Stock Price Prediction Based on Shareholding Network Topology and LSTM Model

Yu Xu[1a] ,Dan Wang[2]* ,Jianshu Hao[1b]

Corresponding author Email: h_wd@hit.edu.cn
[a]xuyu20020131@163.com , [b]20b910016@stu.hit.edu.cn

[1]School of Management,Harbin Institute of Technology, Harbin, 150000, China
[2]School of Civil Engineering,Harbin Institute of Technology, Harbin, 150000, China

Abstract In the stock market, stock price forecasting is important for investors' decisions. Due to the nonlinearity, high noise, and strong temporal variability of stock price data, traditional methods have shortcomings in forecasting tasks**.** Since based on the structure and function of recurrent neural networks, the Long Short Term Memory (LSTM) model has stronger analytical capabilities for this type of time series data. In stock prices prediction, it is important to consider external factors as well as network structures within shareholdings. In this paper,the directed holding network will be vectorized using the LINE algorithm, and four A-share listed companies will be analyzed and compared using three distinct stock price datasets. The study indicates that companies' stock prices are significantly impacted by the relationships between shareholdings and the fluctuations in related company stock prices from the perspective of network topology, leading to risk spreads within the shareholding network. Additionally, category characteristics can serve as effective assessment features.

Keywords: network topology, LINE, Long Short-Term Memory (LSTM) model, stock price prediction, risk transfer

1 Introduction

Since the advent of stock markets, accurate prediction of stock prices has been a primary concern for investors. However, traditional econometric methods like ARIMA and GARCH models were found to be inadequate in capturing the unique features of stock price data and thus had lower accuracy. The introduction of machine learning algorithms, such as recurrent neural networks (RNN) and support vector machines (SVM) [1], as well as deep neural network models, has garnered significant interest, given their ability to predict stock prices effectively. Meanwhile, researchers are exploring factors that influence stock prices, such as investor sentiment and news publications[2-3].

Studies indicate that RNN models have limitations in accurately predicting stock prices, owing to problems of gradient disappearance and explosion[4]. In response, LSTM models were introduced. The combination of LSTM with other models has resulted in increased accuracy of predictions. For example, Peng utilized LSTM and Transformer models to predict bank stock prices in China's A-share market[5], while Liwei Tian's LSTM-BO-LightGBM model has been shown to outperform LSTM[6], and Wen Y's model used PCA and LSTM, resulting in more accurate predictions of stock price fluctuations compared to traditional models[7].

ICPDI 2023, September 01-03, Chongqing, People's Republic of China

DOI 10.4108/eai.1-9-2023.2338695

Against the backdrop of global economic integration, companies have developed increasingly interconnected relationships with each other, with shareholding relationships representing the most prevalent form of association. In fact, studies have already been conducted wherein credit risk contagion among listed companies has been predicted and validated through the use of shareholding networks [8]. Consequently, it is reasonable to assume that the correlation between stocks is an essential factor to consider when forecasting stock price trends. Unfortunately, prior research on stock price trends neglects the commercial relationships existing between companies such as shareholding, cooperation, or supply-demand relationships.

The graph embedding method mentioned in this paper is a technique for representing real-world graph structure information as vectors or tensors that can be input into machine learning models[9]. The Deepwalk[10] algorithm was the first proposed graph embedding method. However, deepwalk is only suitable for undirected graphs, while shareholding networks have directed relationships between companies. To address such shortcoming, this study introduced the LINE algorithm to vectorize the shareholding network information.

In this paper, we proposed a novel methodology that utilizes graph embedding of shareholding networks to transform the valuable information contained within the structure of the network into a set of representative features that are apt for analysis. By integrating this technique with LSTM neural networks, the proposed approach predicts fluctuation trends in stock prices with enhanced levels of accuracy. We scrutinize whether risks stemming from stock price volatility can be transmitted through the shareholding network and undertake this research with the aim of verifying that considering shareholding network relationships in stock price prediction can lead to an improved level of prediction accuracy.

2 Methodology

2.1 LINE

The LINE model [11] constructs a neighborhood using the breadth-first search (BFS) method. This model's unique feature is that it identifies first-order proximity and second-order proximity. The LINE algorithm models nodes that possess first-order and second-order proximity relationships separately. Two embeddings are obtained through the minimization of the KL-divergence between the probability and empirical distributions. Subsequently, two separate embedding vectors generated by different objective functions are linked to each vertex, better representing the input graph. To provide a more complete network description, LINE combines the first-order and second-order proximity in the graph structure. It can be applied to directed, undirected, and weighted graphs as well [12]. Based on the merits of LINE, we will use it to vectorize the holding network in this paper.

2.2 Long Short-Term Memory model (LSTM)

The LSTM[13] represents a novel deep learning architecture. RNNs are characterized by a chain-like structure that incorporates a repeated neural network module. Similarly, the LSTM architecture is based on the RNN structure, but the repeated module has a distinct composition. Specifically, LSTM comprises four interrelated layers, including three *Sigmoid* and one *Tanh*

layer, which interact in a particular fashion. Notably, the LSTM design features three gate structures that are responsible for managing and regulating the “cell”, namely the forget gate, the input gate, and the output gate. LSTM provides improved handling of longer intervals and delays in the time series, and successfully addresses the issue of gradient disappearance. According to the existing studies and the analysis of the advantages of LSTM, we use LSTM for the time series of stock price prediction in this paper.

3 Experiment

3.1 Data Selection and Pre-processing

The primary source of data utilized in this study was obtained from CSMAR. The company's featured dataset "*content*" is comprised of essential characteristics drawn from 746 companies' fundamental information. This dataset constitutes eight features, including the registered capital and asset-liability ratio of the listed companies. Another dataset "*cite*" encompasses the holding relationships among the companies. This paper will utilize the stock price datasets of four Chinese A-share listed companies: Huaneng International Power Co., LTD (600011), Poly Development Holding Group (600048), ICBC (601398), and Zijin Mining Group Co., Ltd (601899). The dataset will cover the period (from March 11, 2019 to March 3, 2023) and will undergo standardization processing to eliminate possible index dimensionality effects on prediction [6]. Huaneng International and Zijin Mining are categorized identically. The selection of these two companies is intended to increase the dependability of the outcomes of the same category. Additionally, the stock price datasets of two companies from different categories (Poly Developments and ICBC) were selected for comparative experiments.

Each listed companies possess three different datasets α, β, γ. Dataset α only consists of seven features such as the opening price and Shanghai Composite Index. β dataset includes category feature 1 in addition to α and γ dataset includes category feature 2, which means that the stock price dataset includes ownership network structural information.

3.2 Model Features

The LSTM model will use basic financial indicators, specifically including opening price, highest price, lowest price, Shanghai Stock Exchange Composite Index(SSEC), categorical features to predict the closing price of the company.

The Shanghai Stock Exchange Composite Index is a comprehensive representation of the stock market environment. Consequently, a surge in the index is indicative of a favorable development trend for the stock market as a whole. After adding the shareholding network structure and performing the clustering operation, different listed companies will be assigned to different companies under the same category, the companies within the category will not only be influenced by the intra-class environment, but also may be influenced by the large environment outside the class. The incorporation of SSEC enables us to account for the influence of the stock market environment on stock prices.

Categorical features, also defined as trends within the same category, are employed in the LSTM model to predict the closing price. K-means clustering is applied to eight fundamental infor-

mation variables of 746 companies, and one category with high intra-category similarity is selected as a target for prediction. A target company is selected from the group, and its corresponding companies with similarities are identified, and fluctuations in their stock prices are calculated. The categorical feature is formed based on the trend of a given day. If the upward trend exceeds 60%, the feature value is set to 1. Similarly, if the downward trend crosses 60%, it is set to -1. In other cases, it is set to 0. This approach results in the formation of a new assessment feature {-1, 0, 1}, which is referred to as "trends within the same category". A second "trend within the same category" has been constructed in the content dataset by adding an embedding containing the shareholding network information.

3.3 Model Procedure

The study involved incorporating a shareholding network structure into the LSTM model. The LINE method was used to vectorize the shareholding network structure and add it to the original company features. By clustering similar trends, same-class trends were generated and treated as new evaluation features that were subsequently added to the LSTM model. The stock price prediction accuracy was compared before and after including the new features. Figure 1 shows the whole idea of experimental modeling.

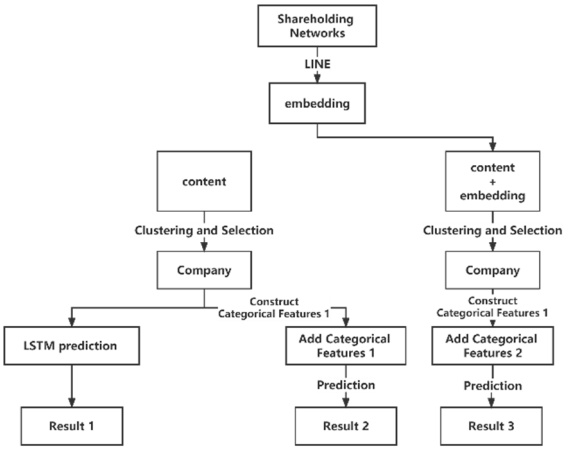

Fig. 1. Modeling Procedure

where embedding refers to the vectorization of the shareholding network structure, which contains information about the shareholding relationship between enterprises.

4 Result

Given that algorithms in the R language exhibit a clearer representation of interclass homogeneity and interclass heterogeneity [13], this study employs the *k-means* clustering algorithm in R language to classify companies. The validity of results will be measured using the ratio of the sum of squared intergroup distances to the total sum of squared distances. A higher score would imply greater distance between categories, indicating more effective clustering outcomes. The

first clustering situation yielded an indicator of 77.7%, while the latter scored 77.2%, both of which meet the standard of validity for clustering results.

In order to verify the effectiveness of vectorizing the shareholding structure using the LINE method, in this process we judge it by observing how much the loss value decreases. The results show that two loss function, in the first-order and second-order[11], with KL-divergence are trained and the loss decreases, and the vectorization of shareholding structure is effective. Figure 2 shows the experimental results of KL-divergence.

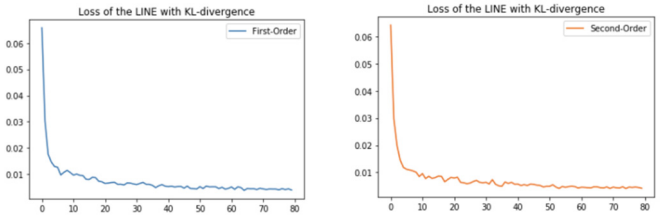

Fig. 2. Loss of the LINE with KL-Divergence

To evaluate the fitting and generalization capabilities of the aforementioned LSTM model, we employed mean squared error as the loss function in training. This result shows that the LSTM model's training situation was normal, and the training model was effective with strong generalization capabilities. To highlight these findings, we present the loss values of the actual and predicted values of α, β, γ datasets from ICBC(601398). The images from left to right are the results of datasets α, β, γ.

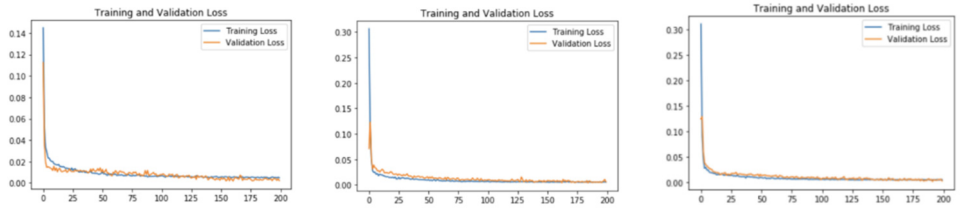

Fig. 3. Training and validation loss of ICBC's dataset

Following the confirmation of the LSTM model's efficacy, we selected the opening price, highest price, lowest price, and other pertinent features from the test sets of each company's α, β, γ datasets as inputs to the network for iterative prediction. Then we examined the fitting between the predicted and actual values, selected and presented three companies' fitting outcomes for different datasets in Figures 3. The images from left to right are the results of datasets α, β, γ.

Fig. 4. Stock price forecasts of Huaneng International's datasets

Fig. 5. Stock price forecasts of Poly Development's datasets

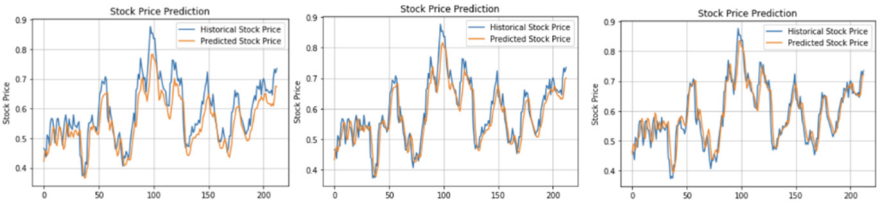

Fig. 6. Stock price forecasts of Zijin Mining's datasets

Table 1. Comparison of prediction results of different datasets

Company	Dataset	R^2	MAE/%	RMSE/%
Zijin Mining	**α**	0.818	3.6	4.3
	β	0.841	3.4	4.0
	γ	0.871	3.0	3.6
ICBC	**α**	0.790	5.1	5.9
	β	0.815	4.4	5.5
	γ	0.840	4.0	5.1
Poly Development	**α**	0.748	4.5	6.0
	β	0.764	4.4	5.8
	γ	0.804	4.1	5.5
Huaneng International	**α**	0.801	3.6	4.6
	β	0.822	3.3	4.3
	γ	0.850	3.0	3.9

Figures 4 to 6 show the results of three different companies on different data sets. Table 1 demonstrates that the LSTM model without category features produced the lowest R2, MAE and RMSE values in predicting the four companies. However, when dataset β, containing category feature 1, was used in predictive analysis, there was a slight improvement in R2, MAE and RMSE. On average, R2 values improved by 2.3%, suggesting that companies belonging to the same category exhibit similar features, and the changes in their stock prices correspond to the changes of other companies in the same category. The empirical analysis indicates that ICBC was grouped with the Agricultural Bank of China (601288), China Construction Bank (601939), and the Bank of China (601988) during the first clustering. The stock price data between March 9, 2020, and June 19, 2020, was selected and compared for changes. Over 80% of the data of the four banks' price changes demonstrated consistency. The findings indicate that over 80% of the data of the four banks' price changes demonstrated consistency. Nevertheless, some results have less distinct differences and similarities mainly because this clustering is based on the companies' characteristics. Some companies shared greater similarities, but because they are in different industries or have no business relationship with each other, they will not be influenced

by other similar enterprises. Therefore, adding category feature 1 did not significantly improve the results.

Finally, we conducted predictive analysis on the dataset γ with category feature 2. According to the results, this dataset achieved the best performance on R2, MAE, and RMSE compared to the other two datasets. R2 improved by 5.3%, 5%, 5.6%, and 4.9% respectively, accompanied by a notable reduction in MAE and RMSE values compared to the α dataset. This indicates that adding category feature 2 as a feature variable to the model can improve the prediction accuracy to a certain extent.

The shareholding network comprises various directed relationships among different nodes. A company with abundant funds and significant influence may hold shares in other companies and concurrently be owned by multiple enterprises. At this stage, the company acts as a "context" [14] and may influence the companies linked to it to varying extents. If uncontrollable circumstances cause a decline in the company's share price, the negative impact will propagate to other companies through the connections in the shareholding network. More developed and mature companies can reduce the risk to their businesses resulting from this by relying on their resilience[15]. Conversely, small companies may struggle to undertake commercial activities effectively or may face insolvency due to factors such as stock sales or investment losses. Based on the experimental outcomes and the previously mentioned explanation, a company's share price may be affected by its own investments or the stock price fluctuations of the businesses in which it has invested. The integration of shareholding network data to the model can enhance the precision of share price prognostications to some extent.

5 Conclusion

The prediction of the stock price data for four A-share listed companies confirms that incorprating shareholding network structure data enhances the precision of stock price prediction while introducing the "similar trend" as a categorical feature. The paper draws its conclusions on the forecast outcomes and analysis of three separate datasets using the LINE and LSTM model:

1. The LINE algorithm effectively converts directed shareholding networks into vectors by representing information between nodes in the network.
2. Incorporating category features obtained by transforming shareholding network information enhances the accuracy of the LSTM model for stock price forecasting. Taking the aforementioned four selected listed enterprises as an example, the R^2 value improved by an average of 5.2%. This indicates that changes in holding and related companies' stock prices influence the stock price of a company. These findings corroborate that adding the shareholding network structure to the forecast model can greatly improve the accuracy of the stock price prediction and the effectiveness of the newly introduced "similar trend" evaluation feature.
3. Fluctuations in stock prices pose a risk, which tends to spread through the shareholding network. Combining shareholding network structure with stock price fluctuations introduces categorical features. As the successful introduction of categorical features demonstrates its effectiveness, it also indirectly proves the close association between shareholding networks and stock price changes.

Investors may receive limited guidance from knowing only the recent trends and prices of individual stocks. To maintain stable returns, it is essential to comprehend other related stocks' trends and use their recent price fluctuations as a reference to improve the overall prediction's accuracy.

In addition, our study shows that the fluctuation in stock prices can spread throughout the holding network, with resulting effects on both the holding companies and their affiliated firms. Future studies can explore several areas, including providing appropriate warnings for relevant firms through risk propagation in the holding network. Other research areas include investigating whether the intensity of stock price risk varies based on holding relationship complexity; whether a corporation's resiliency impacts stock price risk transmission; and determining whether additional optimization of the holding network's structure using new graph neural network methods leads to more precise predictions. Researchers can apply the methods and models used in this paper in other context, for example, at present, China government has paid more attention to the supply chain field, many scholars also began to carry out relevant research in this field, for the inspiration brought to us by this article, we can examine the risk propagation and prediction in complicated supply chain landscapes. These avenues necessitate comprehensive research and investigation.

References

[1] Ding Z. Application of support vector machine regression in stock price forecasting[C]//Business, Economics, Financial Sciences, and Management. Springer Berlin Heidelberg, 2012: 359-365.DOI: 10.1007/978-3-642-27966-9_49.

[2] Jin Z, Yang Y, Liu Y. Stock closing price prediction based on sentiment analysis and LSTM[J]. Neural Computing and Applications, 2020, 32: 9713-9729.https://link.springer.com/article/10.1007/s00521-019-04504-2.

[3] Mohan S, Mullapudi S, Sammeta S, et al. Stock price prediction using news sentiment analysis[C]//2019 IEEE fifth international conference on big data computing service and applications (BigDataService). IEEE, 2019: 205-208.DOI: 10.1109/BigDataService.2019.00035.

[4] Salehinejad H, Sankar S, Barfett J, et al. Recent advances in recurrent neural networks[J]. arXiv preprint arXiv:1801.01078, 2017.https://doi.org/10.48550/arXiv.1801.01078.

[5] Peng Z Y, Guo P C. A data organization method for LSTM and transformer when predicting Chinese banking stock prices[J]. Discrete Dynamics in Nature and Society, 2022, 2022: 1-8.https://doi.org/10.1155/2022/7119678

[6] Tian L, Feng L, Yang L, et al. Stock price prediction based on LSTM and LightGBM hybrid model[J]. The Journal of Supercomputing, 2022, 78(9): 11768-11793.https://link.springer.com/article/10.1007/s11227-022-04326-5

[7] Wen Y, Lin P, Nie X. Research of stock price prediction based on PCA-LSTM model[C]//IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2020, 790(1): 012109.DOI 10.1088/1757-899X/790/1/012109

[8] Zhang W, Yan S, Li J, et al. Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data[J]. Transportation Research Part E: Logistics and Transportation Review, 2022, 158: 102611.https://doi.org/10.1016/j.tre.2022.102611

[9] Cai H, Zheng V W, Chang K C C. A comprehensive survey of graph embedding: Problems, techniques, and applications[J]. IEEE transactions on knowledge and data engineering, 2018, 30(9): 1616-1637.DOI: 10.1109/TKDE.2018.2807452

[10] Perozzi，B，Al-Rfou R，Skiena S .Deepwalk：Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining，New York，2014：701-710．https://doi.org/10.1145/2623330.2623732

[11] Tang J,Qu M,Wang M, et al. Line: Large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015: 1067-1077.https://doi.org/10.1145/2736277.2741093

[12] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey[J]. Knowledge-Based Systems, 2018, 151: 78-94.https://doi.org/10.1016/j.knosys.2018.03.022

[13] Ihaka R, Gentleman R. R: a language for data analysis and graphics[J]. Journal of computational and graphical statistics, 1996, 5(3): 299-314.DOI: 10.1080/10618600.1996.10474713

[14] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.https://doi.org/10.1162/neco_a_01199

[15] Ribeiro L F R, Saverese P H P, Figueiredo D R. struc2vec: Learning node representations from structural identity[C]//Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017: 385-394.https://doi.org/10.1145/3097983.3098061