Incorporating Corporation Relationship via Graph Convolutional Neural Networks for Stock Price Prediction

Yingmei Chen
School of Data Science
Fundan University, China
17210980030@fudan.edu.cn

Zhongyu Wei*
School of Data Science
Fundan University, China
zywei@fudan.edu.cn

ABSTRACT

In this paper, we propose to incorporate information of related corporations of a target company for its stock price prediction. We first construct a graph including all involved corporations based on investment facts from real market and learn a distributed representation for each corporation via node embedding methods applied on the graph. Two approaches are then explored to utilize information of related corporations based on a pipeline model and a joint model via graph convolutional neural networks respectively. Experiments on the data collected from stock market in Mainland China show that the representation learned from our model is able to capture relationships between corporations, and prediction models incorporating related corporations' information are able to make more accurate predictions on stock market.

CCS CONCEPTS

• **Networks → Network algorithms**; • **Social and professional topics → Economic impact**;

KEYWORDS

Node Embedding; Corporation Similarity; Graph Convolutional Neural Networks; Stock Price Prediction

ACM Reference Format:
Yingmei Chen and Zhongyu Wei. 2018. Incorporating Corporation Relationship via Graph Convolutional Neural Networks for Stock Price Prediction. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269269

1 INTRODUCTION

The past ten years witness the rising of using machine learning approaches for the automatic prediction of the stock market. Some researcher utilizes time series information such as historical price [6]. Other researchers dig into news information to identify indicative features, including bags-of-words, noun phrases, named entities [8] and sentiment information [10]. Beside these basic textual features, Ding et al. [1] explore to model event from news via deep learning approaches. Although some improvement has been made in terms of predicting accuracy, many interpretable factors from financial market are largely ignored, e.g., connection among corporations.

With the development of financial market, corporations are connected with each other broadly via various relationships. The Efficient Market Hypothesis (EMH) implies that financial market is informationally efficient [2]. Therefore, it is natural to believe that the change of the stock price of a target corporation would be affected by corporations that are related. In order to take this issue into consideration, we explore to model corporation relationships based on information from the real market and incorporate such information for better stock price prediction.

There are two major challenges: it is non-trivial to model the relationship between corporations; it is difficult to integrate corporation relationship into existing prediction model. To tackle these two challenges, we build a graph for corporations via a self-constructed dataset consisting of financial investment fact and propose two methods to incorporate the information of related corporations for the stock price movement prediction of a target company:

- A pipeline prediction model integrating corporation relationship via node embedding similarity. Several node embedding methods (DeepWalk [7], LINE [9] and node2vec [4]) are utilized to learn the distributed representations for corporations. And the closeness scores among corporations are computed using these representations. We then select top-n related corporations in terms of closeness scores for the target company and combine the features of involved corporations together for its stock price prediction.
- A joint prediction model based on graph convolutional networks (GCN). In order to incorporate information of more companies and companies without direct connection, we utilize graph convolutional networks [5] to use information of the whole network.

Experiments on financial datasets of listed companies in China A shares show that incorporating information from related companies is able to improve the prediction accuracy of the target company's stock price movement. Besides, further analysis also reveals that representation learned for each corporation can reflect the relationship between companies in real-world.

2 MODELS

The task of automatic prediction of stock price can be considered as a binary classification problem. Given Y_i as the prediction target (positive or negative) in day t, we use two kinds of information for the target corporation X_i:

- **Historical information**: looking back d days, we can form a feature vector as $X_{i,t} = (x_{i,t-d}, ..., x_{i,t-2}, x_{i,t-1})$ from day $t - d$

[*] Corresponding author

CIKM '18, October 22–26, 2018, Torino, Italy
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3269269

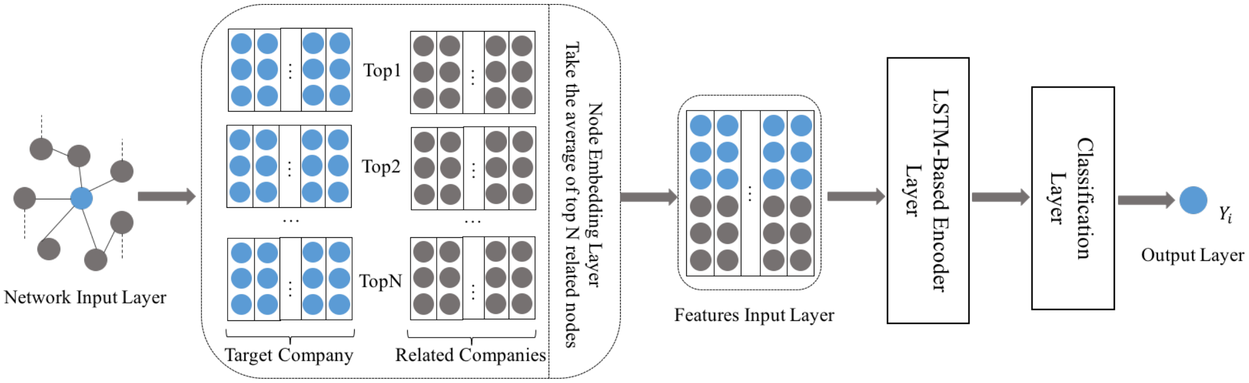

Figure 1: The pipeline prediction model

to $t-1$. Each $x_{i,t-j}, (j = 1, 2, ..., d)$ contains multiple historical indicators of the stock, such as price,volume, etc.

- **Corporation relationship information**: we use the adjacency matrix A to represent the relationship among companies. Each row stands for a company and each entry in the matrix indicates the relationship between two corresponding corporations.

2.1 Historical Information Encoder

The Recurrent neural network(RNN) is proved to be effective for processing sequential data. Following [6], we also use a Long Short Term Memory (LSTM) to encode the historical features for corporations. In our case, $X \in \mathbb{R}^{d \times k}$is an input vector, $W \in \mathbb{R}^{k \times d}$ is the weight matrix, $b \in \mathbb{R}^{k}$ is the bias vector, and$f = tanh$ is the activation function.

2.2 Construction of Corporation Network

We build a graph for corporations based on financial investment fact. More specifically, the data consists of the listed companies and their top 10 stockholders in 29th April 2017 collected from WIND[1]. There are 3,024 listed companies in total, resulting in a graph of 20,836 nodes (some companies share stockholders). In the graph, each node stands for a corporation and edges connecting nodes indicates the relationship between corporations. It is a weighted graph, and each edge represents the shareholding ratio between two nodes.

2.3 Pipeline Prediction Model

In the pipeline model, we first learn the representation for each corporation based on the graph, and select top relevant companies based on such representation. Prediction is then performed combining features of the target company and its related ones. The overall structure can be seen in Figure 1.

Three node embedding methods are utilized in this paper, including DeepWalk [7], node2vec [4] and LINE [9].

Motivated by the idea of skip-gram models, DeepWalk [7] generates sequences of nodes based on random walk. Using such sequences, it learns the node representations in a network, which is able to preserve the neighbor structures of nodes. On the basis of DeepWalk, Node2vec [4] contains a second random walk strategy to ample the neighborhood nodes, which can smoothly interpolate between breadth-first sampling(BFS) and depth-first sampling(DFS). LINE [9] is proposed for large scale network embedding, and the nodes' representations of this method can preserve the first and second order proximities.

Through the network embedding layer (operated via one of the three methods above mentioned), each node, i.e., company obtains its representation vector. With such representation, we can calculate the cosine similarity between nodes and identify the most related ones for a target node. We then take the average of features of the top N related corporations, and combine it with the feature vector of the target company. So, the input for LSTM-based encoder becomes $XX'_i = (xx'_{i,t-d}, ..., xx'_{i,t-2}, xx'_{i,t-1})$, the dimension of each $xx'_{i,t-j}, (j = 1, 2, ..., d)$ is $2 * k$. The new weight matrix and bias vector will be $W \in \mathbb{R}^{2k \times d}$ and $b \in \mathbb{R}^{2k}$. In this way, relevant companies' information is incorporated into model.

2.4 Joint Prediction Model Based on GCN

The pipeline model considers features of the top N relevant companies. Furthermore, we propose a joint prediction model based on GCN that is able to incorporate information of all the relevant companies. The overall structure can be seen in Figure 2.

The input for GCN model contains two aspects: features X and adjacency matrix A. A GCN layer can distribute the information of each node to its neighbors. Integrating the information of neighbors, each GCN layer generates a updated representation for each node. The form of l_{th} GCN layer can be represented as follows:

$$H^{l} = ReLU(\widehat{A}H^{l-1}W^{(l-1)}), \tag{1}$$

where $\widehat{A} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$. Here, $\widetilde{A} = A + I_N$ is the adjacency matrix of the undirected graph G with added self-connections, I_N is the identity matrix, $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$ and W^{l-1} is a layer-specific trainable weight matrix. In our case, we use $ReLU(\cdot)$ as the activation function.

In this paper, we build a three-layer GCN model for stock price movements prediction. And it meets the following form:

$$Y = softmax(\widehat{A}ReLU(\widehat{A}ReLU(\widehat{A}X'W^{(0)})W^{(1)})W^{(2)}), \tag{2}$$

[1] WIND is one of the leading financial data providers in China. http://www.wind.com.cn/

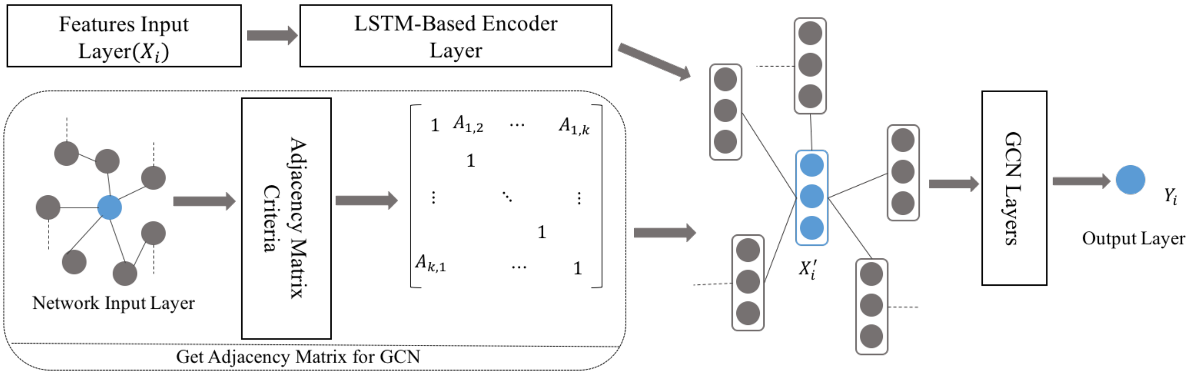

Figure 2: The joint prediction model based on GCN

where $W^{(0)} \in \mathbb{R}^{C \times H}$ is an input-to-hidden weight matrix for a hidden layer with H feature maps. $W^{(1)} \in \mathbb{R}^{H \times H}$ is a hidden-to-hidden weight matrix and $W^{(2)} \in \mathbb{R}^{H \times F}$ is a hidden-to-output weight matrix. The softmax activation function, defined as $softmax(x_i) = \frac{1}{\sum_i exp(x_i)} exp(x_i)$, is applied row-wise.

We train the network using cross-entropy loss over all companies in the network:

$$Loss = -\sum_{i \in Z_L} \sum_{f=1}^{F} Z_{if} lnY_{if}, \tag{3}$$

where Z_L is the set of companies with price information, Z_i is the class of stock price movement direction.

3 EXPERIMENTS

3.1 Data Collection and Experiment Setup

We use a publicly available API tushare [2] to get historical price of listed companies between 29th April 2017 and 31st December 2017. This resulted in 2988 companies. Detail statistics of the training, test sets are shown in Table 1. We extract five numeric features for each company, including *open, high, low, close, volume*. Note that, it is easy to extend our models with other text features. Relationship information for some companies are relative sparse and this would lead to unstable prediction results. To avoid this, we only evaluate our model on companies in CSI 300 [3].

We look back 7 days to construct historical information vector. In particular, we use its historical information in $t-7 \sim t-1$ as inputs X_i to predict target stock price movement in day t. The classification result $Y_i = 1$ represents that the *close* price of the stock increases compared with the *open* price in day t, otherwise, $Y_i = 0$. We use accuracy to evaluate our model.

Table 1: Statistics of experiment datasets

	Training	Test
Number	31,066	13,315
Up rate	0.528	0.501
Time interval	29/04/2017-13/10/2017	16/10/2017-31/12/2017

3.2 Baselines and Proposed Models

For comparison, we propose two baselines that use information of the target company for stock price movement prediction.

- LR [3]: uses the historical numerical information as features for stock price movement prediction via logistic regression model.
- LSTM [6]: encodes the historical numerical information via recurrent neural network.

For our model, we integrate corporation relationships via various strategies.

- DeepWalk+LSTM: pipeline model and the top-N relevant companies are chose based on DeepWalk [7].
- node2vec+LSTM: pipeline model and related companies are chose based on node2vec [4].
- LINE+LSTM: pipeline model and related companies are chose based on LINE [9].
- GCN: a three-layer GCN model[5] is used to re-construct the historical information of the target company incorporating corporation relationships.
- LSTM+GCN: the result of GCN and LSTM are concatenated for stock price movement prediction.

3.3 Experiment Results

The overall experiment results can be seen in Table 2. We have following findings:

- Comparison between *LR* and *LSTM* shows that LSTM model has better performance for stock price prediction. This indicates that it is better to process historical features in a sequential way.
- Taking the relationship between corporations into consideration can greatly improve the performance of stock price prediction. Both pipeline models and joint models can improve the prediction accuracy compared to two baseline models.
- Comparison between *GCN* and *LSTM+GCN* shows that the LSTM model can extract effective information from the company's stock features and this can provide complement information.
- *LSTM+GCN* produces the most impressive performance. The main reason is that it incorporates all other companies' information with the target company, which is helpful for learning better company features embedding.

We further compare the performance of different groups of related companies in term of their effects on target company's stock

[2] http://tushare.org/
[3] https://en.wikipedia.org/wiki/CSI_300_Index

Table 2: Experimental results on stock prices movement of CSI 300. Bold number is the best performance.

Methods	Accuracy
LR	52.07%
LSTM	53.17%
DeepWalk+LSTM	56.93%
node2vec+LSTM	56.61%
LINE+LSTM	57.00%
GCN	54.44%
LSTM+GCN	**57.98%**

price movements prediction. We rank related corporations in terms of their cosine similarities with the target companies and take the top-ranking 10 companies, middle-ranking 10 companies and low-ranking 10 as group Top10, Mid10 and Last10 respectively. The results are shown is Table 3. Comparison between different network embedding methods shows that *LINE* has better performance among these three methods. The performance of *Top10* group of three network embedding methods is the best in three groups. In general, with the similarity between companies increasing, the performance of prediction is getting better.

Table 3: Experimental results on node similarity affection in pipeline model for stock price prediction

Methods		Accuracy
DeepWalk+LSTM	Last10	55.86%
	Mid10	55.79%
	Top10	**56.93%**
node2vec+LSTM	Last10	55.83%
	Mid10	56.38%
	Top10	**56.61%**
LINE+LSTM	Last10	54.52%
	Mid10	55.11%
	Top10	**57.00%**

Case Study of Company Embedding: Figure 3 shows the visualization results of our node embedding learned for corporations. Take the target *Bank of China* as an example, we can see that the 5 nearest companies are *China Construction Bank*, *Industrial and Commercial Bank of China*, *Agricultural Bank of China*, *New China Life Insurance* and *China Everbright Bank*. All of them are in the financial industry, and four of them are state-owned banks in China, the same as *Bank of China*. This indicates that the representation learned from network embedding approaches is able to reflect the similarity of companies in real world.

4 CONCLUSION AND FUTURE WORKS

This paper proposed to incorporate related companies' information for better stock prices movements prediction. Two strategies are utilized to integrate corporation relationships: a pipe-line model and a joint model based on GCN. Experimental results on stock prediction show that using the relationship between companies can improve the performance of stock prediction.

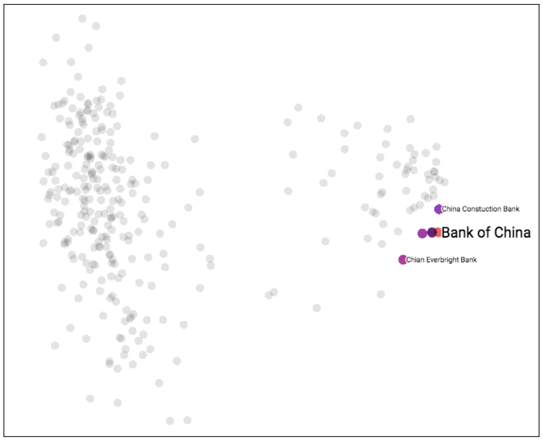

Figure 3: Two-dimensional PCA projection of 128-dimensional company embeddings

Future works will be carried out in two directions. First, we can explore different ways to construct the relationship graph, such as, co-occurrence in the news information, interaction via financial event, etc. Second, currently only numerical information is considered for stock price movement prediction. It is worthy to integrate heterogenous information sources for better prediction. How to incorporate information from multiple sources can be a challenging but useful research direction.

ACKNOWLEDGMENTS

Thanks for the contribution from Yiying Yang, Xuanjing Huang and Libo Wu in Fudan University. The work is partially supported by National Natural Science Foundation of China (Grant No. 61702106), Shanghai Science and Technology Commission (Grant No. 17JC1420200, Grant No. 17YF1427600 and Grant No.16JC1420401).

REFERENCES

[1] Xiao Ding, Yue Zhang, Ting Liu, , and Junwen Duan. 2015. Deep learning for event-driven stock prediction. *In Proceedings of IJCAI* (2015), BueNos Aires, Argentina, August.

[2] Eugene F Fama. 1965. The behavior of stock-market prices. *The journal of Business* (1965), 38(1):34–105.

[3] Jibing Gong and Shengtao Sun. 2009. A New Approach of Stock Price Trend Prediction Based on Logistic Regression Model. *In New Trends in Information and Service Science* (2009), 1366–1371.

[4] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. *In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (2016), 1225–1234.ACM.

[5] Thomas N. Kipf and Max Welling. 2017. Semi-supervised Classfication With Graph Convolutional Networks. *In International Conference on Learning Representations(ICLR)* (2017).

[6] David M. Q. Nelson, Adriano C. M. Pereira, and Renato A. de Oliveira. 2017. Stock Market Price Movement Prediction With LSTM Neural Networks. *In International Joint Conference on Neural Networks (IJCNN)* (2017), 1419–1426.

[7] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (2014), 701–710.ACM.

[8] Robert P Schumaker and Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)* (2009), 27(2):12.

[9] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. *In Proceedings of the 24th International Conference on World Wide Web* (2015), 1067–1077.ACM.

[10] Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. More than words: Quantifying language to measure firms fundamentals. *The Journal of Finance* (2008), 63(3):1437–1467.