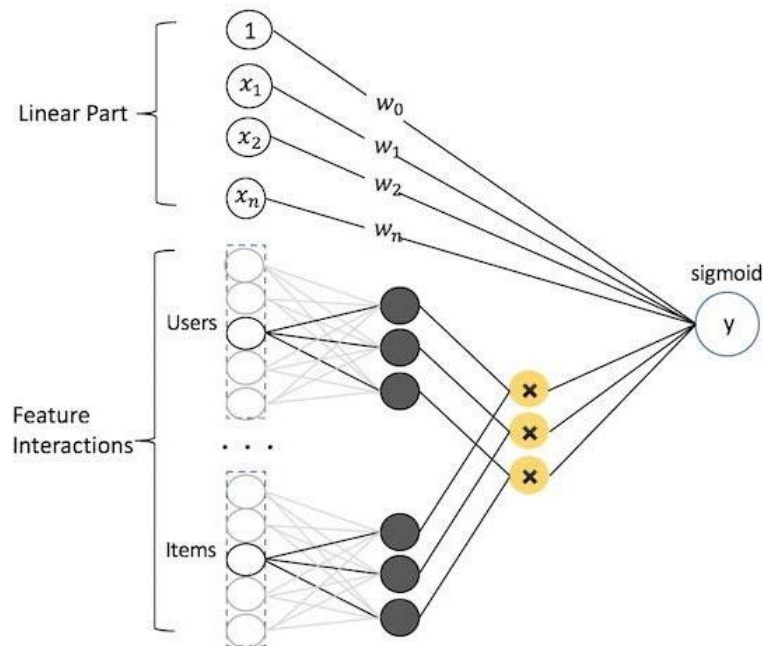


Figure 2. Model input

The input vector of the FFM model must contain the fields, the features under the fields, and the values corresponding to the features, where the new features reconstructed by the GBDT need to be further converted into an acceptable form of the FFM model [8].

The number of subtrees is the same as the number of GBDTs, and the dimensions of the input vector are exactly equal to the sum of the leaf nodes of all GBDT subtrees. In short, it can be expressed as a linear regression problem:

$$r_{i,t} = \alpha_i + \beta_i^{(1)} F_t^{(1)} + \beta_i^{(2)} F_t^{(2)} + \dots + \beta_i^{(k)} F_t^{(k)} + \varepsilon_{i,t} \quad (5)$$

In general, it represents the return of risk asset i at time t, and, $r_{i,t}$ F is called factor return, which represents factor risk exposure. namely: β

$$r_t = \alpha + \beta^{(1)} F_t^{(1)} + \beta^{(2)} F_t^{(2)} + \dots + \beta^{(k)} F_t^{(k)} \quad (6)$$

At this point, a multivariate model equation (6) will be written to represent the sensitivity of stock i to a particular risk at t-time.

$$r_{i,t} = \beta_t + X_{i,t}^{(1)} f_t^{(1)} + X_{i,t}^{(2)} f_t^{(2)} + \dots + X_{i,t}^{(k)} f_t^{(k)} + \varepsilon_{i,t} \quad (7)$$

2.3 LSTM

A typical LSTM unit comprises one or more internal state memory cells, front doors, forgetting gates, and output gates, as shown in Figure 3. Assuming the state of the storage unit at time t, the calculation of the LSTM unit at time is as follows:

$$\begin{cases} s_t = s_{t-1} f_t + i_t g_t \\ o_t = \text{sigmoid}(W^{xo} x_t + W^{ho} x_{t-1} + b_o) \\ h_t = o_t \tan(s_t) \end{cases} \quad (8)$$

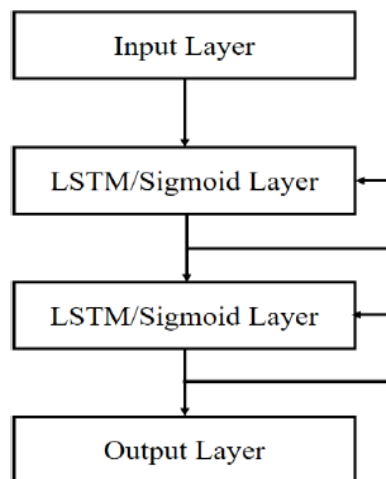


Figure 3. Multilayer LSTM network structure

In this paper, a multilayer LSTM network is used to learn the temporal correlation of time series data, and historical inputs are used to predict the value of the current data, as shown in Figure 4. The input vector is a part of V, there is a certain temporal correlation between the components, the exponent represents a specific time step; is the predicted value of all or part of the first stage; represents the hidden layer, at the time of deployment, and represents the LSTM layer in the step; is the input time step, that is, the step data is used as the input data for the prediction step [9].

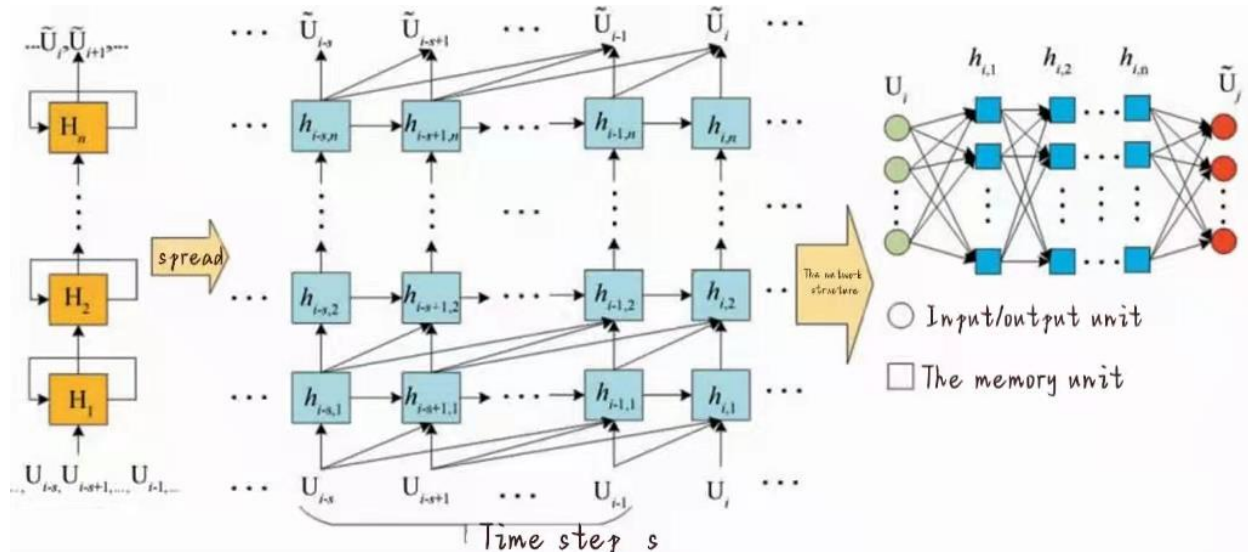


Figure 4. Schematic diagram of multilayer LSTM network structure

Sequence prediction is essentially a regression problem, and for anomaly detection tasks, the normal number of samples is much greater than the number of anomalous samples, at time steps, as input to a multilayer LSTM network. Used as the target output, the loss function is minimized (shown below) $iU_{i-s}, U_{i-s+1}, \dots, U_{i-1}U'_i$

$$Loss = \frac{1}{2}(U'_i - U_i)^2 \quad (9)$$

For the training goal, the backward propagation algorithm is used for training.

3. Model application and sensitivity analysis

3.1 Application of the model

According to the description in the previous section, we use two LSTM models to predict the anomaly data based on the model idea of Figure 4, and by weighting, we aim to improve the accuracy of outlier 1 and normal value 0 at the same time.

The weighted formula is:

$$o_t = \text{sigmoid}(W^{xo}x_t + W^{ho}x_{t-1} + b_o) \quad (10)$$

The numerical pair of accuracy values for a single LSTM model and an improved LSTM model is shown in Table 1.

Table 1. Accuracy comparison

	Acc_0	Acc_1	Acc_all
Single LSTM model (1 day)	0.753	0.619	0.751
Improved LSTM model (1 day)	0.831	0.846	0.847

Through the above comparison, we can conclude that in addition to considering the time span of the model input and the time span of the model output, the improved LSTM model predicts whether abnormal values will occur in the future, and after calculation, the model predicts the F1 value of 0.816, which greatly improves the accuracy of the traditional LSTM model.

Figure 5 shows the predicted convergence curve.

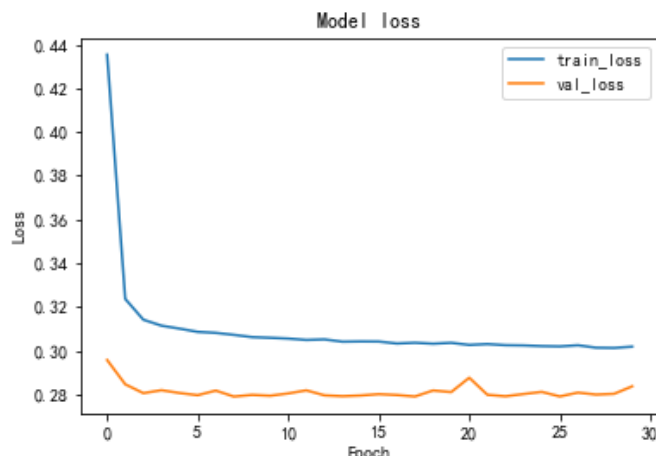


Figure 5. Convergence curve

3.2 Sensitivity analysis

Set 0.5%, 1%, 1.5%, and 2% points by increasing or decreasing the percentage of transaction costs, run the model again, and test the sensitivity size (transaction decision results) under different transaction costs.

Sensitivity analysis is the study and analysis of the sensitivity of changes in the state or output of a system (or model) to changes in system parameters or surrounding conditions. Sensitivity analysis is often used to study the stability of the optimal solution when the raw data is inaccurate or changing, and sensitivity analysis can also determine which parameters have a great influence on the system or model. Local sensitivity analysis, also known as the point variational method, is characterized by a single parameter, taking the central value of other parameters, and evaluating the change in the model results each time the parameter changes, and there are two conversion methods: the first is the factor change method, if the pre-analysis parameters are increased by 10% or reduced by 10%; the other method is to modify the deviation, for example by adding a standard deviation to the pre-analysis parameter or reducing the standard deviation.

The overall sensitivity analysis quantitatively determines the contribution of the parameters of the model to the error of the model results, and the main methods are the sobol method and the fourier amplitude susceptibility test unfolding method, both of which are based on variance. The variance of the model results can fully reflect the uncertainty of the model results, and not only the individual influence of parameters on the model results can be calculated, but also the influence of parameter interactions on the model results can be calculated. Through qualitative global sensitivity analysis, parameters with little impact on model results are screened.

The improved LSTM model principle is shown in Figure 6.

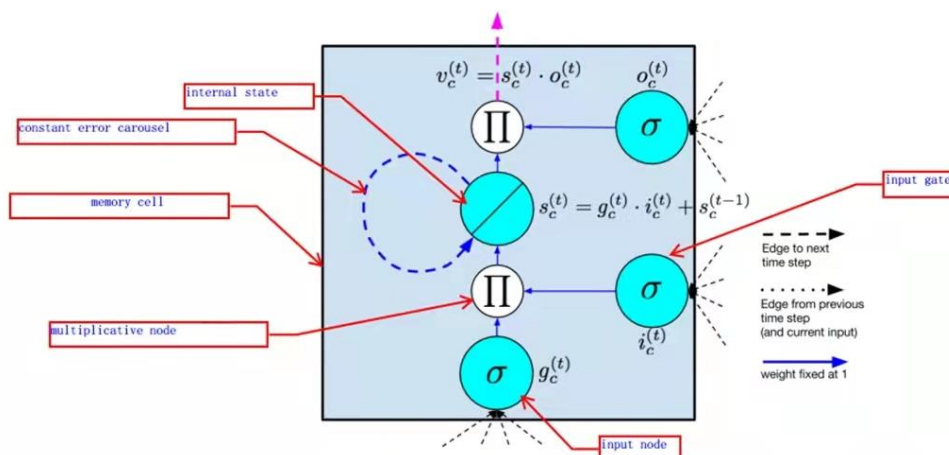


Figure 6. Improved LSTM model

4. Conclusions

The model has certain advantages in sequence modeling problems and has a long-term memory function. It's easy to implement.

The model solves the problems of gradient disappearance and gradient explosion in the training process of long sequences.

There are disadvantages in parallel processing. The effect is average with some of the latest networks.

The stock factor selection model was further studied, and in the first simulation experiment, the yield of a single model tended to be lower than that of multiple models, a idea known as model integration, the basic idea of which was to use multiple models of different branches. The final filtering of the combination of factors is then voted on based on the results of all the models, so we can try more different models, not only GBDT and FFP, but also RF or DNN, but the screening effect of these stock factors is worrying. The increase in the cost of multi-modal transport training is also a factor to consider.

In the course of writing this article, I have constantly consulted and studied journals re-lated to machine learning and quantitative trade, hoping to establish a synchronized link be-tween theory and practice. In order to solve the problems in the increasingly fierce industry competition. The capital market is composed of a variety of influencing factors, the research itself is very difficult and complex, and the understanding and application of quantitative investment and machine learning needs to be improved.

References

- [1] Cao Wen. Empirical research on multi-factor stock selection based on XGBoost algorit-hm [D]. Shandong University of Finance and Economics, 2021.
- [2] Lü Mengdi. Design of quantitative stock selection trading strategy based on price-to-book ratio interpretation model and stochastic forest algorithm [D]. Shanghai Normal University, 2021.
- [3] Wu Bingxin. Research on multi-factor quantitative stock selection strategy based on Lig-htGBM [D]. Northwest University, 2021.
- [4] Xu Zhijian. Research on Quantitative Trading Stock Selection Strategy of Thirty-Three Degrees Asset Management Company Based on Machine Learning [D]. Lanzhou University of Technology, 2020.
- [5] Cai Gaoyuan. Research on quantitative investment model of large-scale assets based on machine learning [D]. Henan University, 2020.
- [6] Wang Huaqin. Quantitative stock selection and effect evaluation based on machine lear-ning [D]. University of Electronic Science and Technology of China, 2020.
- [7] Wei Jing. Research on Quantitative Investment Strategy Based on Integrated Algorithm [D]. Hebei University of Technology, 2020.
- [8] Hü gege. Industry Stratified Stock Selection Strategy Based on Convolutional Neural N-etwork [D]. University of Electronic Science and Technology of China, 2020.
- [9] ZHU Xiaoqiang. Quantitative multi-factor stock selection strategy of dynamic optimizati-on machine learning algorithm [D]. East China University of Political Science and Law, 2020.