2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) | 978-1-6654-6468-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICML57342.2022.10009834

Canonical Correlation Analysis based Bi-Graph Convolutional Network for Stock Price Movement Prediction

Kexin Zhang
School of Statistics and Mathematics, Guangdong University of Finance & Economics
Guangzhou, China
kexinzhang1998@163.com

Jia Cai *
School of Digital Economics, Guangdong University of Finance & Economics
Guangzhou, China
jiacai1999@gdufe.edu.cn
* Corresponding author

***Abstract*—Stock price movement prediction is a very challenging task due to the high volatility and complexity of uncertain financial market. Obviously, the price fluctuations of a target stock is inevitably affected by the price of other related stocks. However, the cross interaction effect among a collection of stocks is not fully explored in the literature. In this paper, we develop a deep learning based novel framework, which combines graph convolutional network, gated recurrent unit and canonical correlation analysis to perform stock movement prediction. Specifically, multiple relationships among stocks are represented by two graphs, namely, industry graph and topicality graph, to model the cross effect problem. Furthermore, gated recurrent unit is utilized to detect the temporal relationships among stocks. In addition, the canonical correlation analysis method is employed to enhance the inter-view correlation between industry factor and topicality factor. Experiments on the well known China Securities Index 300 demonstrate the performance and effectiveness of the proposed approach.**

Keywords-canonical correlation analysis; gated recurrent unit; graph convolutional network

I. INTRODUCTION

Stock trend prediction, as the main concerns of many investors, abound in both traditional finance and modern finance. However, conducting stock price prediction is very challenging due to the complexity and fluctuations of uncertain stock market. Recently, many researchers have proposed various traditional machine learning methods to extract valuable information from various information sources. However, previous approaches assume that stocks are independent of each other, which is inappropriate. The cross-interaction of stocks exist in different forms in reality. Several works try to characterize the complex factors by employing the graph convolutional network (GCN) [26], previous approaches roughly concatenate different graphs, which does not well embody the inter-view relationships.

In this paper, based upon[26] we model various interactions among stocks by utilizing GCN and construct two types of graphs: lead-lag theory [23] based industry graph and common topical news [22] based topicality graph. Gated recurrent unit (GRU) is employed to embody the temporal dependency from historical market data. Moreover, canonical correlation analysis (CCA) is leveraged to capture the inter-view correlation of different information sources.

Specifically, the contributions of the paper are addressed as the following:

- GCN is utilized to construct two graphs: industry graph and topicality graph from the spatial viewpoint whereas GRU is employed to address the temporal relationships among stocks.
- CCA is employed to enhance the correlation of industry factor and topicality factor.
- The proposed Canonical correlation analysis based bi-Graph Convolutional Network (Cab-GCN) is tested on China Securities Index 300 (CSI 300) dataset to demonstrate the performance and effectiveness of the method.

II. RELATED WORK

A. Stock price prediction

Both traditional machine learning methods and deep learning approaches are widely utilized for stock price prediction. Besides historical data information, other types of information sources are utilized to capture the input features in stock price prediction task. For instance, industry information , public news [16], social media [13], and financial performance [2]. Events from news titles were extracted to design a LSTM-based framework [1]. However, there are only a few attempts considering the cross effect [10, 26].

B. Graph convolutional network

The corresponding neural network, namely, GCN has drawn tremendous attention due to its excellent ability in exploiting graph structure information and been widely applied in but not limited to recommender systems computer vision[7]

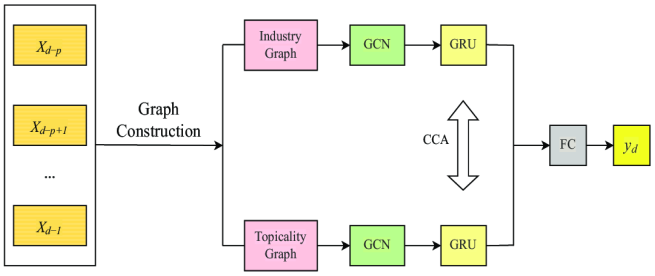


Figure 1: The framework of Cab-GCN.

III. THE PROPOSED APPROACH

A. Preliminary

Stock price prediction is a classic problem in the interdisciplinary fields of finance and computer science. It consists of stock price return prediction and stock movement prediction. Most works in the literature focus on stock movement prediction task. In general, stock price movement aims to predict the trends of an individual stock on a trading day based upon historical price information and other feature information, which can be mathematically formulized as:

$$\hat{y}_d^s = f([x_{d-p}^s, \cdots, x_{d-1}^s], E; \Theta) \quad (1)$$

where $x_t^s \in R^P$ denotes the features of a stock at day t, s is the target stock $\hat{y}_d^s \in [0,1]$ is the predicted probability at day d , p means the lag size, E denotes external feature information, Θ is the set of trainable parameters. However, Equation (1) does not embody the correlations with other related stocks. We employ the idea stated in [26] and encode the complex correlations among stocks as graphs, i.e.,

$$\hat{y}_d = f([x_{d-p}^s, \cdots, x_{d-1}^s], G; \Theta) \quad (2)$$

where $X_t \in R^{N \times P}$ is a matrix of N stocks at day t . G denotes relevant graph. $\hat{Y}_d = [\hat{Y}_d^1, \cdots \hat{Y}_d^N] \in R^N$ is the predicted labels on the d-th day. For example, we let $f(\cdot)$ be the logistic regression function and use least squares method. We need to minimize $R = \sum_{s=1}^{N} (\hat{y}_d^s - y_d^s)^2$ to fit the data. Suppose we predict the opening price of Ping An Bank stock on day 100, so d=100, and the lag size p can be 5 days. As for the graph G can be shareholding graph, industry graph or topicality graph. Θ is the weight parameter matrix. We consider cross entropy function as the loss function:

$$loss^{C-e} = -\frac{1}{N} \sum_{s=1}^{N} [y_d^s \log(\hat{y}_d^s) + (1 - y_d^s) \log(1 - \hat{y}_d^s)] \quad (3)$$

where $y_d = [y_d^1, \cdots y_d^N]$ with its components $y_d^s \in \{0,1\}$

stands for the ground truth series. Here $y_d^s = 1$ means the stock price rises, and $y_d^s = 0$ means the stock price falls down. However, cross entropy loss does not embody the latent correlation between different factors.

B. Canonical correlation analysis based bi-graph convolutional networks

In this part, we address the framework of the developed Cab-GCN, which aims to predict the stock price movement by considering temporal factor and spatial factor jointly. First, we utilize graph to extract the cross interaction relationship among stocks. Second, we use Bi-GCN to calculate two graphs respectively in order to capture spatial correlations. Third, we employ GRU to learn the temporal dependency. we further consider enhancing the inter-view correlation via CCA method. The developed architecture is demonstrated in Fig. 1

Graph Construction: To detect the complicated cross effect among the selected stocks, two types of relationships based on prior financial domain knowledge are extracted. We construct industry graph $G_I = (V, \varepsilon_I, A_I)$ to encode the lead-lag effect within industry and topicality graph $G_T = (V, \varepsilon_T, A_T)$ to encode topical news impact, where |V| = N denotes the number of selected stocks in the constructed graph, $A = (a_{ij})_{N \times N}$ is the adjacency matrix with elements a_{ij} stands for the connection strength between the i-th company and the j-th company. The details are addressed as follows:

(1). Industry Graph: In this paper, an industry graph is constructed to depict the lead-lag relationship. We consider intra-industry lead-lag effect,i.e., there is no edge between two stocks from different industries. Otherwise, the impact from stock j to stock k in the same industry can be described as $a_{ij} = \frac{M_i}{M_j}$, where M stands for the firm size.

(2).Topicality Graph: We collect the dataset from a public API[1] , which contains topical information of each stock. In general, an individual stock has more than one topicality and many stocks may share a common topicality. The connection strength of two stocks are quantified by the number of shared topicalities. Now we measure the connection strength mathematically. Assume company i has M_i topicalities. Company i and Company j share T_{ij} topicalities, then the connection strength from i to j canbe measured by $b_{ij} = \frac{T_{ij}}{M_i}$ .Similarly, the connection strength from j to i can be described as $b_{ji} = \frac{T_{ij}}{M_j}$.

Bi-Graph Convolutional Network:

(1). Graph Convolutional Layer: We get a graph Fourier basis U by decomposing the normalized graph Laplacian matrix L. In this paper, the graph convolutional layer is computed as:

$$H^{(l+1)} = \rho((\sum_{k=0}^{K-1} \theta_k L_\tau^k) H^{(l)} W^{(l)}), \tau \in \{I, T\} \quad (4)$$

$H^{(l)} \in R^{N \times P}$ denotes the input at the ℓ -th layer, $\rho(\cdot)$ is the activation function. $W^{(l)}$ is the parameter matrix.

(2). Bi-GCN: The proposed Bi-GCN can be described as :

$$X_t^{GCN} = \rho(f_\tau(L)\rho(f_\tau(L)X_t W^{(1)})W^{(2)}) \quad (5)$$

where $f_\tau(L) = \sum_{k=0}^{K-1} \theta_k L_\tau^k$ varies according to the selection of τ .

Gated Recurrent Unit:

GRU, as one variant of RNN, the hidden layer of GRU is formulated as:

$$r_t = \sigma([H_{t-1}, X_t, X_t^{GCN}] \cdot W_r + b_r)$$
$$u_t = \sigma([H_{t-1}, X_t, X_t^{GCN}] \cdot W_u + b_u)$$
$$\hat{H}_t = \tanh([r_t \odot H_{t-1}, X_t, X_t^{GCN}] \cdot W_h + b_h)$$
$$H_t = u_t \odot H_{t-1} + (1-u_t) \odot \hat{H}_t$$

where r_t means the reset gate and u_t denotes the update gate. X_t is the record of stocks at time $t \in [d-p, \cdots, d]$. X_t^{GCN} is the output of Bi-GCN, which contains the cross interaction information at instant t. H_{t-1} is the hidden state at instant t− 1. τ is the sigmoid activation function. Then, we achieve the output $X_t^{GRU} = \sigma(H_t W_g) \in R^{N \times G}$.

Inter-view Correlation Enhancement:

We further consider to enhance the inter-view correlation between industry factor and topicality factor via CCA, which aims to maximize the correlations defined as the following:

$$(W_1^*, W_2^*) = \arg\max_{W_1, W_2} corr(W_1^T X_{d,I}^{GRU}, W_2^T X_{d,I}^{GRU}) \quad (6)$$

where $X_{d,I}^{GRU}$, $X_{d,T}^{GRU}$ is the output of GRU for industry and topicality graph at day d. Define R_{11} and R_{22} as the covariance matrices of these, whereas the cross-covariance are denoted as R_{12}. Denote $E = R_{11}^{1/2} R_{12} R_{22}^{1/2}$, then the canonical correlation loss involving $X_{d,I}^{GRU}$ and $X_{d,T}^{GRU}$ is defined as:

$$loss^{corr} = -Trace(E^T E)^{1/2} \quad (7)$$

We minimize $loss^{corr}$ to capture the correlation between industry factor and topic factor. In summary, the whole loss function of the proposed Cab-GCN model can be conducted as:

$$loss = loss^{C-e} + \lambda loss^{corr} \quad (8)$$

where $loss^{C-e} = -\frac{1}{N}\sum_{s=1}^{N}[y_d^s \log(\hat{y}_d^s) + (1-y_d^s)\log(1-\hat{y}_d^s)]$ is given in equation (3). It is the cross entropy function ,and can be seen as the data loss. $loss^{corr}$ is given in equation (7), which is the loss of CCA ,and $\lambda loss^{corr}$ can be seen as the regularization loss, it aims to prevent over-fitting .For example, if the model is over-fitting, λ is a balance coefficient, and the process of regularization can improve the generalization ability of the model. The equation (8) means the sum of the data loss and regularization loss which can combine the machine learning module and the CCA module better.

IV. EXPERIMENTS

A. Settings

To indicate the performance and the effectiveness of the proposed Cab-GCN, we collect the stock dataset from the famous public API tushare according to the China Securities Index300(CSI300). There are three kinds of attributes for each stock in our dataset:(1).Input features: They consist of trading amount, opening price, low price and high price. (2). Relationship features: We use relationship features to construct industry and topicality relationship graphs. (3).Label feature: Closing price is adopted as the label feature.

Denote the closing price of a stock s on day t as F_t , if $F_t > F_{t-1}$,we set $y_t = 1$ to indicate that the price rises at this day, and $y_t = 0$ implies that the price falls down at that day.

We select the historical stock data from November 2020 to December 2021. After adjusting the price, finally, 287 stocks in CSI300 are selected by basic preprocessing. The missing information is filled up with the trading data information in the most recent day. We randomly select 70%, 10% and 20% of the dataset as training data, validation data and test data, respectively. Details of the division are presented in Table 1.

TABLE 1. THE PARTITION OF CSI 300 DATASET

Index	Training set	Validation set	Testing set	Total
CSI 300	52626	7519	15040	75194

Obviously, stock price prediction is a classical binary classification problem, thus we can use the following metrics to measure the effectiveness of all the methods, i.e., Accuracy (ACC), Recall,F1-score, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

B. Baselines

We compare the proposed Cab-GCN model with the following baselines:(1). LR: Logistic regression mode (2). SVM: Classical support vector machine model.(3).RF: Random forest method.(4). LSTM: Long short term memory model, (5). Multi-GCGRU: A model introduced in [26], which combines shareholding graph, industry graph and topicality graph. We test each method ten times and report the averaged results.

C. Results

Experimental results in Table 2 indicate that only considering historical records of the target stock does not perform well. In general, deep learning based methods such as LSTM, MUTI-GCGRU and Cab-GCN perform better than classical LR, SVM and RF methods. This demonstrates that cross effect and temporal dependency play crucial role in the model performance. Moreover, MUTI-GCGRU and Cab-GCN perform better than LSTM. This may be due to the reason that cross effect can help improving the performance since both

https://tushare.pro/.

MUTI-GCGRU and Cab-GCN consider the cross effect. Compared MUTI-GCGRU with Cab-GCN, the performance of Cab-GCN increases nearly 1% in terms of accuracy, which means CCA can indeed enhance the relationship between industry graph and topicality graph. How do the GRU, CCA, industry graph and topicality graph affect the performance of the proposed model? We will elaborate it in the sequel.

D. *Ablation Study*

To testify the effectiveness of industry graph, topicality graph and CCA, we conduct ablation study by removing GRU or CCA or both. In order to demonstrate the effect of relationship graphs, we also conduct experiments by only considering industry graph or topicality graph, i.e., we compare the proposed Cab-GCN method with the following

(1). GCN-I: Only industry graph is considered in the GCN model.

(2). GCN-T: Only topicality graph is considered in the GCN model.

(3). GCN-I-GRU: The model integrate industry graph based GCN with GRU.

(4). GCN-T-GRU: The model integrate topicality graph with GRU.

(5). BGCN: The GCN model with two graphs and GRU.

TABLE 2. THE EXPERIMENTAL RESULTS ON CSI 300 DATASET

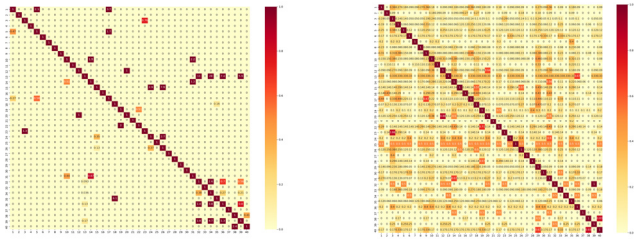
Models	Accuracy	Recall	F1	MSE	RMSE
LR	0.5134	0.5165	0.6218	0.2416	0.4916
SVM	0.5278	0.5127	0.6234	0.2456	0.4955
RF	0.5123	0.5107	0.6234	0.2398	0.4896
LSTM	0.5367	0.5208	0.6364	0.2266	0.4853
MULTI	0.5372	0.5265	0.6332	0.2326	0.4822
Cab-GCN	**0.5428**	**0.5316**	**0.6423**	**0.2313**	**0.4809**

TABLE 3. ABLATION STUDY

Models	Accuracy	Recall	F1	MSE	RMSE
CGN-I	0.5012	0.5065	0.6134	0.2266	0.4853
GCN-T	0.5126	0.5193	0.6127	0.2478	0.4917
GCN-I-GRU	0.5248	0.5172	0.6165	0.2416	0.4915
GCN-T-GRU	0.5317	0.5123	0.6227	0.2423	0.4922
BGCN	0.5301	0.5207	0.6218	0.2325	0.4821
Cab-GCN	**0.5428**	**0.5316**	**0.6423**	**0.2313**	**0.4809**

The results of ablation study are displayed in Table 3, we can see that adding GRU in the model can improve the accuracy for about 2%, which means GRU can well extract the temporal relationships. Moreover, the results in Table 3 also demonstrate that considering industry graph and topicality graph separately lead to bad performance. BGNN achieves higher accuracy, precision, recall and F1 indices, which means industry graph and topicality graph are complementary to each other. In addition, comparing the results achieved by BGNN with those by Cab-GCN, we find that CCA method can effectively exploit the inter-view relationships, which yields at least 1% increase of accuracy, recall, F1, MSE and RMSE. The results in Table 3 also indicates that topicality relationship is better than industry relationship by at least 1% improvement in terms of accuracy. Moreover, common news have more impact n stock price than topicality.

To exploit the correlation between relationship, the industry relationship matrix and topicality relationship matrix are visualized in Fig. 2. We can see that the industry relationship matrix is sparser than the topicality relationship matrix. This may be due to the reason that the dense topicality matrix on contains richer information than the sparse matrix, which can effectively predict the stock price movement.



(a) CSI300-Industry (b) CSI300-Topicality

Figure 2: The Visualization of Relationship Matrices on CSI 300 dataset.

E. *The Length of Historical Information*

In this part, the length of historical information is considered as a crucial factor in the influence of the model performance. Specifically, we consider 3, 5, 7, 9, 11 days experimental results are represented in Table 4 and Fig. 2 The proposed Cab-GCN model achieves the best performance in terms of accuracy 0.5428 and MSE 0.2313 at 7days, whereas Cab-GCN obtains the worst performance in terms of accuracy 0.5264 and MSE 0.2465 at 3 days. Therefore, the length of historical information has crucial impact on stock price movement prediction task.

TABLE 4. CAB-GCN WITH DIFFERENT LAG SIZES

Length	Accuracy	MSE
3-days	0.5264	0.2465
5-days	0.5317	0.2376
7-days	**0.5428**	**0.2313**
9-days	0.5407	0.2321
11-days	0.5397	0.2348

V. CONCLUSION

We developed a novel Cab-GCN framework to conduct stock price movement prediction. Specifically, we first utilize GCN to construct industry graph and topicality graph, which aims to detect spatial localization information of the graphs. Second, GRU is employed to extract the temporal relationships of different stocks. Third, we use CCA method to exploit the inter-view correlation of industry graph and topicality graph. The performance of Cab-GCN is demonstrated on CSI 300 dataset.

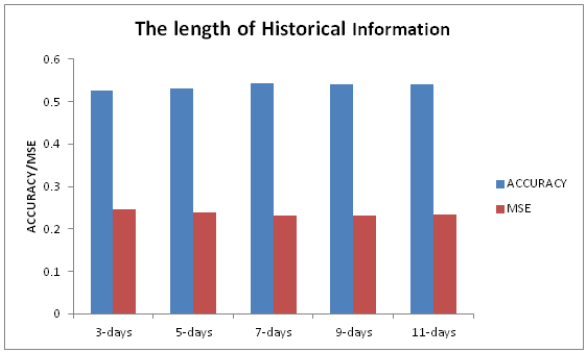


Figure 3: Accuracy and MSE of Cab-GCN with different Lag Sizes.

ACKNOWLEDGMENTS

The work described in this paper was supported partially by the National Natural Science Foundation of China (11871167, 12271111), Special Support Plan for High-Level Talents of Guangdong Province (2019TQ05X571), Guangdong Basic and Applied Basic Research Foundation (2 022A1515011726), Foundation of Guangdong Educational Committee (2019KZDZX1023), Project of Guangdong Province Innovative Team (2020WCXTD011).

REFERENCES

[1] R. Akita, A. Yoshihara, T. Matsubara, and K. Uehara. 2016. Deep learning for stock rediction using numerical and textual information. In 2016 IEEE/ACIS 15th

[2] Edhi Asmirantho and Oktiviani Kusumah Somantri. 2017. The effect of financial performance on stock price at pharmaceutical sub-sector company listed in Indonesia stock exchange. JIAFE (Jurnal Ilmiah Akuntansi Fakultas Ekonomi) 3,2 (2017), 94–107.

[3] Amulya Arun Ballakur and Arti Arya. 2020. Empirical evaluation of gated recurrent neural network architectures in aviation delay prediction. In 2020 5th International Conference on Computing, Communication and Security (ICCCS). IEEE, 1–7.

[4] Mariana Belgiu and Lucian Drăguţ. 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing 114 (2016), 24–31.

[5] I. Bordino, N. Kourtellis, N. Laptev, and Y. Billawala. 2014. Stock trade volume prediction with Yahoo Finance user browsing behavior. In IEEE International Conference on Data Engineering.

[6] J. Bruna, W. Zaremba, A. Szlam, and Y. Lecun. 2014. Spectral Networks and Locally Connected Networks on Graphs. In ICLR.

[7] S. Casas, Cole Gulino, Renjie Liao, and R. Urtasun. 2020. SpAGNN: Spatially-Aware Graph Neural Networks for Relational Behavior Forecasting from Sensor Data. 2020 IEEE International Conference on Robotics and Automation (ICRA) (2020), 9491–9497.

[8] Wesley S. Chan. 2003. Stock price reaction to news and no-news: drift and reversal after headlines. Journal of Financial Economics 70, 2 (2003), 223–260.

[9] C. Chen, L. Zhao, J. Bian, C. Xing, and T. Y. Liu. 2019. Investment Behaviors Can Tell What Inside: Exploring Stock Intrinsic Properties for Stock Trend Prediction. In the 25th ACM SIGKDD International Conference.

[10] Yingmei Chen, Zhongyu Wei, and Xuanjing Huang. 2018. Incorporating Corporation Relationship via Graph Convolutional Neural Networks for Stock Price Prediction. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (Torino, Italy) (CIKM '18). Association for Computing Machinery, New York, NY, USA, 1655–1658. https://doi.org/10.1145/3269206.

[11] Vladimir Cherkassky and Yunqian Ma. 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural networks 17, 1 (2004), 113–126.

[12] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.NIPSWorkshop (2014).

[13] A. Derakhshan and H. Beigy. 2019. Sentiment analysis on stock social media for stock price movement prediction of Artificial Intelligence. 85, Oct. (2019), 569–578.

[14] Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2019. Enhancing stock movement prediction with adversarial training.Proceedings of the 28th International Joint Conference on Artificial Intelligence (2019), 5843–5849.

[15] Thomas Fischer and Christopher Krauss. 2018. Deep learning with long short-term memory networks for financial market predictions. European journal of operational research 270, 2 (2018), 654–669.

[16] M. Hagenau, M. Liebmann, and D. Neumann. 2012. Automated News Reading:Stock Price Prediction Based on Financial News Using Context-Specific Features. Decision Support Systems (2012), 685–697.

[17] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In Proceedings of the eleventh ACM international conference on web search and data mining. 261–269.

[18] Bernard Njindan Iyke and Sin-Yu Ho. 2021. Stock return predictability over four centuries: The role of commodity returns. Finance Research Letters 40 (2021), 101711.

[19] Kewei and Hou. 2007. Industry Information Diffusion and the Lead-lag Effect in Stock Returns. The Review of Financial Studies (2007), 1113–1138.

[20] Thomas Kipf and M. Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In ICLR.

[21] ChristopherKrauss,XuanAnhDo,andNicolasHuck.2017. Deepneuralnetworks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research 259, 2 (2017), 689–702.

[22] Wei Li, Ruihan Bao, Keiko Harimoto, Deli Chen, Jingjing Xu, and Qi Su. 2020. Modeling the Stock Relation with Graph Network for Overnight Stock Movement Prediction.. In IJCAI, Vol. 20. 4541–4547.

[23] Andrew W Lo and A Craig MacKinlay. 1990. When are contrarian profits due to stock market overreaction? The review of financial studies 3, 2 (1990), 175–205.

[24] Andrew W. Lo and A. Craig MacKinlay. 1990. When Are Contrarian Profits Due to Stock Market Overreaction? The Review of Financial Studies 3, 2 (1990), 175–205.

[25] Linkai Luo, Shiyang You, Yanru Xu, and Hong Peng. 2017. Improving the inte-gration of piece wise linear representation and weighted support vector machine for stock trading signal prediction. Applied soft computing 56 (2017), 199–216.

[26] Jiexia Ye, Juanjuan Zhao, Kejiang Ye, and Chengzhong Xu. 2021. Multi-Graph In 2020 25th International Conference on Pattern Recognition (ICPR). 6702–6709. https://doi.org/10.1109/ICPR48806.2021.9412695