Research on stock prediction based on LSTM and CatBoost algorithm

Yu Sun, Liwei Tian*

* Corresponding author: 656453927@qq.com

Guangdong University of Science and Technology, China

ABSTRACT: Stock prediction is a classical problem at the intersection of computer science and finance. How to find an accurate, stable and effective model to predict the rise and fall of stocks has become a hot research topic among financial scholars. In the face of the increasingly prominent demand for stock analysis technology, combined forecasting model began to develop and achieved a lot of results. In this paper, we take the future financial time series up-down trend as the forecast goal, take the stock history data attribute value as the research object, based on the depth machine learning method, the combination model of LSTM and CatBoost optimized by Bayesian algorithm is used to predict the rise and fall of stocks. The model is validated by three evaluation indexes: MSE, MAE and Accuracy, it is concluded that the LSTM-BO-CatBoost model is more stable and feasible than LSTM-CatBoost, LSTM-XGBoost hybrid model, single LSTM network model and RNN network model.

Keywords: LSTM, CatBoost, Bayesian optimization, Stock price forecasting, Time series data.

1 INTRODUCTION

Many uncertainties, such as national policy, news, market sentiment, force majeure and so on, can cause sharp changes in share prices in the short term. Time-price series of stocks are often considered as dynamic non-parametric, chaotic and noisy non-linear series [1]. The prediction of the rise and fall of stock prices has become the most concerned issue for investors and researchers. In the past several decades, most scholars only use one single model to predict the stock market. but the stock price data is nonlinear, and there is some noise, a single model algorithm will lead to the neglect of other key information, resulting in a single model can't get good forecasting results [2]. Therefore, many researchers focus on the combination model to solve the problem of single model, and find that the combination prediction model can make full use of the characteristic information of data samples [3].

The world has entered the era of artificial intelligence, and machine learning algorithms have begun to show their skills in all walks of life. There are also certain research achievements in financial forecasting. In 2020, Zhao H R and Xue L proposed the research of stock prediction based on the LSTM-CNN-CBAM model [4]. In 2021, Meng Y and Xu Q J, put forward a stock forecast based on CNN-BiLSTM and attention mechanism [5].In 2022, Xiong Z and Che W G proposed to combine ARIMA network model, GARCH and M algorithm model, and studied their application in short-term stock prediction [6]. Based on the rolling residual model of

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China

DOI 10.4108/eai.19-5-2023.2334326

ARIMA and SVR, the combination model is used to predict the stock market, and good results are obtained [7]. In 2023 January, He Y and Li H improved NSGA-III-XGBoost hybrid model to predict stock price problems. Actual stock price data sets are used to test the predictive accuracy of the proposed model [8]. On the basis of the above hybrid model research, this paper proposes a stock prediction model based on the combination of LSTM and CatBoost algorithm, the stability and feasibility of the new hybrid forecasting model proposed in this paper are verified.

2 RELATED WORK

2.1 LSTM

LSTM (Long Short-term Memory Networks) was proposed by Hochreiter & Schmid-huber in 1997[9] and modified by Alex Graves in 2012[10]. LSTM is now widely used in artificial intelligence applications. LSTM is a variant of RNN (Recurrent Neural Network) [11], and it can solve this problem which RNN has the defect of not being able to handle long-distance dependencies. The concrete structure of the LSTM is shown in Figure 1.

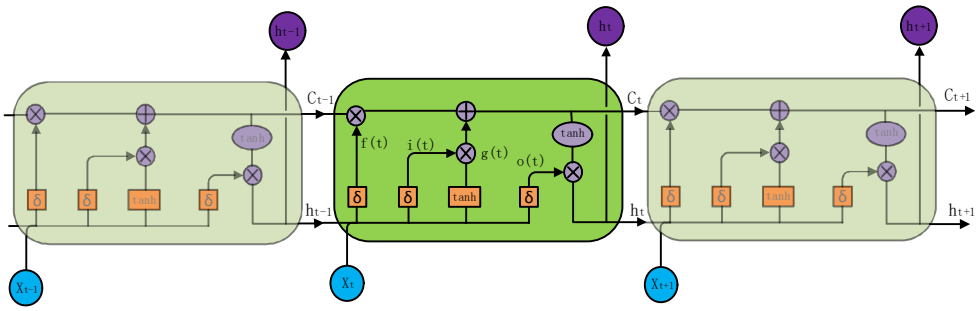

Figure 1. LSTM structure Diagram.

Where X_t represents the input at the current t time, and the following is Calculation formula of LSTM:

$$i_t = \sigma(W_i * [h_{t-1}, X_t] + b_i) \tag{1}$$

$$f_t = \sigma(W_f * [h_{t-1}, X_t] + b_f) \tag{2}$$

$$o_t = \sigma(W_o * [h_{t-1}, X_t] + b_o) \tag{3}$$

$$\hat{C}_t = tanh(W_c * [h_{t-1}, X_t] + b_c) \tag{4}$$

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

Among them, W_i , W_f , W_c and W_o are the weight matrix of input gate, forgetting gate, update gate and output gate respectively. b_i , b_f , b_c and b_o are the offset of input gate, forgetting gate, update gate and output gate respectively, so as to calculate the output h_t at the current t time and the updated cell state C_t at the current t time, and h_t is the output of the hidden layer.

2.2 Catboost model

CatBoost is an optimal implementation of GDBT algorithm based on symmetric decision tree-based learner. It can optimize the class data and overcome the model deviation and gradient error, thus, overfitting is reduced and the generalization performance and precision of the model is enhanced [12].

CatBoost uses an effective strategy called Greedy TS, which can reduce over-fitting on the one hand, and train with all data on the other hand. The data set is randomly sorted first. For each sample, when converting to a numeric type, the label value is averaged based on the label value of the category before the sample, at the same time, the weight coefficient of priority (prior value) is added. If $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_n)$ is a random permutation sequence, then:

$$x_{i,j} = \frac{\sum_{k=1}^{n}\left[x_{k,j} = x_{i,j}\right] * Y_k + \alpha * p}{\sum_{k=1}^{n}\left[x_{k,j} = x_{i,j}\right] + \alpha} \tag{7}$$

$[\cdot]$ denotes the indicator function, $x_{i,j}$ denotes the class i value of the j-th feature, Y_k denotes the corresponding label value, and molecules denote the sum of the corresponding label values of the Class i value of the j-th feature, the denominator denotes the number of class i values for the j-th feature, and P represents a priori. For the regression task, the mean value of the label is calculated as a priori value; for the dichotomous task, the probability of occurrence of a positive class is a priori value. α represents the weight coefficient of priority, and α is the weight coefficient greater than 0. Adding a prior distribution term can reduce data noise and overcome the influence of low frequency category data on the model.

2.3 Bayesian optimization

Bayesian optimization [13] is a very effective global optimization algorithm, the goal is to find the global optimal solution, machine learning algorithm super-parameter selection has an absolute advantage. Bayes optimization can effectively solve the problem of low efficiency of manual parameter adjustment, which is preferred by many researchers. The principle of Bayesian regularization is to analyze and judge the signal according to its uncertainty, estimate and predict some uncertain cases subjectively, and then adjust the event probability according to the Bayesian equation, then the optimization judgment is made according to the adjustment probability of the expected value.

The ideas are as follows:

- Calculate the expression of class conditional probability density parameter and prior probability according to historical data information.
- Use Bayesian formula (formula (9)) into a posteriori probability.
- Make a decision according to posterior probability.

Let D_1, D_2, ... , D_n be a partition of the sample space domain S, if it represents the probability of event Di occurring, then for any event x, P (x) > 0, there are:

$$p(D_j|x) = \frac{p(x|D_j)P(D_j)}{\sum_{i=1}^{n} p(x|D_i)P(D_i)} \tag{8}$$

Then we use the Bayesian regularization algorithm to maximize the posterior probability and get the parameter w, that is:

$$w^* = argmax_w\, p(w|D) = argmax_w \frac{p(D|w)P(w)}{P(D)} \tag{9}$$

In the formula, $p(D|w) = \prod_{k=1}^{n} p(D_k|w)$, $p(D|w)$is the probability of the occurrence of the observed data d of the likelihood function in the case of the parameter vector w, and P(w) is the prior probability of the parameter vector w. The objective function of the neural network can be optimized by determining the prior probability to get the regular term, which can improve the generalization ability of the network and the overall performance of the network.

3 MODEL BUILDING

In this paper, the workflow is to obtain data, analyze correlation, determine training sets and test sets, build models, and compare and analyze prediction results. The concrete flowchart of the model is shown in Figure 2.

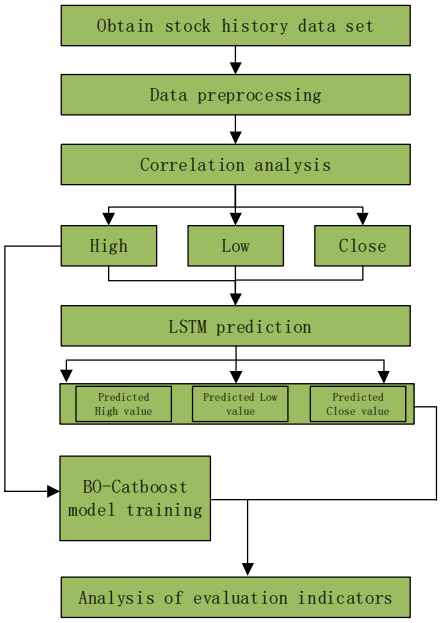

Figure 2. Model building.

The LSTM-BO-CatBoost model was constructed as follows:

(1) The historical data of the stock index was obtained from https://www.kaggle.com/, and the missing values were processed, and the training set was divided into the test set and the first 80% of the data in the data set were used as the training set;

(2) The "Date" attribute in the data set was decomposed into three attributes: "Year", "Month" and "Weekday". This paper makes a correlation analysis between all the attribute data and the "rise and fall" of the stock. After the analysis of the data, remove the Low correlation with the "rise and fall" of the attributes, leave "High", "Low", "Close" and "Adj Close" four attributes. The autocorrelation analysis of "High", "Low", "Close" and "Adj Close" shows that "Close" and "Adj Close" have strong correlation, therefore, only "High", "Low" and "Close" attributes are retained in the final data set;

(3) A two-layer LSTM model is constructed to train the "High", "Low" and "Close" attributes in the stock history data set respectively, build LSTM stock price prediction model, where LSTM model units = 100, batch, epoch = 10, the prediction results were reorganized as a new test set;

(4) Use the CatBoostRegressor method in the Sklearn package to implement the CatBoost regression prediction algorithm, bayesian optimization algorithm is used to optimize the "n_estimators", "max_depth", and "learning_rate" parameters in the CatBoostRegressor model, and train the "High", "Low", and "Close" attributes in the stock historical data set to build the BO-CatBoost stock price prediction model;

(5) Finally, BO-CatBoost stock price prediction model predicts the reconstructed data set after LSTM prediction, compares the difference between the real value and the predicted value, and judges the performance of the LSTM-BO-Catboost model in the stock price prediction.

4 SIMULATION EXPERIMENT

The experimental platform of this experiment is Jupyter Notebook. The experimental data set was downloaded from https://www.kaggle.com/ to obtain the historical time series data of 10 stocks named "MCD", "KO", "BAC", "XOM", "WMT", "ATVI", "CRM", "F", "GOOG" and "TSLA". The 10 stocks start on January 3,2013, and end on December 30,2022, for a total of 2,518 time series data. The historical data for the 10 stocks are shown in Figure 3. We can see that each stock is independent of each other.

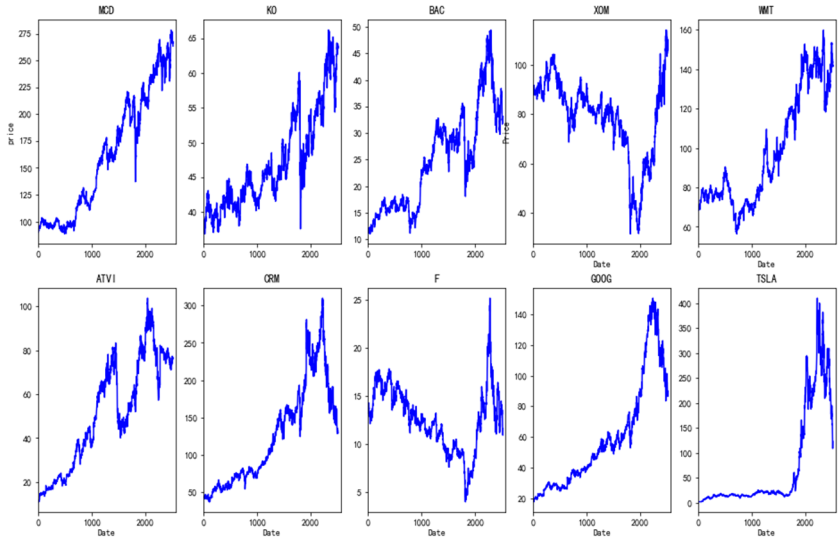

Figure 3. Chart of historical stock data.

4.1 Evaluation indicators

The goal of this experiment is to verify the accuracy and fit of the integrated algorithm of LSTM and CatBoost in stock prediction, therefore, the mean square error (MSE) and mean absolute error (MAE) are used to be fit performance evaluation index, Accuracy is used to evaluate the accuracy of the performance of the model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 \tag{10}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y^{(i)} - \hat{y}^{(i)}\right| \tag{11}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$y^{(i)}$ and $\hat{y}^{(i)}$ are the target real value and predicted value respectively, TP and FP are the number of correct and wrong forecasts of stock price increases, TN and FN are the number of correct and wrong forecasts of stock price declines.

4.2 Comparison of experimental results

In this paper, three machine learning models (XGBoost model, LSTM neural network and RNN neural network), which are commonly used in the field of financial forecasting, are used as nonlinear contrast models. The forecasting ideas of LSTM and RNN are as follows: the time series data of rising and falling attributes of stocks are taken as the input of the model, and the predicted value of the index closing price of the next 1 day is taken as the output of the model. For a more directly comparison of the models, LSTM-XGBoost uses the same input data as the LSTM-BO-CatBoost model presented in this article. In this study, first of all, the 10 stocks were carried out correlation analysis. They are in a similar correlation, and the correlation graph of two randomly selected stocks is shown in Figure 4.

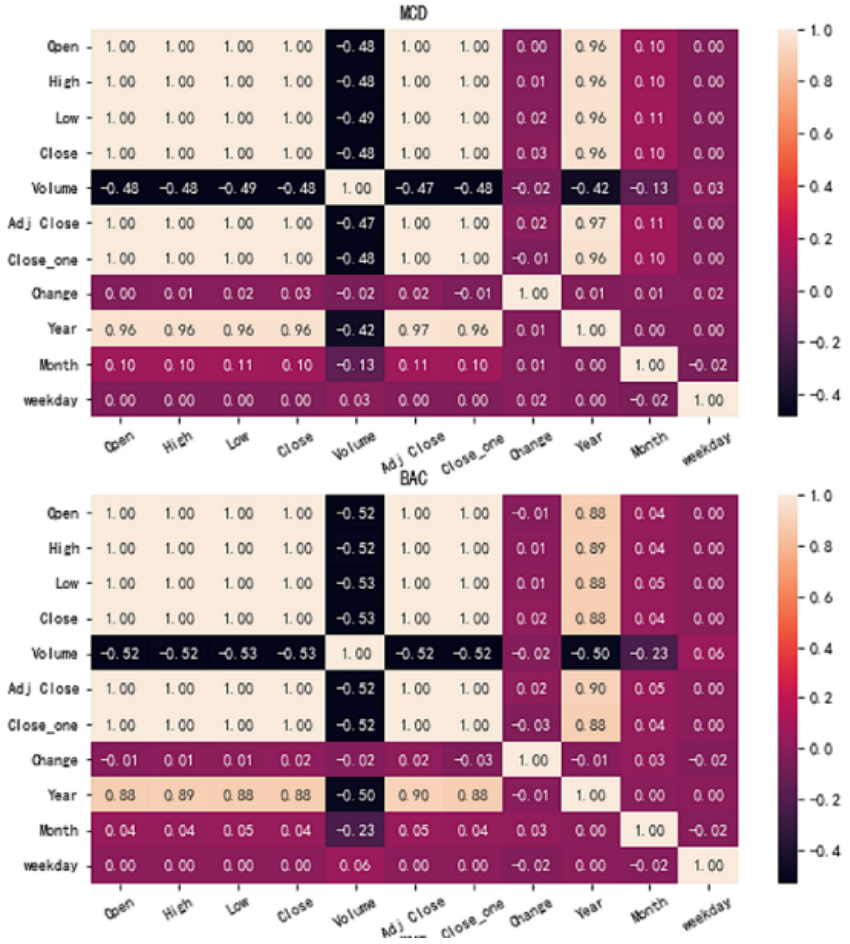

Figure 4. The correlation graph of randomly selected 4 stocks is shown in Figure 4.

As can be seen from Figure 4, Volume has negative correlation, which affects the experimental results. “High”, “Low”, “Close” and “Adj Close” have good correlation with the price tag values in the above 10 stocks, however, “Close” and “Adj Close” are strongly correlated, so “High”,

"Low" and "Close" are retained as the final data training and prediction. After determining the prediction attributes of each stock, the LSTM model is used to train and predict each attribute. Where the sliding window for the LSTM model is set to 1, that is, step prediction, while epochs is set to 10 and batch is set to 64. After the parameters of the LSTM model were set, the Catboost regressor model was trained. The core parameters of the CatBoost algorithm model were optimized by the Bayesian algorithm with 5-fold cross-validation, scoring = "neg_mean_squared_error", the optimal solution of "n_estimators", "max_depth" and "learning_rate" of CatBoost model is obtained, and BO-CatBoost model is constructed.

To verify the generalizability of the model, the model was trained and compared with 10 stocks named "MCD", "KO", "BAC", "XOM", "WMT", "ATVI", "CRM", "F", "GOOG" and "TSLA". The average comparisons of the three measures of MSE, MAE and Accuracy for 10 stocks with 10 times predict results are shown in Table 1.

Table 1. Prediction results of different models of 10 stocks.

Stock names / Indicator	MCD					KO				
	LSTM-BO-CatBoost	LSTM-CatBoost	LSTM-XGBoost	LSTM	RNN	LSTM-BO-CatBoost	LSTM-CatBoost	LSTM-XGBoost	LSTM	RNN
MSE	**7.45**	8.42	8.38	7.94	7.48	**0.50**	0.66	0.74	0.52	**0.50**
MAE	**2.06**	2.19	2.17	2.14	**2.06**	**0.51**	0.59	0.63	0.55	0.54
Accuracy	**0.51**	**0.51**	0.49	0.48	0.49	**0.53**	**0.53**	0.52	0.46	0.44
Stock names / Indicator	BAC					XOM				
	LSTM-BO-CatBoost	LSTM-CatBoost	LSTM-XGBoost	LSTM	RNN	LSTM-BO-CatBoost	LSTM-CatBoost	LSTM-XGBoost	LSTM	RNN
MSE	**0.49**	0.54	0.67	0.50	**0.49**	**2.58**	4.29	5.05	2.68	**2.58**
MAE	**0.54**	0.57	0.64	0.55	**0.54**	**1.22**	1.57	1.74	1.25	**1.22**
Accuracy	0.50	0.51	**0.54**	0.50	0.48	0.47	0.46	0.49	0.47	**0.50**
Stock names / Indicator	WMT					ATVI				
	LSTM-BO-CatBoost	LSTM-CatBoost	LSTM-XGBoost	LSTM	RNN	LSTM-BO-CatBoost	LSTM-CatBoost	LSTM-XGBoost	LSTM	RNN
MSE	**4.01**	5.10	5.64	23.17	31.35	**2.61**	2.95	2.93	8.61	14.29
MAE	**1.37**	1.63	1.70	4.51	5.34	**1.01**	1.10	1.12	2.62	3.55
Accuracy	**0.52**	0.50	0.52	0.48	0.48	0.48	0.48	**0.49**	0.47	0.47
	CRM					F				

Stock names / Indicator	LSTM-BO-CatBoost	LSTM-CatBoost	LSTM-XGBoost	LSTM	RNN	LSTM-BO-CatBoost	LSTM-CatBoost	LSTM-XGBoost	LSTM	RNN
MSE	**25.56**	34.08	44.12	292.43	384.60	**0.25**	0.28	0.28	0.33	0.40
MAE	**3.71**	4.31	4.91	16.45	19.04	**0.37**	0.38	0.40	0.46	0.53
Accuracy	0.53	**0.55**	0.53	0.53	0.51	**0.51**	0.50	0.51	0.50	0.50
Stock names / Indicator	GOOG					TLSA				
	LSTM-BO-CatBoost	LSTM-CatBoost	LSTM-XGBoost	LSTM	RNN	LSTM-BO-CatBoost	LSTM-CatBoost	LSTM-XGBoost	LSTM	RNN
MSE	5.50	5.51	**5.48**	6.02	6.02	**114.21**	114.53	198.60	155.30	215.89
MAE	**1.72**	**1.72**	1.73	1.82	1.83	**7.81**	7.82	11.16	9.55	12.02
Accuracy	**0.53**	**0.53**	0.48	0.50	0.52	**0.52**	0.51	0.48	0.51	**0.52**

According to Table 1 , we can clearly see that for 10 stock forecast data, the LSTM-BO-CatBoost model has better performance in predicting the rise and fall of 9 stocks in the MSE, MAE evaluation indicators, indicating that the model is better than other models in the fitting. Although the performance of MSE in Table 1 is poor, it is only slightly lower than that of LSTM-XGBoost model. Overall, the prediction performance of LSTM-BO-CatBoost model is relatively stable. At the same time, although the accuracy of "BAC", "XOM", "ATVI" and "CRM" does not obtain the best performance in the comparative reference model, the accuracy results are only slightly lower than that of the best model, indicating that the accuracy of this model also has good effect. Overall, the LSTM-BO-CatBoost model is more stable and efficient than the LSTM-CatBoost, LSTM-XGBoost, single LSTM and RNN models in stock price forecasting. Furthermore, the LSTM-BO- CatBoost model presented in this paper is proved to be feasible and stable in the prediction of stock price trend.

5 CONCLUTION

In this paper, a hybrid model (LSTM-BO-CatBoost) is proposed to predict stock price. We analyze the attributes of the stock data set, select the features, and divide the training set and test set. Then, the two-layer LSTM model feature attributes are used to predict and the predicted results are used to build a new test set. At the same time, the CatBoost model is used to train the original training set, the BO-XGBoost model is constructed by using Bayesian optimization algorithm to optimize the parameters of CatBoost model. Finally, BO-CatBoost model is used to predict the new test set. The results of prediction and evaluation are compared with the

LSTM-XGBoost hybrid model, the single LSTM network model and the RNN network model to verify that the LSTM-BO-CatBoost model proposed in this paper has better approximation ability and generalization ability in stock price forecasting. Although the model proposed in this study has some improvement in performance, but because stocks have some noise impact, the overall accuracy rate is not very satisfied, later, we will consider the impact of network public opinion on the stock price to improve the performance of the model to predict the stock price, and provide more valuable reference for people to grasp the overall rise and fall of stock prices.

ACKNOLEDGEMENT: The paper is funded by Dongguan social science and technology development project: Research on financial trend prediction technology based on deep learning (No. 20221800902782) ; Guangdong University of Science and Technology Scientific research projects: Research on financial time series prediction based on ensemble learning algorithm (No. GKY-2022KYYBK-36); Special project in key fields of ordinary colleges and universities in Guangdong Province in 2021 (new generation information technology): Research on cloud storage mechanism based on blockchain technology (2021ZDZX1075).

REFERENCES

[1] Jiang W. Applications of deep learning in stock market prediction: recent progress[J]. Expert Systems with Applications, 184: 115537(2021).

[2] Liu P. Research on the application of portfolio model in stock price forecasting [J]. Computer Simulation, 27(12): 361-364(2010).

[3] Zhao Q. Forecasting and comparative analysis of stock market price based on nonlinear mixed model [J]. Acta Shanghai Business School, 15(4) : 7(2014).

[4] Zhao H, Xue L. Research on stock prediction based on LSTM-CNN-CBAM Model [J]. Computer Engineering and applications, 57(03): 203-207(2021).

[5] Meng Yi, Xu Q. Stock prediction based on CNN-BiLSTM and attention mechanism [J]. Journal of Nanning Normal University (natural science edition), 38(04): 70-77(2021).

[6] Xiong Z, Che W. Application of ARIMA-GARCH-M model in short-term stock prediction [J]. Journal of Shaanxi University of Technology (natural science edition), 38(04): 69-74(2022).

[7] Chen D, Du F, Xia K. Stock prediction based on the combination of Arima and SVR rolling residual model [J]. Computer Age, 76-81(2022).

[8] He Y, Li H. Application of improved NSGA-III-XGBoost algorithm in stock prediction [J/OL]. Computer Engineering and Applications: 1-11(2023).

[9] Hochreiter S, Jü, Schmidhuber R A. Long Short-Term Memory[J]. Neural Computation, 1997.

[10] Graves A, Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 18(5–6):602-610(2005).

[11] Miao Y, Gowayyed M, Metze F. EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding[J]. 2015.

[12] Prokhorenkova LO, Gusev G, Vorobev A, et al. Cat Boost:Unbiased boosting with categorical features. Proceedings ofthe 32nd International Conference on Neural InformationProcessing Systems. Montréal: Curran Associates Inc, 6639–6649(2018).

[13] Eric Brochu,Vlad M. Cora, Nando de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning[J]. CoRR, abs/1012.2599(2010).