Research on Clustering Application Based on Stock Association Network

Chi Ma 1,2, Shengliang Lu 2,1,* and Shaofan Wang 2,1
1. School of computer Science and Engineering, Huizhou University, Huizhou, China.  
2. School of computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China.  
Email: lu_shengliang@163.com

Abstract. This article uses stock market data and text information to construct related complex networks of stock information, and compares the cluster analysis effect of three community discovery algorithms on two networks to discuss how to classify stocks in order to give investors provide better reference. First, transform the heterogeneous data into structured data by obtaining and preprocessing. Then, build an association network based on the similarity of stock price fluctuations and the correlation of stock text information. Finally, using three different cluster analysis algorithms to analyze the stock association network and the text similarity network to compare the effects of different algorithms on stock classification.

1. Introduction

The purpose of this article is to use complex network theory to perform clustering analysis of stocks in the capital market [1]. Through stock market data and individual stock text information such as stock reviews and announcements, construct an association network based on the correlation between stock price and text information. Use the industry data's own attributes to determine the threshold value for optimizing the association network. Base on filtering and optimizing the redundant edges of the network using the threshold method, obtain the final stock association network, and perform clustering on it. Divided with the traditional artificial plate different, This article uses the Newman fast algorithm, Louvain fast algorithm, and the community discovery algorithm based on the LDA topic model to cluster stocks on the obtained stock association network. Based on those models, the clustering of several algorithms is analyzed and compared with the traditional stock division.

2. Related Work

Community discovery algorithm based on modularity is the most classical algorithm series. Newman and Girvan put forward the concept of modularity in 2004. After that, modularity has become a generally recognized evaluation index to measure the quality of network community Division [2]. Many community discovery algorithms have been proposed on this basis. According to the different ways to optimize the objective function, it is generally divided into two types of network community discovery algorithms based on statistical reasoning and heuristic measurement [3]. But at present, the more commonly used community discovery algorithm classification is based on whether a single network node is allowed to exist in multiple network communities, and the discovery algorithm is divided into two types: overlapping and non-overlapping community discovery algorithm. Among them, Newman fast algorithm and GN algorithm are typical non-overlapping community discovery algorithms based on modular degree [4]; simulated annealing optimization algorithm combined with random walk is a non-overlapping community discovery algorithm based on information theory, and
this algorithm has high accuracy [5-9]. Other non-overlapping algorithms also have a good performance in complex network community discovery.

3. Cluster Analysis of Stocks Based on Complex Networks

3.1. Calculate Time Series Similarity of Stocks Based on Dynamic Time Warping (dtw)
Suppose \( X, Y \) as two time series of length \( n \) and \( m \), using the points of the two series constructed matrix \( M \). Using \( M_{ij} \) to express the distance between \( x_i \) and \( y_j \), DTW algorithm is to find a curved path from the distance matrix which can minimize the cumulative cost between \( X \) and \( Y \), and calculate the total cost of this path, as shown in equation 1:

\[
DTW(X, Y) = \min_{w} \left( \sum_{k=1}^{H} w_k \right)
\]

In addition, z-score is used to standardize the time series to eliminate the impact of the dimension and standardize time series data. Equation 2 is the normalized calculation formula. Using \( \mu \) represents sample, \( x \), \( \sigma \) represents mean and standard deviation of the samples. After the dtw distance of any two stocks is determined, equation 3 quantifies the dtw distance into correlation coefficient. Finally, the \( N \times N \) Order symmetric correlation coefficient matrix \( C \) is calculated, as shown in equation 4.

\[
x' = \frac{x - \mu}{\sigma}
\]

\[
\rho_{ij} = \frac{1}{1 + d_{ij}}
\]

\[
C = \begin{cases} 
\rho_{ij}, & i \neq j \\
1, & i = j 
\end{cases}
\]

3.2. Generate Text Similarity Network Based on Text Information Similarity
Softmax is often used for classification problems in neural networks. The function expression is as shown in equation 5, and equation 6 is used for classification calculation.

\[
p(w|y) = \frac{\exp(v'_{w_0}^T v_w)}{\sum_{w=1}^{W} \exp(v'_{w_0}^T v_w)}
\]

\[
\log\delta(v'_{w_0}^T v_w) + \sum_{i=1}^{k} E_{w_j-p(w)} \left[ \log\delta(-v'_{w_0}^T v_w) \right]
\]

In actual operation, a word vector with a 200 dimensions is constructed for each word by word2vec, and by calculating the cosine similarity between the word vectors, the quantized similarity can be obtained to determine the similarity relationship between them. A symmetric matrix \( D \) is used to describe the degree of similarity between stock text information.

3.3. Use Three Community Discovery Algorithms to Perform Cluster Analysis on Networks
Newman fast algorithm uses modular value \( Q \) as quantitative target to measure the advantages and disadvantages of community Division. Newman's fast algorithm initially treats each node of the complex network as a community, and then iteratively merges each generated the two communities with the largest value of \( Q \), until the end of the iteration, the community partition with the largest module degree value is selected as the result. The algorithm is as follows:
Initially set each node of the network as a community, \( n \) communities in total. The initial values of \( e_{ij} \) and \( a_i \) are as flow:

\[
e'_{ij} = \begin{cases} 
\frac{1}{2m} & \text{if } i \text{ and } j \text{ are connected} \\
0 & \text{else}
\end{cases} \tag{7}
\]

\[
a_i = \frac{k_i}{2m} \tag{8}
\]

- Merging pairs of connected communities and calculating the modularity increment after the merger
- Repeat step (2) and continue to merge until merged into a community.

Louvain's fast algorithm is divided into two phases: Traversing the network nodes, merging each node to its neighboring node communities, the modular value gradually increases until all nodes are stable. After that, it reconstructs the network by merging the communities into super nodes. Louvain's fast algorithm is as follows:

- Initialize each point to be divided into different communities;
- Divide each point into the community adjacent to it, and calculate the module degree, then judge the difference between the module degrees before and after the division \( \Delta Q \). If \( \Delta Q \) is positive, accept the division, otherwise, give up the division;
- Repeat step (2) until the modularity no longer increases;
- Construct a new graph. Each point in the new graph represents the community drawn in step (3), and continues to perform (2) and (3) until the community structure no longer changes.

In the above Louvain fast algorithm steps, The calculation formula of modular increment \( \Delta Q \) is shown in equation 9, Where \( \sum_{\text{in}} \) and \( \sum_{\text{tot}} \) represent the sum of all edge weights in the community and all edge weights connected to nodes in Community c respectively. \( k_i \) is the weight of all edges connected with i and \( k_{i,\text{in}} \) is All edge weights connected with i in the community c, \( m \) is the sum of all edge weights in the network graph.

\[
\Delta Q = \left[ \frac{\sum_{\text{in}} + 2k_{i,\text{in}}}{2m} - \left( \frac{\sum_{\text{tot}} + k_i}{2m} \right)^2 \right] - \left[ \frac{\sum_{\text{in}}}{2m} - \left( \frac{\sum_{\text{tot}}}{2m} \right)^2 - \left( \frac{k_i}{2m} \right)^2 \right] \tag{9}
\]

Using the LDA topic model for community discovery. Documents and words are a collection of stock nodes, and stock nodes are used as documents.

\( G = (V, E) \) represents the associated network diagram. \( V = \{v_1, v_2, v_3 \ldots v_m\} \) and \( E = \{e_1, e_2, e_3 \ldots e_n\} \) represent the set of nodes and edges, \( \bar{w}_i \) and \( w_{ij} \) represent the neighbor node of stock node \( v_i \) and its j-th neighbor node respectively respectively. The network interaction model can be expressed in equation 10.

\[
SRI(v_i) = \{(w_{i1}, SRW(v_i, w_{i1})), \ldots, (w_{im}, SRW(v_i, w_{im}))\} \tag{10}
\]

After the model description, the node corresponds to the model SRI can be regarded as a document, the node adjacent to the stock node becomes the word of the document, \( SRW(v_i, v_j) \) is Interaction intensity of \( v_i \) and \( v_j \) as defined in equation 11.

\[
SRW(v_i, v_j) = \begin{cases} 
\rho_{ij} & \text{if } e(v_i, v_j) \in E \\
0 & \text{else}
\end{cases} \tag{11}
\]
4. Experiments and Results

4.1. Evaluation Criteria

Modularity: equation 12 is the modified formula for calculating the weighted modularity, $\Sigma \text{in}$ and $\Sigma \text{tot}$ are the sum of the weights of all the edges in the community and all the edges connected to the inner node.

$$Q = \sum_c \left[ \frac{\Sigma \text{in}}{2m} - \frac{\Sigma \text{tot}^2}{2m} \right]$$

If the modular value of the community structure is larger, it means that the better the community division is, and the links within the community are dense.

Average correlation coefficient: According to formula 13, the correlation degree $\Phi$ of all the stocks in community $c_i$ is measured as the mean value of the number of the price fluctuation relationships among the stocks in community $c_i$. Equation 14 is the average value of the correlation of all communities in the whole community division $C = (c_1, c_2, ..., c_m)$, which is the overall correlation measurement index of the inspection algorithm division, where $m$ is the number of communities obtained by the community division.

$$\Phi(c_i) = \frac{2}{m(m-1)} \sum_{i \neq j} \rho_{ij}$$

$$\bar{\phi}_c = \frac{\sum_m \phi(c_i)}{m}$$

Equation 15 is the coverage calculation formula, Where $c_i$ and $s_j$ represent the classification of stock associations and traditional industry plates obtained by association discovery algorithm respectively.

$$\alpha(c_i, s_j) = \frac{|c_i \cap s_j|}{|c_i|}$$

4.2. Conclusions

The modularity, average correlation, and industry coverage of the three algorithms for clustering analysis of the n1 network are shown in Table 1, and Figure 1.

| Community discovery algorithm | Number of communities | Modularity | Mean correlation coefficient |
|------------------------------|-----------------------|------------|----------------------------|
| Fast Newman Algorithm        | 147                   | 0.562      | 0.778                      |
| Louvain fast algorithm       | 128                   | 0.575      | 0.784                      |
| LDA algorithm                | 134                   | 0.423      | 0.759                      |
The modularity, average correlation and industry coverage of the three algorithms for clustering analysis of n2 networks are shown in Table 2 and Figure 2.

Table 2. Comparison of network n2 application algorithms

| Community discovery algorithm | Number of communities | Modularity | Mean correlation coefficient |
|-------------------------------|-----------------------|------------|-----------------------------|
| Fast Newman Algorithm         | 119                   | 0.482      | 0.445                       |
| Louvain fast algorithm        | 111                   | 0.612      | 0.484                       |
| LDA algorithm                 | 114                   | 0.411      | 0.523                       |

In this chapter, the Newman fast algorithm, Louvain fast algorithm, and LDA topic model community discovery algorithm are used to divide and analyze the clustering of stock association networks. From the experimental results, the LDA topic model community discovery algorithm can obtain higher correlation than Newman fast algorithm and Louvain fast algorithm. At the same time, the clustering results obtained by several community discovery algorithms are compared with traditional plate classification, and experiments have found that the clustering results of several algorithms are higher than the average correlation coefficient of plate division.

5. Acknowledgments
This paper is supported by (1) Foundation of Guangdong Educational committee under the Grant No.2018KTSCX218. (2) The Professorial and Doctoral Scientific Research Foundation of Huizhou University No. 2018JB020.
6. References

[1] HJ Kim, Y Lee, I Kim, B Kahng. Scale-free Network in Financial Correlations[J]. Journal of the Physical Society of Japan, 2001, 71 (9): 2133-2136.

[2] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2004, 69(2): 026113.

[3] Zhang W, Gang P, Wu Z, et al. Online community detection for large complex networks[C] // International Joint Conference on Artificial Intelligence. AAAI Press, 2014: 1903-1909.

[4] Wu J, Du R, Zheng Y Y, et al. Optimal multi-community network modularity for information diffusion[J]. International Journal of Modern Physics C, 2016, 27(08).

[5] Banks J, Moore C, Neeman J, et al. Information-theoretic thresholds for community detection in sparse networks[J]. 2016.

[6] Rosvall M, Bergstrom CT. An information-theoretic framework for resolving community structure in complex networks.[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(18): 7327-7331.

[7] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(4): 1118-1123.

[8] Cheoljun Eom, Jongwon Park, WooSung Jung, et al. The Effects of Market Properties on Portfolio Diversification in the Korean and Japanese Stock Markets[J]. Quantitative Finance, 2009(0902.3836).

[9] V Boginski, S Butenko, PM Pardalos. Statistical analysis of financial networks[J]. Computational Statistics and Data Analysis, 2005, 48(2): 431-443.