An Overview of Clustering Methods in The Financial World

Geyang Tang[1, *, †], Rujian Tian[2, †], Bingdi Wu[3, †]

[1] *Bachelor of Business Administration, University of Toronto, Toronto, ON M1C 1A4, Canada*
[2] *Hangzhou Foreign Languages School, Hangzhou, 310018, China*
[3] *Jiangsu Tianyi High School, Wuxi, 214000, China*
**Corresponding author. Email: geyang.tang@mail.utoronto.ca*
[†] *These authors contributed equally*

ABSTRACT
This paper reviews the widely used clustering methods: K-mean, MST, and the hierarchical approach, along with its application in financial fields. Specifically, it includes a discussion of the application in credit scoring, stock market, portfolio selection, and trading strategy. Moreover, significant challenges and future research directions are also being identified. It is founded that clustering could have wide application in varied financial fields, however, challenges might be solved by future in-depth research. This paper aims at helping other researchers to select the best-fit algorithms to conduct their research and provide an analytical base for people who have an interest in this topic. Additionally, business analysts and quantitative researchers might be able to leverage the insights.

Keywords: *Clustering, Financial application, Machine learning*

1. INTRODUCTION

Clustering, as a type of unsupervised learning, is a beneficial tool in many industries, including computer science, biology, medicine, and the business industry. There are over 100 algorithms that have been developed in the fields, but not all the methods are being widely used, especially in the financial world [1]. People live in a complicated, diverse, enormous, and continually changing financial system. Information is generated at an ever-increasing volume and frequency, posing challenges for data analysis [2]. Although researchers nowadays put more emphasis on deep learning, it does not mean the study on clustering should be paused. There is still a lot to be explored about the application of clustering. However, such a trend poses challenges to getting up-to-date research outcomes, and the scope of review are somehow limited.

This paper aims at studying the commonly used clustering algorithms in the financial system and their area of applications (credit scoring, trading strategy, portfolio analysis, and stock market). The challenges and future research direction are also discussed in the last section.

2. METHODOLOGY

By reading through a considerable amount of literature regarding the application of clusters in the financial world, It is discovered that hierarchical clustering, k-mean, minimum spanning tree (MST)and the hierarchical approach are the primary methods researchers used to explore the financial world. This section will first provide a brief introduction to these three methods, after which summarizing what research topics are linked to the three main methods.

The following example discussed is to show how popular these main methods are in a financial context and does not contain all of the research that is related to these two main methods. However, it contains some of the popular and frequently cited research.

2.1. K-Mean

If a group of observations is given (x1, x2 ...), where x is a real vector, K -mean is to classify all observations into K sets to minimize the variance within the clusters. Numerically, the objective function is [3].

$$J = \sum_{i=1}^{m} \sum_{k=1}^{k} w_{ik} \|x^i - \mu_k\|^2$$

This technique classifies the dataset to k different groups having an almost equal number of values (x). In each cluster, a centroid is a point that represents the cluster, and it is the average of all the points in the set [4]. Since centroids are the mean value of the data set, it might not be the actual point. For example, if all of your data are integer, there is a possibility that the centroid in a cluster is 3.8, which is not an integer and is not any given data in the set.

With regards to the application of K-mean in the financial fields, it has been massively used, especially in the stock market. For example, Zheng et al. used the algorithms to predict the stock market fluctuation and they then used the closing price and price per share to verify the results [4]. In the application section of this paper, [5-7] all used the K-mean to conduct research.

2.2. MST

A spanning tree consists of edges (line segments) connecting all nodes with no cycles. Each edge is weighted in different values, and span trees with the lowest cost/distance are called the minimum spanning trees [8]. To detect clusters of heterogeneous nature, MST-based clustering algorithms have been successfully employed. MST-based algorithms typically produce a graph first, then form MST from the graph, based on data of n random points. One classical MST algorithm in a cluster involves generating a k-partition of the set of data. The algorithm builds an MST of the data set and clear edges that satisfy a predetermined set of criteria. A repetition of the process leads to k clusters being generated.

As for the method of MST, Ren et al. rely on the structures of minimum spanning tree networks in stock markets to find out a dynamic portfolio strategy [9]. There are also some rebates on the utilization of MST, some of the researchers suggested investing money in the centrality of MST because eccentricity-based risk budgeting portfolios have improved return to risk ratios [10]. On the other hand, many other researchers demonstrated that peripheries of the MST are more worthy of being invested in [11,12].

2.3. The hierarchical technique

Each value will be treated as an individual group at the beginning. Then, the similar group will combine to form a new group called A based on distance difference. However, the distance calculation here is not based on MST methods, several alternative ways (mathematical formulas) are used to calculate the distance [13]. A will then compare with other clusters in terms of similarity. The process will continue until all the groups are combined and form a single cluster. The number of clusters is not predetermined, and the hierarchical relationship of this approach can be shown by the dendrogram.

The application of the hierarchical approach is also vast. Gava et al. used hierarchical algorithms to a group of government bond factors, and finds the one factor that benefit most from a positive carry [14]. In terms of product selection, hierarchical clustering was used to determine what product market structure worked best for financial services [15]. Additionally, Lemieux et al. proved the fact that by using the different clustering methods, the results varied significantly even if applied to the same data set. Such a conclusion was achieved by comparing the results from the Hierarchical approach to K mean and K-medoids [13].

3. FINANCIAL APPLICATIONS

Clustering is really necessary for financial analysis. Some papers talk about several clustering applications. These applications can be categorised based on different fields, such as credit scoring, stock selection, portfolio selection, trading strategies, etc.

3.1. Credit Scoring

Credit scoring is one of the considerable methods to predict financial trouble for listed companies. Furthermore, individuals can decide whether to extend or deny credit. As a result, credit scoring can be necessary.

Tripathi et al. utilise an approach based on feature clustering for feature selection, and the dataset with the selected features are applied on five base classifiers, and output obtained by base classifiers are aggregated by weighted voting approach for prediction of final output as Tripathi et al. try to combine the benefits of both feature selection as well ensemble classification to improve the performance of credit scoring model. The clustering methods play a crucial role in this article. In the end, they proposed that the features within a cluster having a ratio between inter and intraclass correlation coefficients nearer to 1 are considered as best features [5].

Hsieh utilises a hybrid mining approach in the design of credit scoring models to support credit approval decisions. The clustering methods play a crucial role in this article for credit scoring. Hsieh addresses the benefit by investigating a simple but effective hybrid utility of clustering and neural network techniques designing a credit scoring model [17].

As mentioned above, both Tripathi et al. and Hsieh utilise the clustering methods in the credit scoring area.

3.2. Stock market

The stock market has been really common nowadays. It is simple for people to operate their stock via clustering methods.

B.S and Mathew utilise the K-mean and density-based techniques to extract and categorise useful stock data from an unknown, substantially helpful dataset. The clustering methods play a crucial role in stock data selection here in this article. The selected data then is being used as the input for stock price prediction [6].

Suresh Babu, Dr N. Geethanjali, and Prof B. Satyanarayana utilise the HRK, which is an advanced technique to predict the short-term stock price movements based on the release of financial reports. This clustering method plays a vital role to predict the stocks for people to select in this article. At the end of the article, Babu et al. employ the representative feature vectors from the clustering methods to predict the stock price movements [7].

Gupta and Dr Samidha D. Sharma utilise a hybrid combinatorial method of clustering and classification to enhance the accuracy of predicting the future value of the stock market. This clustering method plays an important role in the stock prediction. The dataset is first clustered by K-means clustering algorithm, and the values obtained are classified by horizontal partition decision tree [18].

In conclusion, B.S and Mathew, Suresh Babu et al., and Gupta and Dr Samidha D. Sharma successfully predict the stock data better via the clustering methods.

3.3. Portfolio Selection

A portfolio is a collection of financial investments like stocks, bonds, commodities, cash, and cash equivalents. Therefore, people can profit from selecting a portfolio by applying clustering methods.

S.R. Nanda, B. Mahanty, and M.K. Tiwari utilise a data mining approach finally to select stocks from the clusters to build a portfolio, minimising portfolio risk and comparing the returns with that of the benchmark index. This clustering method plays a vital role in portfolio management in this article. Also, the authors demonstrate the implementation of stock data clustering using well-known clustering techniques, namely K-means, self-organising maps (SOM), and Fuzzy C-means in this article [19].

Tolaa, Lilloc, Gallegatia, and Mantegnac applied clustering algorithms to improve the reliability of the portfolio in terms of the ratio between predicted and realised risk by assuming idealised conditions of perfect forecast ability for the future return and volatility of stocks and short selling. The clustering algorithms play a crucial in the optimisation of a financial portfolio in this article [20].

Dose and Cincotti utilise the methodology which first selects a subset of stocks and then sets the weight of each stock as a result of an optimisation process to solve the index tracking problems. The clustering methods play a vital role for applications to index and enhance the index-tracking portfolio in this article [21].

In summary, S.R. Nanda et al., Tolaa et al., Dose, and Cincotti all utilise the clustering methods to improve the applications in portfolio selection.

3.4. Trade Strategies

People can utilise the clustering methods for the trade strategies to better utilise the trade strategies in their work.

Narayan and Popp imply that price clustering can potentially be a source of oil market inefficiency, influencing trading strategies. Narayan and Popp document evidence of price clustering behaviour in the oil futures market. They recognised the importance of price clustering phenomena in the oil futures market and investigated this empirically [22].

Fricke utilises the trading strategy of a particular member institution is defined as the sequence of (intra-) daily net trading volumes within a certain semester. The authors show that there are significant and persistent bilateral correlations between institutions 'trading strategies. They analyse the correlations in trading patterns for members of the Italian interbank trading platform e-MID [23].

As mentioned above, both Narayan and Popp, and Fricke utilise the clustering methods to improve the trading strategy.

4. CHALLENGES AND FUTURE DIRECTION

4.1. Challenges

4.1.1. Uncertainty about the quality and speed of data generation

Data is one of the most significant figures in clustering. Data quality challenges arise when a trade happens more and more frequently [2]. Specifically, when involving a large number of orders, cancellations, and other transactions. Long-term time series may be required for financial stability analysis. Data reliability can be significantly affected by factors such as missing values for specified periods, changes in the meaning of values, and human errors when operating the system. Additionally, the rapid changes in data at various speeds make it much harder to extract data in real-time. On the one hand, it is difficult to determine when is the cut-off point for analysing the existing data as the data is changing all the time. On the other hand, when the data

is created faster than it can be extracted, the trend would change and the clustering methods may fail to predict precisely.

4.1.2. The lack of uniformity

A random choice of clustering methods or/and cluster patterns will generate various clustering results, which brings inconsistency. For example, different runs of the algorithm of K-means will generate different results, thus choosing an appropriate number of splits in HAC which improves the quality of the clustering generated by the K-means clustering, is also a challenge. Moreover, DiwakarTripathi et al. tried to combine the benefits of both feature selection as well ensemble classification to improve the performance of the credit scoring model[5]. For feature selection, a feature clustering-based approach is proposed to find the optimal set of features. Furthermore, a dataset with selected features is forwarded to heterogeneous classifiers, and outputs predicted by various classifiers are aggregated by weighted voting. Some pre-processing steps are also performed before the knowledge discovery process [5]. The various combinations as well generate inconsistent results.

4.1.3. Challenges in defining and measuring systemic risk

Due to the multifaceted nature of the financial system that there is such diversity in definitions. It is very important to consistently commit to specific measures, especially for data-driven calculations. Systemic risk is characterized by its multifaceted nature. Although there are already some agreements on concepts including leverage, liquidity, etc, the agreement always breaks down when it comes to selecting risk measurement. Bisias, et al. point out some disagreements about the definition of systemic risk.[30] Bisias, et al. state one of the definitions of systemic risk that, “Any set of circumstances that threatens the stability of or public confidence in the financial system” [31], While The European Central Bank (ECB) (2010b) defines it as a risk of financial instability. Bisias, et al also find those who focused on more specific mechanisms, which includes imbalances [30], correlated exposures[32], spillovers to the real economy [33], information disruptions[25], feedback behaviour[26], asset bubbles[27], contagion [28], and negative externalities [29].

Such ambiguity is widespread in that it impairs the functioning of a financial system to the point where economic growth and welfare suffer materially.

4.2. Future Directions

Development of evaluation techniques [2]-Visualisation tools can support a wide scale of applications even for more specific purposes, especially for depicting degrees of risk and uncertainty in financial data, moreover, identifying gaps and quality issues in raw source data. As tooling emerges to address these various concerns, evaluation techniques will be needed to assess effectiveness [26].

Definition of systemic risks-The calculation of the average liquidity coverage ratio or risk-weighted capital ratio for each subset is one of the significant keys in where economic growth and welfare suffer. Therefore, it is important to unify the definition of systemic risk and the concept of more specific mechanisms to achieve the protection of the functioning of a financial system.

5. CONCLUSION

This paper mainly reviews the applications of various clustering methods, which include K-mean, MST, and the hierarchical approach, This paper first summarises the links between the research topics and the three main methods, then gives introduction to these three clustering methods.This paper briefly analysis the applications of clustering methods in the financial system, which includes credit scoring, trading strategy, portfolio analysis, and the stock market.It pointed out the future research direction of clustering, for example, the further construction of evaluation techniques, et.It concludes that the research on clustering methods should be more in-depth, and the applications of clustering can still be applied more widely. This paper assists other researchers choose the most suitable algorithms for their explorations and provides an analytical basis for those who are interested in this topic. This paper is dedicated to the wide uses of different clustering methods in various fields in the future, and contribute to future economic development.

REFERENCES

[1] Marti, G., Very, P., Donnat, P., & Nielsen, F. (2015). A proposal of a methodological framework with experimental guidelines to investigate clustering stability on financial time series. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). https://doi.org/10.1109/icmla.2015.11

[2] Flood, M. D., Lemieux, V. L., Varga, M., & Wong, B. L. (2014). The application of visual analytics to financial STABILITY MONITORING. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2438194

[3] Forgy, Edward W. (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics. 21 (3): 768–769.

[4] Fang, Z., & Chiao, C. (2021). Research on prediction and recommendation of financial stocks based on K-means clustering algorithm optimization. Journal of Computational Methods in Sciences and Engineering, 1–9. https://doi.org/10.3233/JCM-204716

[5] Tripathi, D., Edla, D. R., Kuppili, V., Bablani, A., & Dharavath, R. (2018). Credit scoring model based on weighted voting and cluster-based feature selection. Procedia Computer Science, 132, 22–31. https://doi.org/10.1016/j.procs.2018.05.055

[6] Bini, B. S., & Mathew, T. (2016). Clustering and regression techniques for stock prediction. Procedia Technology, 24, 1248–1255. https://doi.org/10.1016/j.protcy.2016.05.104

[7] Clustering approach to stock market prediction. (n.d.). Retrieved October 23, 2021, from https://www.ijana.in/papers/V3I4-10.pdf.

[8] Pettie, Seth; Ramachandran, Vijaya (2002), "Minimizing randomness in the minimum spanning tree, parallel connectivity, and set maxima algorithms", Proc. 13th ACM-SIAM Symposium on Discrete Algorithms (SODA '02), San Francisco, California, pp. 713–722.

[9] Ren F, Lu Y-N, Li S-P, Jiang X-F, Zhong L-X, Qiu T (2017) Dynamic Portfolio Strategy Using Clustering Approach. PLoS ONE 12(1): e0169299. https://doi.org/10.1371/journal.pone.0169299

[10] H. Kaya, Eccentricity in asset management, Journal of Network Theory in Finance 1 (2014) 1–32.

[11] F. Pozzi, T. Di Matteo, T. Aste, Spread of risk across financial markets: better to invest in the peripheries, Scientific reports 3 (2013).

[12] G. Peralta, A. Zareei, A network approach to portfolio selection, Journal of Empirical Finance (2016).

[13] Musmeci, N., Aste, T., & Di Matteo, T. (2014). Relation between financial market structure and the real economy: Comparison between clustering methods. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2525291

[14] J. Gava, W. Lefebvre, J. Turc, Beyond carrying and momentum in government bonds, Available at SSRN 3446653 (2019).

[15] N. Musmeci, T. Aste, T. Di Matteo, Risk diversification: a study of persistence with a filtered correlation-network approach, Network Theory in Finance 1 (2015) 77–98.

[16] Lemieux, V., Rahmdel, P. S., Walker, R., Wong, B. L., & Flood, M. (2014). Clustering techniques and their effect on portfolio formation and risk analysis. Proceedings of the International Workshop on Data Science for Macro-Modeling - DSMM'14. https://doi.org/10.1145/2630729.2630749

[17] HSIEH, N. (2005). Hybrid mining approach in the design of credit scoring models. Expert Systems with Applications, 28(4), 655–665. https://doi.org/10.1016/j.eswa.2004.12.022

[18] Clustering-classification based prediction of stock market ... (n.d.). Retrieved October 23, 2021, from http://www.ijcsit.com/docs/Volume%205/vol5issue03/ijcsit2014050328.pdf.

[19] Nanda, S. R., Mahanty, B., & Tiwari, M. K. (2010). Clustering Indian stock market data for portfolio management. Expert Systems with Applications, 37(12), 8793–8798. https://doi.org/10.1016/j.eswa.2010.06.026

[20] Tola, V., Lillo, F., Gallegati, M., & Mantegna, R. N. (2007, April 2). Cluster Analysis for portfolio optimization. Journal of Economic Dynamics and Control. Retrieved October 23, 2021, from https://www.sciencedirect.com/science/article/pii/S0165188907000462.

[21] Dose, C., & Cincotti, S. (2005). Clustering of financial time series with application to index and Enhanced Index Tracking portfolio. Physica A: Statistical Mechanics and Its Applications, 355(1), 145–151. https://doi.org/10.1016/j.physa.2005.02.078

[22] Narayan, P. K., Narayan, S., & Popp, S. (2011). Investigating price clustering in the oil futures market. Applied Energy, 88(1), 397–402. https://doi.org/10.1016/j.apenergy.2010.07.034

[23] Fricke, D. (2012). Trading strategies in the overnight money market: Correlations and clustering on the e-MID trading platform. Physica A: Statistical Mechanics and Its Applications, 391(24), 6528–6542. https://doi.org/10.1016/j.physa.2012.07.045

[24] Caballero, R. J. (2009), "The 'Other' Imbalance and the Financial Crisis," Working paper No.09-32, Massachusetts Institute of Technology.

[25] Mishkin, F. S. (2007), "Systemic Risk and the International Lender of Last Resort," technical report, Board of Governors of the Federal Reserve, Speech delivered at the Tenth Annual International Banking Conference, Federal Reserve Bank of Chicago, September 28, 2007.

http://www.federalreserve.gov/newsevents/speech/mishkin20070928a.htm

[26] Kapadia, S., Drehmann, M., Elliott, J. and Sterne, G. (2009), "Liquidity Risk, Cash Flow Constraints, and Systemic Feedbacks," technical report, Bank of England.

[27] Rosengren, E. S. (2010), "Asset Bubbles and Systemic Risk," technical report, Federal Reserve Bank of Boston, Speech delivered at the Global Interdependence Center's Conference on "Financial Interdependence in the World's Post-Crisis Capital Markets," Philadelphia, March 3, 2010.http://www.bos.frb.org/news/speeches/rosengren/2010/030310/030310.pdf

[28] Moussa, A. (2011), "Contagion and Systemic Risk in Financial Networks," PhD thesis, Columbia University.

[29] Financial Stability Board (2009), "Guidance to Assess the Systemic Importance of Financial Institutions, Markets and Instruments: Initial Considerations," technical report, Financial Stability Board. http://www.financialstabilityboard.org/publications/r_091107c.pdf

[30] Bisias, D., Flood, M., Lo, A. and Valavanis, S. (2012), "A Survey of Systemic Risk Analytics," Annual Review of Financial Economics, 4, 255–296.

[31] Billio, M., Getmansky, M., Lo, A. W. and Pelizzon, L. (2010), "Econometric measures of systemic risk in the finance and insurance sectors," NBER working paper 16223, National Bureau of Economic Research.

[32] Acharya, V., Pedersen, L., Philippon, T. and Richardson, M. (2010), "Measuring Systemic Risk," Working paper, New York University.

[33] Group of Ten (G-10) (2001), "Report on Consolidation in the Financial Sector: Chapter III.Effects of consolidation on financial risk," technical report, International Monetary Fund. http://www.imf.org/external/np/g10/2001/01/eng/index.htm