International Journal of Innovative
Computing, Information and Control
Volume **18**, Number **3**, June **2022**

ICIC International ©2022 ISSN 1349-4198
pp. 999–1008

PREDICTION ALGORITHM OF STOCK HOLDINGS OF HONG KONG-FUNDED INSTITUTIONS BASED ON OPTIMIZED PCA-LSTM MODEL

Jianming Li[1], Tingting Zhou[1] and Xiangpei Hu[2]

[1]School of Computer Science
[2]School of Management and Economics
Dalian University of Technology
No. 2, Linggong Road, Ganjingzi District, Dalian 116024, P. R. China
{ lijm; drhxp }@dlut.edu.cn; 18342207476@163.com

Received September 2021; revised December 2021

Abstract. *As the largest incremental capital in the A-share market, the trend of Hong Kong capital plays a very important guiding role in selecting industry investment opportunities and guiding market sentiment. Securities investors generally conduct quantitative research on Hong Kong capital as a whole. Through the analysis of the historical transaction data of Hong Kong-funded institutions, this paper finds that the trading characteristics of Hong Kong-funded institutions with different attributes are different, and the profit and loss and position of stocks will affect their future trading operations. In this paper, a multi-dimensional factor dataset of Hong Kong-funded institutions is constructed, principal component analysis (PCA) is used to reduce the noise of the dataset, feature extraction is carried out through long short-term memory (LSTM), etc., and finally a prediction model based on optimized PCA-LSTM is proposed. The results show that compared with the original factor dataset, the multi-dimensional factor dataset can better quantify the transaction features, and the optimized PCA-LSTM model also has a better learning effect than the original LSTM.*
Keywords: Multi-dimensional factor dataset, PCA, LSTM

1. **Introduction.** With the advent of the era of big data, the flow of Hong Kong capital and changes in returns have become the focus of attention of stock market researchers and securities institutions. After the Shanghai-Hong Kong Stock Connect policy was put forward, the overall stock price of China's A-share market has been significantly improved in reflecting the speed of market public information [1]. On the basis of empirical research, the researchers believe that Hong Kong capital is indeed a "smart money". In the short term, due to the quantitative trading of Hong Kong capital, it is easy to grasp the "buying points" and "selling points"; in the long term, Hong Kong capital has a sound investment style in stock selection and value investment [2,3].

In the research of financial big data, multi factor model has always been one of the main quantitative models. Typical models include momentum factor model [4,5], illiquidity factor model [6], heterogeneous volatility factor model [7,8], and five factor model adding investment factor and profit factor on the basis of three factor model [9,10]. In addition, a large number of research results of behavioral finance have also been applied to factor construction, and the most representative is the investor sentiment factor model [11,12].

The emergence of Hong Kong capital means that a new kind of financial factor data can be mined and analyzed. The researchers found that Hong Kong capital, as a composite factor, has alpha characteristics that traditional factors do not have, and the machine

DOI: 10.24507/ijicic.18.03.999

learning classification algorithm can be used for factor screening [13]. In the field of stock time series data prediction, a Re-LSTM for stock price movement prediction is proposed, and its performance is better than the original LSTM [14]. In terms of using LSTM to process Hong Kong capital data, the researchers constructed the Hong Kong capital position dataset and proposed a stock price fluctuation prediction algorithm based on multi factor LSTM model [15].

After in-depth study of Hong Kong-funded institutions, this paper finds that not all Hong Kong-funded institutions have a good performance in terms of income. After observing the holdings and returns of Hong Kong-funded institutions over the years, it can be found that different institutions have different trading characteristics, and their prediction and preference for stocks are different. Therefore, it has become a new research direction to study the transaction dynamics of Hong Kong-funded institutions. This paper constructs the multi-dimensional factor dataset of Hong Kong-funded institutions from different angles. Then, the shortcomings of LSTM's inability to effectively extract local features are improved, and the optimized PCA-LSTM model is used to predict transaction dynamics.

The structure of this paper is as follows: The second section briefly describes the technical part used in this paper; The third section describes the construction of multi-dimensional factor dataset in detail; The fourth section is the experimental model and results; The fifth section is the summary of this paper.

2. **Related Technology.**

2.1. **Principal components analysis (PCA).**

PCA is a dimensionality reduction statistical method. With the help of an orthogonal transformation, it transforms the original random vector whose components are related into a new random vector whose components are not related. This is expressed algebraically as transforming the covariance matrix of the original random vector into a diagonal matrix, and geometrically as transforming the original coordinate system into a new orthogonal coordinate system. Make it point to the P orthogonal directions where the sample points are scattered most.

2.2. **Long short-term memory (LSTM).**

LSTM is a special recurrent neural network (RNN), which can learn long-term dependence information. RNN has been widely used in many research fields such as language recognition and text classification [16]. RNN has the problems of gradient explosion and gradient disappearance, that is, the long-term dependence on historical data cannot be effectively solved. In order to solve these two problems, machine learning researchers have developed the LSTM model [17]. The most obvious improvement of LSTM model compared with RNN is the addition of 1 cell state C and 3 valves. The three valves are forgetting gate F, output gate O and input gate I, which solve the problem of gradient explosion and disappearance, that is, effectively deal with the redundancy of relevant information in historical data [18,19].

2.3. **Convolutional neural network (CNN).**

CNN is a special kind of multi-layer neural network. Like other neural networks, CNNs are also trained using the backpropagation algorithm. The difference is in the structure of the network. The network connection of CNN has the characteristics of local connection and parameter sharing. The use of local connections and parameter sharing reduces the number of parameters that need to be trained, and the amount of computation is gradually reduced, while CNN preserves global features by increasing the number of network layers. This paper uses one-dimensional convolution to process factors, considering the problem that the data source is financial time series data. When dealing with sequence problems, it is necessary to consider that

time passes in one direction (similar to a hidden Markov chain), that is, the feature is the result of convolution of the output at time t with the elements before t. So the experiment cannot use ordinary one-dimensional convolution, and uses causal convolution. Its calculation method is shown in Formula (1).

$$p(x) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1}) \tag{1}$$

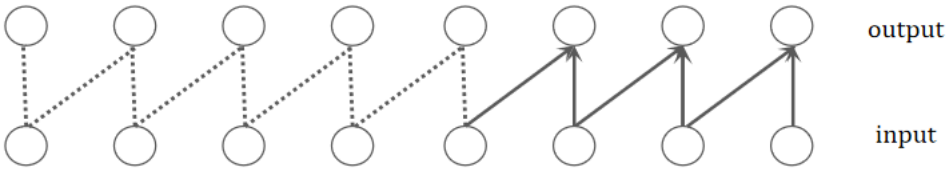

FIGURE 1. Causal convolution

3. **Construction of Multi-Dimensional Factor Dataset.** This paper selects 15 mainstream Hong Kong-funded institutions to hold. The market value accounts for about 90% of the total market value held by Hong Kong investors. A total of 16,928,910 transaction data of Hong Kong-funded institutions have been obtained. Obtain basic market data and basic financial indicator data through Tushare financial data interface package. A multi-dimensional factor dataset is constructed on the basis of the above-mentioned raw data. The following introduces the relevant information of several important dimensions and the calculation method of the relevant important factors. After the factor calculation is completed, the factor set of 15 institutions (the top 100 stocks with the most trading volume of each institution) is regarded as a data sample. In the same data sample, the input dataset is divided according to the time_steps required by the model.

3.1. **Transaction characteristics of Hong Kong-funded institutions.** This paper divides Hong Kong capital into allocation-type and transaction-type, which are represented by the factor *trade_charact*. Allocation-type tends to hold stocks for a long time, and the turnover rate is low. Institutions holding capital tend to carry out continuous inflow transactions, and the representative institutions are Standard Chartered Bank, HSBC, etc.

Transaction-type is greatly affected by market sentiment, and is easily affected by risk appetite and continues to flow in and out. Therefore, capitals are very active and have a high turnover rate. The representative institutions are China International Finance Hong Kong Securities Co., Ltd., etc. Figure 2 shows the comparison of the two types.

3.2. **Profit and loss of Hong Kong-funded institutions.** The profitability of an institution is closely related to the dynamic changes of stock market prices. Institutions judge the market outlook and make trading strategies based on their own profit and loss conditions. When in a state of loss for a period of time, it is necessary to consider whether to stop losses in time or gradually increase positions to reduce costs; while in a state of profit, analyze the possibility of the stock continuing to rise and the risk of retracement to decide whether to stop profit in time or continuously increase positions in pursuit of profit maximization. Therefore, investment institutions can find the optimal positions in the A-share market to maximize the long-term growth rate of assets and carry out capital allocation management according to the profit and loss situation, which can effectively respond to market changes, increase investor returns, and reduce investment

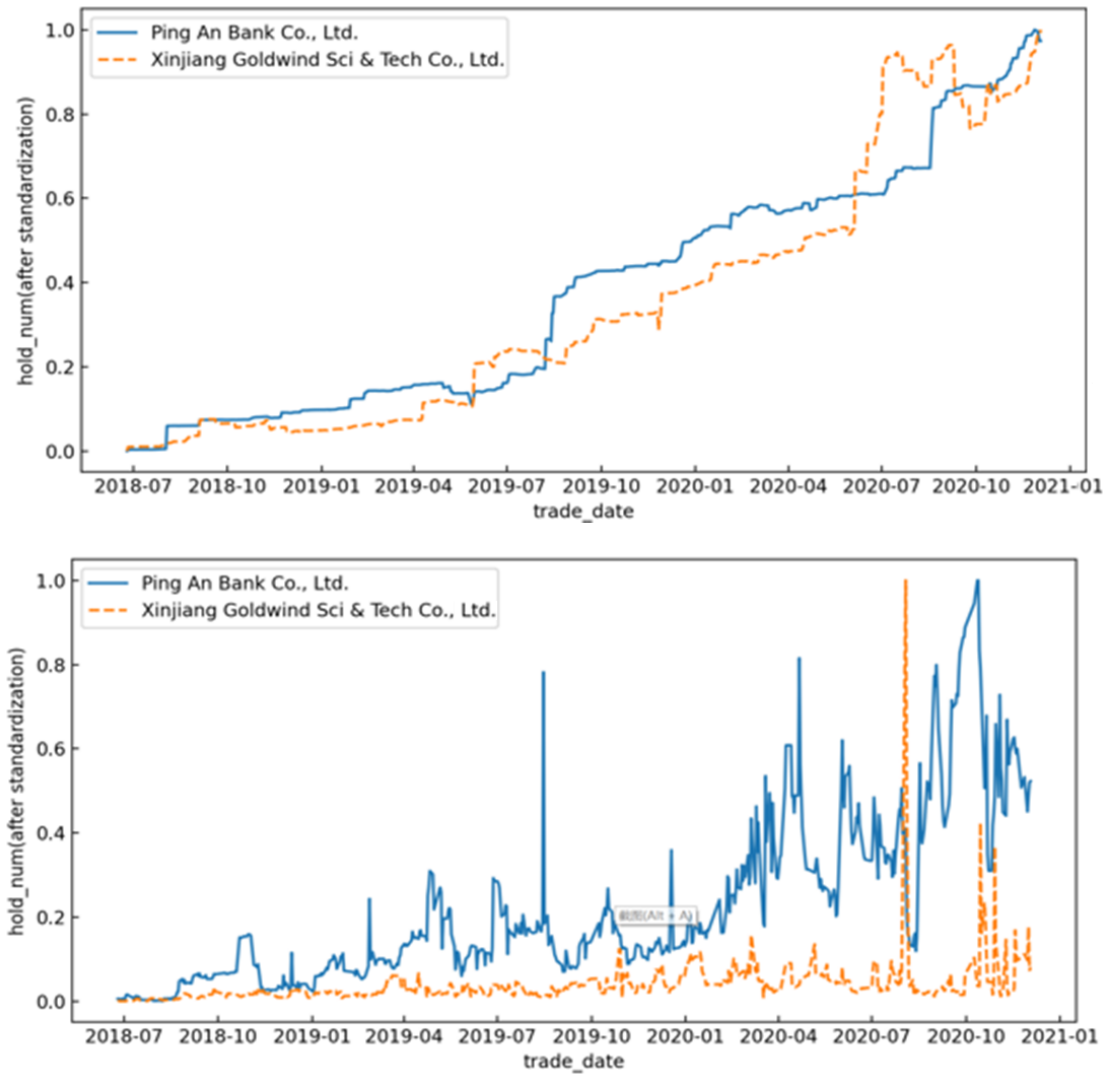

FIGURE 2. Shareholding curve of Standard Chartered Bank & China International Finance Hong Kong Securities Co., Ltd.

risks [20]. Factors that can express profit and loss include profit and loss ratio, profit and loss amount, rate of return, average daily rate of return, amount of income, etc. Formulas (2) and (3) are the most representative calculation methods for the profit and loss ratio factor, which is the ratio of the trading profit and loss to the position cost.

$$profit\_and\_loss = cost_{buy} \times profit\_loss\_ratio \tag{2}$$

$$profit\_loss\_ratio = (close - price_{cost})/price_{cost} * 100\% \tag{3}$$

3.3. **Excessive transactions of Hong Kong-funded institutions.** Market analysis points out that Shanghai-Hong Kong Stock Connect and Shenzhen-Hong Kong Stock Connect have a keener sense of A-shares. The historical data shows that the sudden continuous net buying phenomenon is often an important information for the start of a short-term rebound of A-shares. Due to the maturity of Hong Kong capital, it can always sniff out the impact of the international situation and domestic policies on economic changes, and plan the entry and return of funds in advance. The abnormal excess trading behavior of Hong Kong capital also affects stock price changes. Research data shows that when the trading volume of a single Hong Kong-funded institution or multiple major

institutions reaches 5% of the total stock trading volume, it will have a certain impact on the stock price, reaching 10% can affect the price performance, and reaching 20% can almost control the price direction [21]. Therefore, grasping the flow of historical excess capital is an important direction for market forecasting. The formula for calculating the excess_factor factor is as follows. Among them, Formula (4) is the judgment of excess inflow behavior, and Formula (5) is the judgment of excess outflow behavior. $vol_{\max}$ is the maximum number of changes in holdings in the past month.

$$(vol_{buy\_\max} - vol_{buy\_avg})/vol_{buy\_avg} > 30\% \tag{4}$$

$$(vol_{sell\_\max} - vol_{sell\_avg})/vol_{sell\_buy} <= -30\% \tag{5}$$

3.4. **Positions of Hong Kong-funded institutions.** The position level is divided into two dimensions: stock and increment. The stock shows the overall capital investment. For different types of stocks, the amount of capital allocated by the institution is different. For example, blue-chip stocks with a significant dominant position and good performance in their industry are suitable for higher trading quotas than common stocks. The factors that reflect the stock are the circulation ratio factor, market value of holdings, etc.; the increment shows the forecast for the market, the proportion of additional or reduced positions for investment. And the increment can reflect the institution's views and investment decisions on the future trend of the stock. The factors that reflect the increment include the Masukura ratio and so on. Therefore, the factor of the position dimension can show the recent preference and position adjustment direction of Hong Kong-funded institutions. Formulae (6) and (7) introduce the calculation method of the circulation ratio factor and the Masukura ratio factor.

$$circulation\_ratio = vol_{ins}/vol_{A\text{-}share} \tag{6}$$

$$Masukura\_ratio = vol_t - vol_{t-1}/vol_{t-1} \tag{7}$$

4. Prediction Algorithm Based on Optimized PCA-LSTM.

4.1. **Factor screening.** Although this paper constructs a factor dataset from multiple dimensions, some factors may bring additional financial noise to the study; the correlation between the factor data also makes the information reflected in the data overlap to a certain extent, and cannot well show the deep hidden laws; such a large amount of factor data makes the model run too much computation and increases the experimental cost. Therefore, certain screening of the obtained factor data has become the key to constructing a good dataset [22,23]. Considering the interaction between institutional trading behavior and stock prices, the factors in the multi-dimensional factor dataset are combined with stock market factors and stock financial factors. Use random forest for factor importance ranking for all factors, and select the top 20 factors for subsequent training.

4.2. **Label calculation.** Traditional LSTM is mostly used to predict specific values, but considering the actual demand, this paper believes that it is not necessary to predict the actual value, but only the growth rate in the next period of time. Therefore, it is first necessary to classify the dynamic changes of Hong Kong-funded institutions transactions. There are three types of situations: excess inflow, normal transaction, and excess outflow. Since the transaction situation of Hong Kong-funded institutions has trend characteristics, the change in transaction volume over a period of time should be considered. The calculation method is as follows.

$$label = \begin{cases} 0 & \Delta V < \delta_1 \\ 1 & \delta_1 < \Delta V < \delta_2 \\ 2 & \Delta V > \delta_2 \end{cases} \tag{8}$$

ΔV is the greatest potential for Hong Kong capital to flow into or out of A-shares. In order to make the number of categories relatively uniform, set (δ_1, δ_2) as the threshold of the trading range. The thresholds for different Hong Kong-funded institutions are different.

4.3. **Model introduction.** Figure 3 shows a schematic diagram of the model structure. Since the factor feature information of each dimension in the composition of the original dataset overlaps, in order to remove redundant information, PCA is used to reduce the dimension of the factor dataset, so that the reconstructed factors are independent. Firstly, the factor data was standardized by SPSS software, and then the Bartlett sphericity test and KMO test were carried out. So there is a significant correlation between the factors.

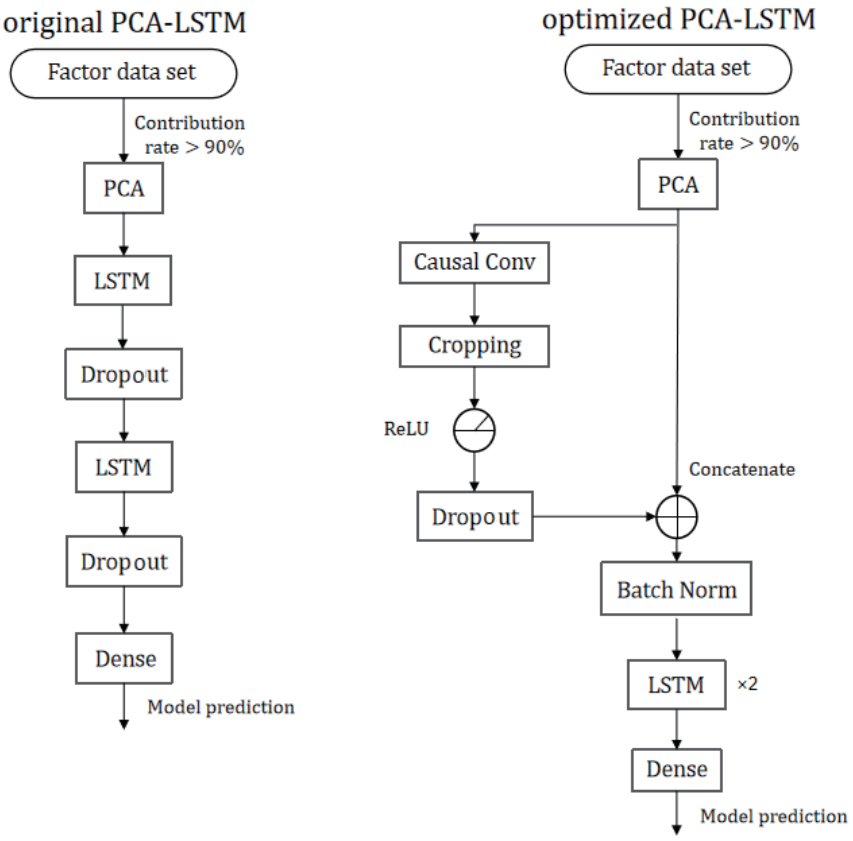

FIGURE 3. Model comparison

TABLE 1. KMO and Bartlett's test

Kaiser-Meyer-Olkin measure of sampling adequacy		0.723
Bartlett's test of sphericity	Approx. Chi-Square	406493.484
	df.	171
	Sig.	0.000

Principal components are extracted from the data, and 90% contribution rate is taken for the extracted principal components to obtain 10 principal components. In order to make the PCA interpretation more efficient and minimize the data redundancy of the model, this paper retains the parameter data with values greater than 0.4.

This paper optimizes the original PCA-LSTM model. Considering that although LSTM can obtain global time series information, the special transaction behavior of Hong Kong capital, such as capital flow that lasts for several days, or sudden capital inflow and outflow

operations is extremely important to the task. Use causal convolutional units as feature extractors to capture transactional or factorial features local to the dataset. After the convolution operation, the size of the new data after convolution is different from the size of the input data due to padding. After the convolution, the data size needs to be modified, a cropping function is constructed to change the size, and then the ReLU and dropout layers are added. Dropout will randomly disconnect input neurons with a certain probability every time the parameters are updated during the training process to prevent over-fitting. The processed information is spliced with the original information, normalized and then input into the two-layer LSTM network to extract time series features and speed up network training and convergence. The output feature data needs to go through the nonlinear transformation of the fully connected layer, and then input it to the Softmax layer to obtain the probability distribution of different categories.

4.4. **Experiment results.** The following will take Standard Chartered Bank as an example to study the dynamic changes of transactions. Different hyper parameters have a significant impact on the predictive ability of the model [24]. Table 2 shows the parameter settings of the optimized PCA-LSTM model.

Table 2. Parameter settings of the optimized PCA-LSTM model

Convolution kernel size	LSTM layers	Learning rate	Dropout	Hidden layer neurons	Classification function	Categories
5	2	0.005	0.3	128	Softmax	3

The first 70% of the factor dataset, that is, the first 25,396 data, is used as the training set, and the remaining 30% of the data is used as the prediction set. In order to verify the general validity of the model proposed in this paper, SVM, XGBoost, LSTM, TreNet (a time series trend prediction classification network in the financial field), original PCA-LSTM are used for comparative experiments under the same dataset input. The experimental results are shown in Table 3, it can be seen that all evaluation indicators of the optimized PCA-LSTM model can achieve better results, and compared with the best baseline model, the ACC is increased by 1.98%, the macro_F1 value is increased by 2.09%, and the kappa coefficient is 0.6131 (0.61∼0.80 means the results are highly consistent); from the experimental results, we can see that the model in this paper has higher classification accuracy. In addition to the SVM, the accuracy rates of the other models are all above 60%, which proves that the dataset has good separability.

Table 3. Experimental result 1

Model	ACC	Precision	Recall	Macro_F1	Kappa	Ham_distance
SVM	0.5255	0.5235	0.5263	0.5229	0.2888	0.4744
XGBoost	0.6221	0.6217	0.6223	0.6219	0.5332	0.3278
LSTM	0.6415	0.6419	0.6417	0.6415	0.5748	0.3034
TreNet	0.6573	0.6568	0.6574	0.6570	0.5859	0.2926
PCA-LSTM	0.6565	0.6550	0.6567	0.6567	0.5848	0.2934
Model	0.6771	0.6789	0.6772	0.6779	0.6131	0.2803

Perform ablation experiments on the factors to test whether each dimension factor is effective. Using the optimized PCA-LSTM model, each dimension factor that needs to be verified is eliminated in turn to test the accuracy of the model. except_1 is the transaction type factor ablation, except_2 is the profit and loss situation factor ablation, except_3 is the

TABLE 4. Experimental result 2

Model	ACC	Precision	Recall	Macro_F1	Kappa	Ham_distance
except_1	0.6534	0.6541	0.6529	0.6540	0.5829	0.2996
except_2	0.6224	0.6222	0.6231	0.6215	0.5335	0.3211
except_3	0.5406	0.5481	0.5492	0.5498	0.3245	0.4393
except_4	0.5788	0.5783	0.5785	0.5784	0.3482	0.3811

excess transaction factor ablation, and except_4 is the position situation factor ablation. According to the experimental results, it can be seen that the excess transaction factor and the position factor have the greatest impact on the model, and the accuracy is 13.65% and 9.83% lower than the full factor model. Factor ablation experiments demonstrate the effectiveness of the dataset, and more relevant factors can be mined in subsequent studies to enrich our multi-dimensional factor dataset.

5. **Conclusion.** This paper uses the optimized PCA-LSTM model to dynamically predict the transactions of Hong Kong-funded institutions, introduces basic technical indicators of stocks, builds a multi-dimensional factor dataset, and verifies the validity of the dataset. The PCA is added to the model to use the extracted principal components and the calculated data as a new training sample set. This method improves the sample quality and eliminates the correlation of the input features. While improving the simplicity of the input data, it also simplifies the overall network structure. The model uses feature extraction tools such as CNN and LSTM to extract various effective features of datasets such as local change information and time series information. By comparing the experimental results, it is found that the model proposed in this paper predicts better than other baseline models. Not only the prediction accuracy is improved, but the prediction effect is also more stable.

As an incremental factor affecting stock price changes, Hong Kong capital can be used for stock price prediction to better determine buying and selling points for trading strategies. How to combine the behavior of Hong Kong capital that affects stock price changes with the two elements of the stock market will be the focus of the next research.

REFERENCES

[1] D. Lv, Q. Ruan and X. Wan, An empirical study on the impact of Shanghai-Hong Kong Stock Connect on the price discovery speed of Shanghai underlying stocks, *Business Research*, no.7, pp.34-43, 2017.

[2] K. Zhong, C. Sun, Y. Wang and H. Wang, The opening of capital markets and the heterogeneous fluctuation of stock prices: Empirical evidence from the "Shanghai-Hong Kong Stock Connect", *Financial Research*, no.7, pp.174-192, 2018.

[3] L. Cai and S. Liu, Is "northbound capital" really smart money – Empirical research based on A-share listed companies, *China Prices*, vol.12, 2020.

[4] M. M. Carhart, On persistence in mutual fund performance, *Journal of Finance*, vol.52, no.1, pp.57-82, 1997.

[5] F. Li, F. Jiang and H. Yang, The impact of investors' rational characteristics on momentum effect: Evidence based on China's A-share market, *Macroeconomic Research*, no.11, pp.112-122, 2019.

[6] Y. Amihud, Illiquidity and stock returns: Cross-section and time-series effects, *Journal of Financial Markets*, vol.5, no.1, pp.31-56, 2002.

[7] A. Ang, R. J. Hodrick, Y. Xing and X. Zhang, The cross-section of volatility and expected returns, *Journal of Finance*, vol.61, no.1, pp.259-299, 2006.

[8] L. Zhou and Y. Wang, Research on the low-risk anomaly of China's stock market, *Financial Theory and Practice*, no.3, pp.90-96, 2020.

[9] E. F. Fama and K. R. French, Incremental variables and the investment opportunity set, *Journal of Financial Economics*, vol.117, no.3, pp.470-488, 2015.

[10] S. Zhao, H. Yan and K. Zhang, Is the Fama-French five-factor model better than the three-factor model: Empirical evidence from China's A-share market, *Nankai Economic Research*, no.2, pp.41-59, 2016.
[11] A. Sun, M. Lachanski and F. J. Fabozzi, Trade the Tweet: Social media text mining and sparse matrix factorization for stock market prediction, *International Review of Financial Analysis*, no.48, pp.272-281, 2016.
[12] C. Yu, Y. Gong, F. Wang et al., Stock trend prediction based on text price fusion model, *Data Analysis and Knowledge Discovery*, vol.2, no.12, pp.33-42, 2018.
[13] S. He, *Research on Multi-Factor Stock Selection Strategy Based on Hong Kong-Funded Position Trading Behavior*, Master Thesis, Shanghai Normal University, 2021.
[14] Y. Zhang, S. H. Tan, J. Yang, T. Kim and J. Bae, Stock price movement prediction based on re-extract feature LSTM, *ICIC Express Letters*, vol.16, no.2, pp.187-194, 2022.
[15] Y. Zhang, *Stock Forecasting System Based on LSTM Analysis of Hong Kong-Funded'S Multi-Factor*, Master Thesis, Dalian University of Technology, 2020.
[16] H. Han, G. Liu, T. Sun et al., Text sentiment analysis based on multi attention level neural network, *Computer Engineering and Application*, vol.56, no.10, pp.100-105, 2020.
[17] D. Pei and M. Zhu, Stock price prediction based on multi factor and multi variable long term and short term memory network, *Computer System Applications*, vol.28, no.8, pp.30-38, 2019.
[18] A. Zeng and W. Nie, Stock recommendation system based on deep bidirectional LSTM, *Computer Science*, vol.46, no.10, pp.84-89, 2019.
[19] J. Ren, J. Wang, C. Wang et al., Stock forecasting system based on ELSTM-L model, *Statistics and Decision*, vol.35, no.21, pp.160-164, 2019.
[20] W. Zhang, Risk analysis and profit and loss management of stock investment, *Economist*, no.10, pp.112-113, 2008.
[21] X. Zhao, Layout routine of northward capital, *Manage Money Matters*, no.5, pp.52-53, 2017.
[22] C. Cortes and V. Vapnik, Support vector network, *Machine Learning*, vol.20, no.3, pp.273-297, 1995.
[23] T. K. Ho, The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol.20, no.8, pp.832-844, 1998.
[24] T. Xu, *Research on Stock Price Rise and Fall Prediction Based on LSTM Neural Network Model*, Master Thesis, Shanghai Normal University, Shanghai, 2019.

Author Biography

Jianming Li received the bachelor's degree in ship engineering from Dalian University of Technology, China, 1999; the M.Sc. degree in computer application technology from Dalian University of Technology, China, 2002; the Ph.D. degree in computer application technology from Dalian University of Technology, China, 2007.

Dr. Li is currently a full-time associate professor at the Dalian University of Technology, China. His main research interests include the machine learning, classification and prediction algorithms of deep learning, software automation, and quantitative analysis and strategy research in the financial field. He has published over 50 papers in journals and conferences.

Tingting Zhou obtained a bachelor's degree in engineering and a bachelor's degree in Japanese literature, majoring in computer science and technology, from September 2015 to June 2020, Dalian University of Technology.

Ms. Zhou is currently studying for a master's degree in Dalian University of Technology. Her main research field includes machine learning, classification and prediction algorithms of deep learning, and quantitative analysis and strategy research in the financial field.

Xiangpei Hu received his BS (1983), MS (1987) and Ph.D. Degree (1996) from Harbin Institute of Technology, China, respectively. He is a Professor of Management Science at Dalian University of Technology, China, "Distinguished Young Scholars" of National Natural Science Foundation of China (NNSFC), "Chang-jiang Scholars Distinguished Professor" of Ministry of Education (MOE) of China.

His research and teaching interests include electronic commerce, supply chain and logistics management, intelligent operations research and the real-time optimization control for dynamic systems. He has published over 200 scholarly papers in refereed journals.