Correlation Based Clustering on Stock Market Data

Piyush Nayak[1]
kayan_piyush@tamu.edu

Rajesh Satpathy[1]
rajeshsatpathy@tamu.edu

Sanjay Nayak[1]
sanjaynayak@tamu.edu

Sindhuja Reddy Kamidi[1]
sindhujareddykamidi@tamu.edu

[1] **Department of Computer Science & Engineering, Texas A&M University, College Station**

Abstract

Graph clustering has become synonymous in understanding many aspects of our day-to-day life, ranging from social networks to image segmentation. Analysis and knowledge extraction from financial data is one of the integral tasks everybody in the financial sector is interested in. Stock investors are interested in gaining knowledge on stocks, a representation of the financial state of company in the financial market. Various data mining methods are being used and experimented on, with an aim to extract useful information out of it. This paper focuses on clustering methods that use correlation to provide a sense of similarity and dissimilarity between various stocks. Relations between different features of stocks like closing price, highest price can be used for creating a graph. We demonstrate on how the daily changes in stock prices of 'S&P 500' can be represented as a correlation matrix between stocks and analyze their behavior using correlation-based clustering algorithms, namely, DBSCAN, Hierarchical Clustering, and Correlation Clustering.

1 Introduction

Graph clustering is a technique where the main objective is to partition nodes of a graph into clusters, such that, the connections among the nodes in a cluster are dense and connections among the nodes in different clusters are sparse. A graph can be modelled in many ways, by utilizing the different interactions between entities in the graph. The entities are represented as nodes in the graph and interactions between them are represented as edges between them. A few examples are the Social network graph [11](persons are nodes and social interactions are edges), Computer Network graph (network devices are nodes and communication between them are edges) and Biological Species Network graph(organism type are nodes and their relation and category of species are edges). Stock data can be represented similarly by representing companies as nodes and correlation/interaction between them as edges. The similarity between nodes is one of the main criteria for clustering and is calculated differently for different problems, relevant to the purpose of clustering. A way of calculating the similarity between nodes is by calculating the correlation between them. The correlation between nodes is affected by what the edges in the graph represent. It may be interactions, relations, transactions, dependencies, etc. When clustering on a graph is performed based on the correlation factor between nodes, it is referred to as correlation-based clustering.

1.1 Related Work

Correlation based clustering is more of a general idea that signifies *clustering based on correlation coefficients*. Correlation clustering is a specific type of objective, where the goal is to cluster a signed

graph in such a way that clusters are "correlated" with edge signs, i.e., negative edges tend to be between clusters and positive edges tend to be within clusters. Correlation clustering is used when more than one distance measure needs to balance between two possible contradicting measures, and is used to solve constraint-clustering problems.

Correlation-based clustering can be used to solve real-time problems like web document clustering, parallelizing batch processes, detecting brain activation based on a stimulus, etc, and also use a variety of techniques like Affinity matrix, NC-spectral clustering, and Spatio-temporal clustering techniques. A brief overview on the mentioned clustering techniques and problems mentioned is discussed. Web document clusters are generated using weblogs [12], where the web documents are nodes of the graph and the probability of visiting one document from another, the edges. The nearest correlation method named NC-spectral clustering claims to have a better performance than conventional distance-based methods introduced in [8]. The paper discusses NC-spectral clustering on a case study of parallelized batch processes, by taking an affinity matrix to represent the graph. To detect brain activation on the application of a stimulus task [9] discusses converting FMRI images to p-dimensional vectors for p sequential images, and also to cluster them using correlation values of sensor data [14]. An alternative spatio-temporal clustering technique [10] based on the spatial correlations over time, was demonstrated on big data and proved to circumvent the memory limitations.

Correlation clustering was introduced by [4] and was also proven to be NP-Hard. The authors provided constant-factor approximation for complete graphs and tried to provide an optimal number of clusters. An $O(log\ n)$ approximation algorithm based on "region-growing" technique was proposed by [6]. A 4C(Computing Correlation Connected Clusters) [5] method was introduced, which is a combination of PCA and DBSCAN, to find local subgroups on complex correlations. The word "correlated" is used differently in different approaches. In CURLER [13], an algorithm for finding and visualizing nonlinear correlation clusters in the subspace of high-dimensional databases is discussed, along with the co-sharing level concept. HiCO (hierarchical correlation ordering) [1] is a correlation clustering technique where a hierarchical approach was proposed. COPAC (COrrelation PArtition Clustering) [2] is a correlation clustering algorithm where local correlation dimensionality is assigned to each object of the database and represents the dimensionality of the correlation cluster. ORCLUS is a generalized projected clustering method that picks seeds, assigning data objects to these seeds according to the criteria of distance calculation by considering eigenvectors.

This paper discusses how Graph Clustering can be used to cluster financial network graphs, where the value and performance of a company are characterized by the stock price change in a day. Stock investment portfolio should be diversified for better risk management as it reduces the variance in returns and increases profits. The literature covers a range of solutions to diversify the portfolio such as [15] where the author builds the portfolio by (i) picking stocks randomly, (ii) by choosing one stock from each industry group, (iii) by choosing one stock from each cluster after correlation clustering is performed, (iv) by choosing one stock from each cluster without repeating stocks from the same industry group. It is observed that choosing clusters randomly and choosing stocks randomly from the selected cluster without repeating the industry group forms a more diversified portfolio that gives more returns. This shows that forming a good cluster is critical for building a more diversified portfolio.

1.2 Our Approach

For our paper, we explored on the following methods, namely, DBSCAN, Hierarchical clustering method, and a greedy approximation method of correlation clustering, using which clusters can be created without providing predefined number of clusters to the algorithm. All the three algorithms were implemented over the stock data of 'S&P 500' and the results were analyzed. The greedy approximation algorithm used for the correlation clustering was changed a little by the addition of a threshold point of correlation value to determine whether a stock should be added into the current cluster.

1.3 Organization

In Section 2, we have provided notations, useful definitions, and problem statements. Section 3 provides us more information on the working of the three different algorithms that are used for the

analysis. In Section 4, the results of the three algorithms are provided. In Section 5, we discuss the improvement opportunities and future work.

2 Preliminaries and Definitions

Let $G = (V, E)$ be a complete graph defined by nodes $V = \{S_1, S_2, ..., S_{500}\}$, where S_i represent the stock name and the weight $w_{i,j}$ on the edges $e(i, j) \in E$ represents the correlation between the nodes i and j. Let $\mathcal{N}^+(i) = \{j \in V : (i, j) \in E,\ e(i, j) > 0\}$ and $\mathcal{N}^-(i) = \{j \in V : (i, j) \in E,\ e(i, j) < 0\}$ denote the positive and negative neighbors of i respectively, and A represent its adjacency matrix.

Let w_0 denote the threshold of correlation that is used in correlation based clustering algorithm and $w_{i,j}$ denotes the correlation associated with edge $e(i, j)$. In a cluster $\mathcal{C}$, an edge $e(i, j) \in E$ is a positive mistake if $w_{i,j} > w_0$ and negative mistake if $w_{i,j} < w_0$, and let the positive mistake be denoted by $\mathcal{P}^+$ and negative mistake by $\mathcal{P}^-$. We denote the neighbors of node i as $\mathcal{N}(i)$ and the minimum number of neighbor nodes is represented by $\mathcal{N}_{min}$. The correlation value between two stocks i and j is denoted by $corr_{i,j}$.

3 Methodology

3.1 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN algorithm [7] is a spatial clustering algorithm which finds regions of higher density and separates regions with lower density. Concepts of density-reachable and density-connected are important to define cluster nodes. Here the density of a node i in a region is defined by the number of nodes within a given radius (ϵ) of the node. The algorithm is randomly seeded with a node that has not been visited and the neighborhood nodes are retrieved within the radius from the node. If this node has more than a minimum number of neighbor nodes, cluster formation is initiated and the node is labeled as a core node, else it is labeled as noise. The noise-labeled node can be a part of a neighborhood of another core node. The neighboring nodes of a core-labeled node within the radius, also become a part of the cluster and are labeled as core nodes. The algorithm continues till a density-connected cluster is found.

Our implementation of DBSCAN takes a distance matrix as input that is computed from the correlation matrix, generated using Pearson method to calculate the correlation between the stock data. This generated distance matrix was passed into the DBSCAN algorithm provided in the *sklearn* library of Python. The conversion of correlation matrix to distance matrix is done using the formula

$$d_{i,j} = \sqrt{2 \times (1 - corr_{i,j})} \tag{1}$$

Lemma 1. *Density-Reachable: Node i is said to be directly density reachable from node j when,*

1. $|\mathcal{N}(j)| \geq \mathcal{N}_{min}$ *i.e., j is a core node*
2. $i \in \mathcal{N}(j)$ *i.e., i is in ϵ neighborhood of j*

Lemma 2. *Density-Connected: A density connected cluster is defined as, a node i is density connected to a node j with respect to ϵ and $\mathcal{N}$, if there is a point x such that, both i and j are density reachable from x.*

3.2 Hierarchical Clustering

Here, we have followed the agglomerate version of hierarchical clustering. Clustering algorithm starts with each data point as an individual cluster. At each iteration similar clusters are merged. Proximity of newly merged clusters is calculated and the algorithm is repeated until a single cluster is formed or required k number of clusters are formed. This method was implemented using the *sklearn* library of Python.

Table 1: Correlation based clustering analysis

	Analysis			
Algorithm	# Clusters	$\mathcal{P}^+$ (%)	$\mathcal{P}^-$ (%)	Total $(\mathcal{P}^+ + \mathcal{P}^-)$ (%)
DBSCAN	16	60.9	0.4	61.4
Hierarchical Clustering	9	73.7	0.2	73.9
Mod-CC-PIVOT	17	68.1	2.8	70.9

3.3 Correlation Clustering

The main objective of the correlation clustering is to minimize the negative mistake $\mathcal{P}^-$ within cluster and positive mistakes $\mathcal{P}^+$ between clusters. For the implementation, the CC-PIVOT algorithm [3] is used with modifications in the condition during cluster formation and how to choose the initial node for cluster, which is described in Algorithm 1. The correlation matrix is passed into the algorithm and algorithm continues to create clusters till all nodes have been visited. The whole algorithm was implemented in Python and experiments were conducted using multiple threshold values, w_0.

Algorithm 1 Mod-CC-PIVOT(G):

Pick a pivot that has maximum $\mathcal{N}^+$
$C = \{\text{i}\}, C' = \{\phi\}$
for all $j \in V, j \neq i$ **do**
  if $e(i,j) \geq w_0$ **then**
    $C \leftarrow j$
  else if $e(i,j) < w_0$ **then**
    $C' \leftarrow j$
  end if
end for
Return C, Mod-CC-PIVOT(C')

4 Experimental Results

The heat map corresponding to the correlation matrix obtained from the Pearson method is shown in Figure 1 consisting of first 10 stocks data. The Table 1 provides insight to the number of clusters, the positive penalty and the negative penalty incurred when the three algorithms were used on the stock market data set consisting of 500 different stocks. The $\%$ calculated in the Table 1 is with respect to the total number of edges (positive and negative). The graph of 500 nodes is a fully connected graph and contains a total of $125,000$ edges. It can be observed that DBSCAN algorithm performed the best on the used data set, followed by the Mod-CC-PIVOT algorithm.

From the experiments conducted on the implemented algorithms, it was observed that as the ϵ value in DBSCAN algorithm is increased, the number of clusters created is increased but the number of nodes within each cluster are decreased. The same trend was observed in the Mod-CC-PIVOT algorithm as the threshold value, w_0, is increased. It was also observed that on making $w_0 = 0$, the number of clusters formed by mod-CC-PIVOT algorithm was 3 with most of the nodes in a cluster. This can be attributed to the way the pivot is selected to form a cluster. The number of positive edge degree is always checked to create a pivot. This can be one of the reason for its high penalty.

5 Conclusion and Future Work

This project gave us an insight as to how the stock data set can be formulated into a graph problem and how graph clustering techniques can be applied in order to generate meaningful results. Preliminary work conducted on the stock data, consolidated from the last three years, gave us an insight how

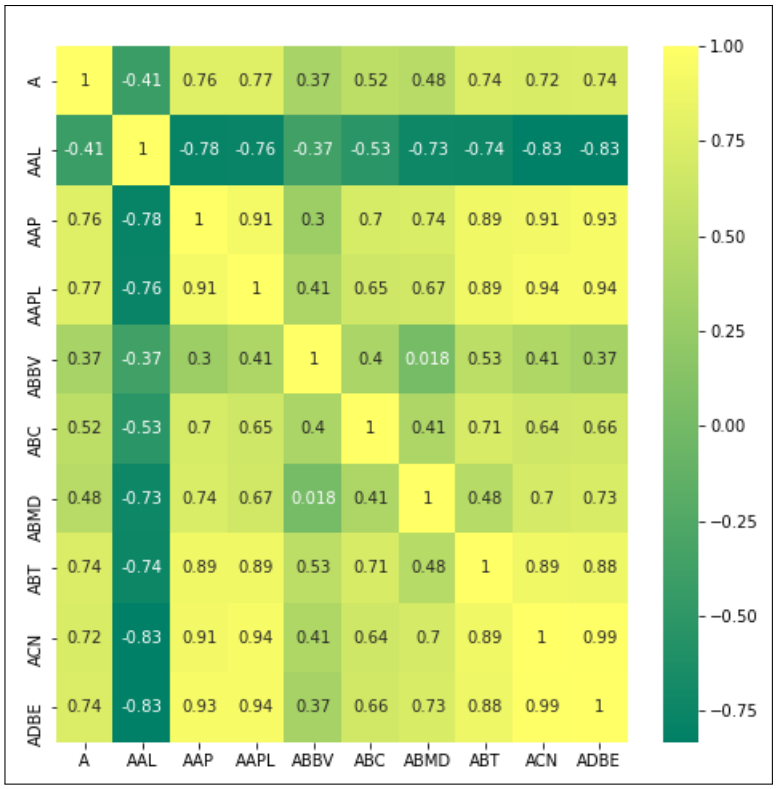

Figure 1: Heat map denoting the correlation of first 10 stocks of S&P 500. The darkest green band signifies the least correlation and the lightest green band/yellow band signifies the highest correlation.

correlation based clustering works over the stock data set. In this paper, we also came up with a modified version of the CC-PIVOT algorithm, a kind of correlation clustering algorithm. This algorithm can also be developed further by improving the techniques used to select the starting vertex that is used to create the cluster. We could further take this work by applying other correlation based clustering techniques and gain insight from the results.

References

[1] E. Achtert, C. Bohm, P. Kroger, and A. Zimek. Mining hierarchies of correlation clusters. In *18th International Conference on Scientific and Statistical Database Management (SSDBM'06)*, pages 119–128, 2006.

[2] Elke Achtert, Christian Böhm, Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Robust, complete, and efficient correlation clustering. Society for Industrial and Applied Mathematics, Apr 2007.

[3] Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: Ranking and clustering. In *Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '05, page 684–693, New York, NY, USA, 2005. Association for Computing Machinery.

[4] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine learning*, 56(1):89–113, 2004.

[5] Christian Böhm, Karin Kailing, Peer Kröger, and Arthur Zimek. Computing clusters of correlation connected objects. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, page 455–466, New York, NY, USA, 2004. Association for Computing Machinery.

[6] Erik D. Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. Correlation clustering in general weighted graphs. volume 361, page 172–187. Elsevier BV, Sep 2006.

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[8] Koichi Fujiwara, Manabu Kano, and Shinji Hasebe. Development of correlation-based clustering method and its application to software sensing. *Chemometrics and Intelligent Laboratory Systems*, 101(2):130–138, 2010.

[9] Xavier Golay, Spyros Kollias, Gautier Stoll, Dieter Meier, Anton Valavanis, and Peter Boesiger. A new correlation-based fuzzy logic clustering algorithm for fmri. volume 40, page 249–260. Wiley, Aug 1998.

[10] Marc Hüsch, Bruno U. Schyska, and Lueder von Bremen. Corcluststt—correlation-based clustering of big spatio-temporal datasets. volume 110, page 610–619. Elsevier BV, Sep 2020.

[11] Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert E. Tarjan. Clustering social networks. In *Algorithms and Models for the Web-Graph*, pages 56–67, 2007.

[12] Zhong Su, Qiang Yang, Hongjiang Zhang, Xiaowei Xu, and Yuhen Hu. Correlation-based document clustering using web logs. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, pages 7 pp.–, 2001.

[13] Anthony K. H. Tung, Xin Xu, and Beng Chin Ooi. Curler. ACM Press, 2005.

[14] Myung Ho Yeo, Mi Sook Lee, Seok Jae Lee, and Jae Soo Yoo. Data correlation-based clustering in sensor networks. IEEE, Oct 2008.

[15] Hannah Cheng Juan Zhan, William Rea, and Alethea Rea. An application of correlation clustering to portfolio diversification. 2015.

A Appendix

The whole project was built and executed using Python 3.9. *sklearn* library was used in order to implement the DBSCAN and Hierarchical clustering algorithm. We tried to modify the CC-PIVOT clustering algorithm in order to get a better version, but further work needs to be done in order to improve the modified algorithm.

Github Link for the code - https://github.com/natsu1628/Correlation_Based_Clustering