Int. J. Simul. Multidisci. Des. Optim. **14**, 8 (2023)

https://doi.org/10.1051/smdo/2023009

IJSMDO
Available online at:
https://www.ijsmdo.org

RESEARCH ARTICLE OPEN ACCESS

Application of PCA-LSTM algorithm for financial market stock return prediction and optimization model

Yanxiang Mi [1], Donghai Xu [2,*], and Tielin Gao[2]

[1] Huihua College of Hebei Normal University, Shijiazhuang 050091, China
[2] Business College of Hebei Normal University, Shijiazhuang 050024, China

Received: 8 May 2023 / Accepted: 28 July 2023

Abstract. Accurately predicting stock returns can help reduce market risk. This paper briefly introduced the long short-term memory (LSTM) algorithm model for predicting stock returns and combined it with principal component analysis (PCA) to improve the prediction accuracy. Simulation experiments were conducted on 80 stocks, and the PCA-LSTM model was compared with back-propagation neural network (BPNN) and LSTM models. The results showed that the PCA analysis method effectively identified the principal components of variable indicators. During the training iteration convergence, the PCA-LSTM model not only converged faster but also had smaller errors after stabilization. Moreover, the PCA-LSTM model had the highest prediction accuracy, the LSTM model was the second, and the BPNN model was the worst.

Keywords: Stock return / Prediction / Long short-term memory / Principal component analysis

1 Introduction

The rapid development of the market economy has also led to the development of the stock market. Although the stock market cannot directly create wealth, it promotes economic development and facilitates a certain degree of wealth redistribution [1], while also provides a platform for increasing individual wealth value. However, operating in the stock market is complex and volatile, and the changes in the stock returns of each listed company are affected by various objective and subjective factors [2]. Simply put, the stock market is a risky place, and it is almost impossible to consistently make a profit without experiencing losses. The only way to minimize the risk is by considering relevant factors that influence the market. One method of reducing risk is through forecasting stock returns [3], which allows investors to make informed decisions based on the results. With the development of computer technology, intelligent algorithms are gradually being used for stock forecasting. Intelligent algorithms can take advantage of data mining of big data to calculate the hidden laws of stock changes, which can then be used for accurate stock forecasting [4]. Zhang et al. [5] proposed an improved stock forecasting model. The Shanghai Stock Exchange Composite Index and the Taiwan Stock Exchange Capitalization Weighted Stock Index were used to validate the performance of the proposed method. The experimental results showed that the method outperformed other baseline methods. Yang et al. [6] developed a convolutional neural network (CNN)-based model for predicting time series using multi-factor analysis. They found that the prediction accuracy of the model was higher than the other models. Xiao et al. [7] proposed a cumulative autoregressive moving average method for basic stock market forecasting and found from simulation experiments that their multi-model fusion algorithm could achieve expected results, indicating universal applicability, market applicability, and stable feasibility. This paper briefly introduced the long short-term memory (LSTM) algorithm model for forecasting stock returns and combined it with principal component analysis (PCA) to improve accuracy. Simulation experiments were then carried out with 80 stocks. Moreover, the PCA-LSTM model was compared with back-propagation neural network (BPNN) and LSTM models. The novelty of this article lies in the combination of PCA and the LSTM model, using PCA to screen factors that affect stock returns, reducing data dimensions, and facilitating fast computation. Its contribution also lies in combining PCA with the LSTM model, reducing computational complexity by selecting important indicators and minimizing interference from redundant indicators, thereby improving the accuracy of the LSTM algorithm for predicting stock returns and providing an effective reference for predicting changes in the stock market.

* e-mail: gtl1990@hebtu.edu.cn

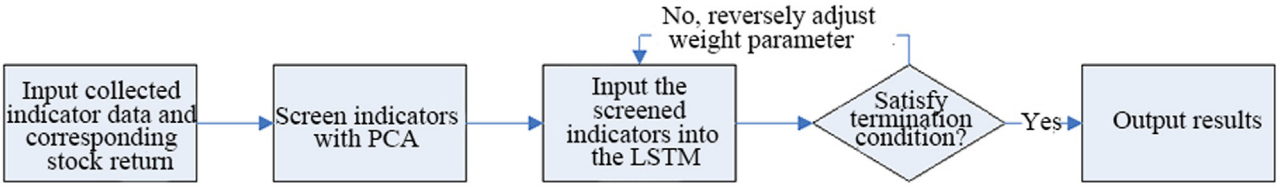

Fig. 1. Training flow of the LSTM algorithm after combined optimization.

2 Predictive models for stock returns

Intelligent algorithms that can be applied to stock returns include correlation rule analysis, support vector machines, and neural networks, among which the neural network algorithm uses the nonlinear characteristics of the activation function to fit hidden patterns [8]. All the aforementioned intelligent algorithms require collecting relevant indicator factors for predicting stock returns and then utilize them in making predictions [9].

The neural network algorithms used in this paper effectively fit the nonlinear law for predicting stock returns [10]. While the BPNN algorithm is a common choice for stock return prediction, it fails to consider the time series pattern of stock returns being related to previous time periods [11]. Therefore, the LSTM algorithm, which is suitable for processing time series data, is ultimately selected for predicting stock returns. The LSTM algorithm is an extension of the recurrent neural network algorithm, which can solve the problem of gradient explosion occurs when dealing with long series data. There are three gates and one neuron state in each neuron of the LSTM algorithm. The "forgetting gate" [12] is used to either forget or store past data. The "input gate`" is used to input current moment data (a collection of indicators used to predict stock returns). The "output gate" is used to output the weighted combination of the "forget gate" and "input gate" [13].

In the actual use of the LSTM algorithm for stock return prediction, many indicators need to be input, and some of them may not have a significant impact on stock returns. This can instead cause interference in the prediction process. Therefore, in this paper, to further improve the prediction accuracy of the intelligent algorithm, the PCA [14] and the LSTM algorithm are combined. The PCA is used to screen indicators that have the greatest influence on the direction of stock returns from the input indicators. After screening, these indicator data are used to make predictions. The flowchart of the optimized combined algorithm is shown in Figure 1.

☒ The indicator data of the collected training sample and the corresponding stock returns are input.

☒ The PCA screens the indicator data in the training sample [15]. Firstly, the sample data constitutes a matrix of size $i \times j$, where i denotes the number of samples and j denotes the number of indicators in each sample. Then, the data matrix is standardized to remove differences in magnitude between different indicators. Next, the correlation coefficient matrix of the standardized data matrix is calculated to determine each indicator's variance contribution rate [16], and the indicators that make the cumulative variance contribution rate exceed 85% are used as principal components. The variance contribution rate can reflect the degree to which various indicators affect stock returns during the process of change in this article. The variance contribution ratio is calculated by the following formula:

$$a_j = \frac{\lambda_j}{\sum_{k=1}^{p} \lambda_k}, \tag{1}$$

where a_j is the variance contribution of the jth indicator, λ_j and λ_k denote the eigenvalue of the jth and the kth indicator, respectively, and p is the total number of eigenvalues.

☒ The PCA-filtered indicator data are input into the LSTM algorithm for forward calculation:

$$\begin{cases} f_t = \sigma(b_f + U_f x_t + W_f h_{t-1}) \\ s_t = f_t s_{t-1} + g_t \sigma(b + U x_t + W h_{t-1}) \\ g_t = \sigma(b_g + U_g x_t + W_g h_{t-1}) \\ h_t = \tanh(s_t) q_t \\ q_t = \sigma(b_q + U_q x_t + W_q h_{t-1}) \end{cases} \tag{2}$$

where f_t indicates the forget gate output, b_f, U_f, and W_f indicate the bias term in the forget gate [17], the input term weight, and the forget gate weight, respectively, s_t indicates the cyclic gate output, b, U, and W are the corresponding weight in the cyclic gate, g_t indicates the external input gate unit, b_g, U_g, and W_g are the corresponding weight in the input gate, q_t indicates the output gate unit, b_q, U_q, and W_q are the corresponding weight in the output gate, and x_t is the sample input at the current moment.

☒ Whether the LSTM algorithm has reached the termination condition or not is determined. If it has [18], the training is stopped; if not, the weight parameters in the neurons of the LSTM algorithm are adjusted in reverse using stochastic gradient descent approach. The termination conditions include the number of training times reaching the preset threshold or the stable convergence of the difference between the results obtained from the forward calculation of the LSTM algorithm and the labels to the preset threshold [19].

The indicators screened by the PCA are input into the LSTM algorithm for forward calculation to obtain the final prediction results after undergoing repeated training in the aforementioned steps.

Table 1. Data of part of the Shanghai Stock Exchange 50 index constituent stocks.

Trade date	Opening index	Highest index	Lowest index	Closing index
2015-03-02	3071.23	3154.36	3071.23	3089.57
2015-03-03	3089.57	3198.74	3075.36	3102.54
2015-03-04	3063.57	3099.71	3053.47	3053.47
.....	.....	.....	.....	.....
2019-03-06	3087.36	3154.36	3021.38	3054.71
2019-03-07	3075.47	3178.65	3047.39	3057.98
2019-03-06	3087.36	3154.36	3021.38	3054.71
.....	.....	.....	.....	.....
2021-03-01	3026.35	3187.25	3012.47	3057.8
2021-03-02	3028.98	3127.35	3011.47	3011.47
2021-03-03	3024.78	3179.68	3024.78	3087.47
.....	.....	.....	.....	.....

3 Simulation experiments

3.1 Data sources

The constituent stocks of the Shanghai Stock Exchange 50 index were selected as the research subject from the CSMAR database [20], with a time range from January 1, 2015 to December 31, 2021. After excluding suspended and delisted stocks within this period, data from randomly selected 80 stocks were used for analysis, some of which are shown in Table 1. CSMAR database is a research-oriented and precise database in the field of economic and finance, developed by Shenzhen CSMAR Data Technology Co., Ltd. based on academic research needs, drawing on professional standards from authoritative databases such as CRSP, COMPUSTAT, TAQ, THOMSON, and combining with the actual national conditions of China. After 20 years of continuous accumulation and improvement, the CSMAR database has covered 18 series including factor research, character features, green economy, stocks, companies, overseas, information, funds, bonds, industries, economy and commodity futures. It contains over 150 databases with more than 4,000 tables and over 40,000 fields. In addition to various indicators, stock returns are also be affected by the previous time period, making it a characteristic of time series. Thus, when constructing training and testing sets after collecting stock data, it is necessary to consider the continuity of the time series. Therefore, this paper took one year as the length of the sliding window and half a year as the sliding step length of the window. One sliding window was one cycle of training and testing, then there were 13 periods. In each period, the first nine months were used as training period and the last three months were used as testing period.

The variable indicators used for input in the training and test datasets are shown in Table 2. OP_t represents the stock opening index at time t, CP_t represents the stock closing index at time t, HP_t represents the highest stock index at time t, LP_t represents the lowest stock index at time t, $True-range$ represents the true range of change between the previous and current moment of the stock, MA represents the moving average, MTM represents the momentum indicator of the stock price, and RSI represents the relative strength indicator of the stock price.

3.2 Experimental setup

The relevant parameters of the LSTM algorithm combined with PCA were set. Four hidden layers are set in the LSTM algorithm, and the number of nodes in each layer was 1,024. The activation function in the hidden layer was the sigmoid function. The stochastic gradient descent approach was used for training. The learning step length was set as 0.02. The maximum number of training was 1,000.

In addition to the PCA-LSTM algorithm, two neural network algorithms, LSTM and BPNN, were also tested. The parameter settings of the LSTM algorithm were consistent with those of the PCA-LSTM algorithm. The number of nodes in the input layer of the BPNN related parameters depended on the number of variable indicators. The number of nodes in the hidden layer was set to 512 based on experience and the orthogonal experiment method. The output layer had one node, which was used to output the predicted stock return at the next moment. Weight parameter adjustment and training session numbers during the BPNN training process were consistent with the previous two prediction algorithms.

3.3 Experimental results

Table 3 shows the results of the PCA-LSTM algorithm after performing PCA on the input indicators. The variance contribution rates for all 14 input indicators are shown in Table 3. After arranging different indicators in descending order according to their variance contribution rate and calculating the cumulative variance contribution rate, the top k variables with a cumulative variance contribution rate exceeding 85% were used as principal component variables. After calculation, it was

Table 2. Variable indexes to be input.

Serial number	Variable index	Calculation formula	Serial number	Variable index	Calculation formula
1	r_t	$\ln(CP_t) - \ln(CP_{t-1})$	8	h_{t-1}	$\ln(HP_{t-1}) - \ln(OP_{t-1})$
2	r_{t-1}	$\ln(CP_{t-1}) - \ln(CP_{t-2})$	9	l_t	$\ln(LP_t) - \ln(OP_t)$
3	c_t	$\ln(CP_t) - \ln(OP_t)$	10	l_{t-1}	$\ln(LP_{t-1}) - \ln(OP_{t-1})$
4	c_{t-1}	$\ln(CP_t) - \ln(OP_{t-1})$	11	$True - range$	/
5	o_t	$\ln(OP_t) - \ln(CP_{t-1})$	12	MA	/
6	o_{t-1}	$\ln(OP_{t-1}) - \ln(CP_{t-2})$	13	MTM	/
7	h_t	$\ln(HP_t) - \ln(OP_t)$	14	RSI	/

Table 3. The PCA results.

Serial number	Variable index	Variance contribution rate/%	Serial number	Variable index	Variance contribution rate/%
1	r_t	***15.9***	9	l_t	***4.8***
5	o_t	***13.8***	7	h_t	***4.4***
11	$True - range$	***12.6***	2	r_{t-1}	4.2
12	MA	***10.5***	4	c_{t-1}	3.5
13	MTM	***8.5***	8	h_{t-1}	3.2
14	RSI	***8.5***	10	l_{t-1}	2.5
3	c_t	***6.5***	6	o_{t-1}	1.1

found that nine indicators, i.e., No. 1, 5, 11, 12, 13, 14, 3, 9, and 7 indicators in Table 3, were the main component variables.

Figure 2 shows the convergence curves of three stock return forecasting models, BPNN, LSTM, and PCA-LSTM models, during training. It was observed in Figure 2 that the mean square error (MSE) of all three algorithm models converged during the training process, the PCA-LSTM model converged the fastest, stabilizing after about 600 times, followed by the LSTM model which stabilized after about 800 times, and the BPNN model stabilized after about 900 times. In addition, when stabilized, the MSE of the BPNN model was the largest, while the MSE of LSTM and PCA-LSTM models was relatively close; however, it was still evident that the former had a larger MSE.

As there were 13 periods and 80 stocks selected as research subjects, there is not enough space here to show the predicted and true values for all test sets. Therefore, only the real return of stock number 600029 over a period of 30 days and its corresponding predictions by the three forecasting models are shown in Figure 3. It can be observed from Figure 3 that the stock return fluctuated around zero, with the predictive value of the PCA-LSTM model being close to the true value, while the BPNN model showed significant deviation.

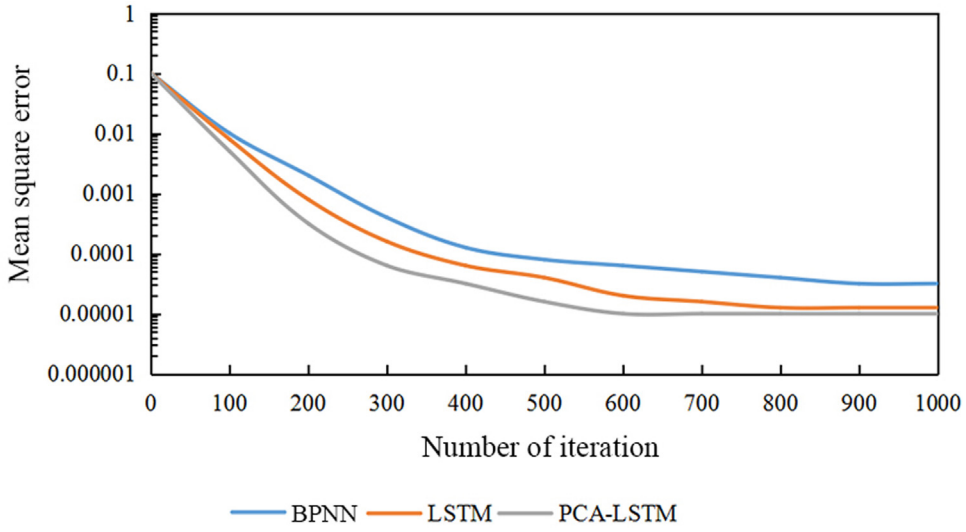

Fig. 2. The convergence curve of three prediction algorithms during training.

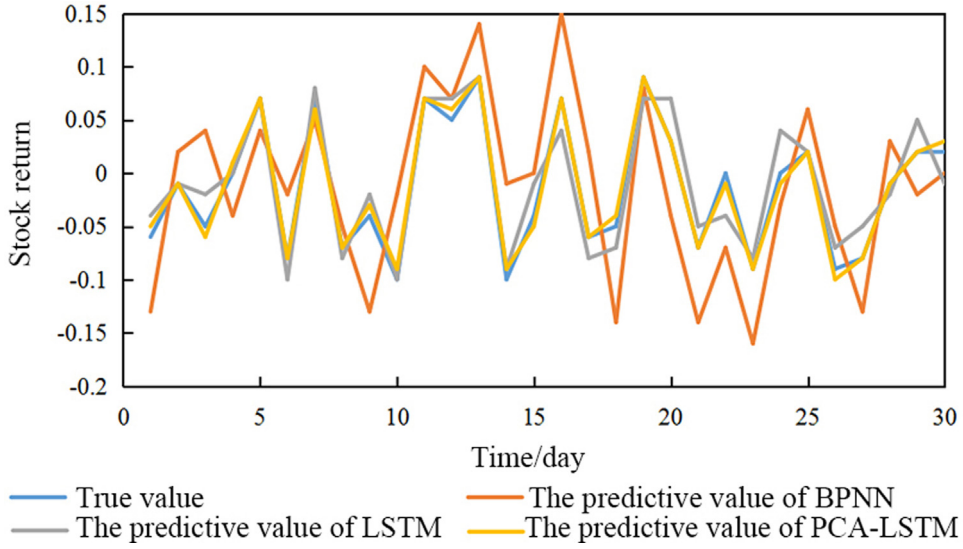

Fig. 3. The predictive value of the stock return for the No. 600029 stock over a 30-day period predicted by three models and the true value.

Table 4. Statistical results of the prediction accuracy of the prediction model.

Prediction model	MSE	MAPE	Minimum relative error	Maximum relative error
BPNN	66.125	1.2687	0.0031	0.1218
LSTM	55.871	1.0003	0.0014	0.0211
PCA-LSTM	30.231	0.9571	0.0007	0.0187

Table 4 displays the prediction accuracy of three models obtained from test results statistics. The BPNN model had an MSE of 66.125, MAPE of 1.2687, minimum relative error of 0.0031, and maximum relative error of 0.1218; the LSTM model had an MSE of 55.781, MAPE of 1.0003, minimum relative error of 0.0014, and maximum relative error of 0.0211; and the PCA-LSTM model had an MSE of 30.231, MAPE was 0.9571, minimum relative error was 0.0007, and maximum relative error of 0.0211. The MSE of the PCA-LSTM model was 30.231, the MAPE was 0.9571, the minimum relative error was 0.0007, and the maximum relative error was 0.0187. The comparison of the data in Table 4 showed that the PCA-LSTM model had the highest prediction accuracy, the LSTM model was the second, and the BPNN model was the lowest.

4 Conclusion

This paper briefly introduced the LSTM model for predicting stock returns and combined it with PCA to form the PCA-LSTM prediction model. Simulation experiments were then conducted on 80 stocks. The PCA-LSTM model was compared with the BPNN and LSTM models. The findings are shown below. (1) In the results of PCA, 9 out of 14 variable indicators were identified as the main component variables with serial numbers 1, 5, 11, 12, 13, 14, 3, 9, and 7. (2) The training errors of all three prediction algorithm models converged with the increase of training times; among them, the PCA-LSTM model had the fastest convergence rate followed by the LSTM model and the BPNN model which was slowest to converge; when convergence was stable, the BPNN model had the largest error while LSTM model had second highest error and PCA-LSTM model had smallest error. (3) In terms of the prediction accuracy of stock return, the PCA-LSTM model had the highest performance, followed by the LSTM model and then the BPNN model.

References

1. Q. Liu, Q. Yao, G. Zhao, Model averaging estimation for conditional volatility models with an application to stock market volatility forecast, J. Forecasting **39**, 841–863 (2020)
2. J. Wang, Q. Cui, X. Sun, M. He, Asian stock markets closing index forecast based on secondary decomposition, multi-factor analysis and attention-based LSTM model, Eng. Appl. Artif. Intel. **113**, 1–21 (2022)
3. S.D. Chen, Y.L. Sun, Y. Liu, Forecast of stock price fluctuation based on the perspective of volume information in stock and exchange market, China Finance Rev. Int. **8**, 297–314 (2018)
4. A. Kanwal, M.F. Lau, S.P.H. Ng, K.Y. Sim, S. Chandrasekaran, BiCuDNNLSTM-1dCNN – A hybrid deep learning-based predictive model for stock price prediction, Expert Syst. Appl. **202**(Sep.), 1–15 (2022)
5. W. Zhang, S. Zhang, S. Zhang, D. Yu, N. Huang, A multi-factor and high-order stock forecast model based on Type-2 FTS using cuckoo search and self-adaptive harmony search, Neurocomputing, **240**(May31), 13–24 (2017)
6. S. Yang, H. Guo, J. Li, CNN-GRUA-FC stock price forecast model based on multi-factor analysis, J. Adv. Comput. Intell. Intell. Inform. **26**(4 TN.157), 600–608 (2022)
7. C. Xiao, W. Xia, J. Jiang, Stock price forecast based on combined model of ARI-MA-LS-SVM, Neural Comput. Appl. **32**, 5379–5388 (2020)
8. Z.K. He, Prediction of amazon's stock price based on ARIMA, XGBoost, and LSTM models, Busin. Econ. Res. **5**, 127–136 (2022)
9. C.R. Ko, H.T. Chang, LSTM-based sentiment analysis for stock price forecast, PeerJ Comput. Sci. **7**, 1–23 (2021)
10. Q. Wang, K. Kang, Z. Zhang, D. Cao, Application of LSTM and CONV1D LSTM network in stock forecasting model, Adv. Artif. Intell. **3**, 36–43 (2021)
11. Y. Yan, D. Yang, A stock trend forecast algorithm based on deep neural networks, Sci. Programming **2021**, 1–7 (2021)
12. M. Liu, J. Ye, L. Yu, Volatility prediction via hybrid LSTM models with GARCH type parameters, Busin. Econ. Res. **5**, 37–46 (2022)
13. J. Li, T. Zhou, X. Hu, Prediction algorithm of stock holdings of hong kong-funded institutions based on optimized PCA-LSTM model, Int. J. Innov. Comput. I. **18**, 999–1008 (2022)
14. L. Sun, W. Xu, J. Liu, Two-channel attention mechanism fusion model of stock price prediction based on CNN-LSTM. ACM T, Asian Low-Reso. **20**, 1–12 (2021)
15. C. Anand, Comparison of stock price prediction models using pre-trained neural networks, J. Ubiq. Comput. Commun. Technol. **3**, 122–134 (2021)
16. D. Saravagi, S. Agrawal, M. Saravagi, Indian stock market analysis and prediction using LSTM model during COVID-19, Int. J. Eng. Syst. Model. **12**, 139–147 (2021)
17. H.K. Konan, F.A. Kouassi, M. Coulibaly, O. Asseu, Prediction of stock prices via recurrent neural networks with LSTM (long shortterm memory) architecture, Far East J. Math. Sci. **128**, 53–65 (2021)
18. X. Zhang, W. Qi, Z. Zhan, A study on machine-learning-based prediction for bitcoin's price via using LSTM and SVR, J. Phys. Conf. Ser. **1732**, 1–5 (2021)
19. L.Z. Tao, Predicting Google's stock price with LSTM model, Busin. Econ. Res. **5**,82–87 (2022)
20. N. Jing, Z. Wu, H. Wang, A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction, Expert Syst. Appl. **178**, 115019 (2021)

Cite this article as: Yanxiang Mi, Donghai Xu, Tielin Gao, Application of PCA-LSTM algorithm for financial market stock return prediction and optimization model, Int. J. Simul. Multidisci. Des. Optim. **14**, 8 (2023)