International Journal of Research and Review
Vol. 10; Issue: 9; September 2023
Website: www.ijrrjournal.com
Review Paper *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

A Review on Appropriateness of Partitional Clustering Algorithms in Handling Transactional Data

Andre Hasudungan Lubis[1], Elysa Ramayana[2]

[1,2]Faculty of Engineering,
[1,2]Universitas Medan Area, Medan, Indonesia

Corresponding Author: Andre Hasudungan Lubis

DOI: https://doi.org/10.52403/ijrr.20230918

ABSTRACT

Clustering is an unsupervised learning that widely used in vast researches area. The technique also utilized in any disciplines that involves multivariate data analysis. In term of transactional data handling, the partitional clustering is promoted as the one method to explore knowledge from several attributes that are related the business. In this paper, we investigate the use of partitional clustering algorithms including k-means, k-medoids, Fuzzy C Means, CLARA, and CLARANS. The present article delineates the various stages that are integral to accomplishing a review. These stages encompass data collection, data pre-processing, determination of the number of clusters, implementation of algorithms, and evaluation of clustering. The study pointed out that k-medoids as the most potential to be implemented in handling transactional data. The algorithm has achieved commendable scores in two out of three metrics, namely Calinski-Harabasz Index and Silhouette Index but not for the Davies-Bouldin Index. Nevertheless, k-medoids algorithm emerges as a formidable tool in handling the transactional data and facilitating enhanced decision-making. It is our hope that the knowledge acquired from this research will leading to progress in diverse fields where transactional data holds a crucial position.

Keywords: Partitional Clustering, Transactional Data, k-means, k-medoids, Fuzzy C Mean, CLARA, CLARANS

INTRODUCTION

Over the past few decades, machine learning has made significant strides in various fields, such as industry, environment, and business. The advent of machine learning has had a profound impact on scrutinize data and tackle intricate predicaments, thereby ushering in a novel epoch of ingenuity and prospects. (Cioffi et al., 2020). Machine learning usually divided into two approaches, the supervised learning and the unsupervised learning. As one of unsupervised learning approach, clustering is a fundamental concept in entailing the exploration and organization of data without the aid of labelled outcomes. In this approach, the algorithm endeavors to reveal concealed patterns and relationships within the dataset by grouping similar data points into distinct clusters (Liu et al., 2022).

Furthermore, clustering is frequently used in any discipline that involves multivariate data analysis (Lipovetsky, 2022). Hence, it leads to the popularity of this technique that brings the creation a substantial body of literature that underscores the significance of data grouping. Additionally, a various of scientific and applied domains have leveraged the clustering methodologies, brings numerous algorithms having been implemented in several cases in this regard (Ariza Colpas et al., 2020). The utilization of clustering analysis has become an essential tool in the extraction of patterns from bulk

data, thereby facilitating the process of valuable knowledge discovery. This is attributed to the substantial amount of data that is maintained across various domains (Ghosal et al., 2020).

In term of business, clustering also can be used to modify the parameter configuration to suit the specific attributes of the dataset, with the objective of extracting customized patterns from transactional data. Therefore, it provides the knowledge or references to the stakeholder as the decision support (Guidotti et al., 2017). Transactional data refers to the information that is gathered from transactions. This data encompasses various details such as the time and quantity of product sale, the price of the purchased item, the payment method employed, and other attributes that are related the business (Ajah & Nweke, 2019). The collection and analysis of transaction data is crucial for businesses as it provides valuable insights into consumer behaviour, preferences, and trends. By leveraging this data, businesses can make informed decisions regarding their marketing strategies, inventory management, and pricing policies, among other things, to enhance their overall performance and profitability (Azcoitia & Laoutaris, 2022). Therefore, the clustering algorithms should be carefully selected to ensure the appropriateness towards the transactional data.

There are various types of clustering algorithms that can be implemented in certain cases. The selection of a particular clustering algorithm depends on the nature of the data and the desired outcome (Mahdi et al., 2021). For instance, partitional clustering is considered as the one to handling the business sector. This is purpose to gain insights from patterns within sets of transactions, including uncover groups of similar transactions, revealing common behaviours, preferences, or trends among customers (Zada et al., 2022). There are several algorithms that categorized as the partitional clustering. However, the algorithms of k-means, k-medoids, Fuzzy C Means, CLARA, and CLARANS are widely recognized as prominent partitioning clustering (Mahdi et al., 2021).

This paper provides a comprehensive review of the most suitable partitional clustering algorithm for managing transactional data. Specifically, this investigation assesses the appropriateness of those algorithms in effectively partitioning transactional data into meaningful clusters that can reveal valuable insights. This review aims to furnish guidance on the selection of the most appropriate partitional clustering algorithm for the purpose of extracting actionable knowledge from transactional datasets.

MATERIALS & METHODS

Research Stages

Several stages are determined to explain the flow in obtaining the research objective. The first stage is involving the data collection, followed by data pre-processing, clusters determination, and the algorithms implementation towards dataset. In the final stage, clustering evaluation is performed to rank appropriateness of the algorithms for the transactional data. Figure 1 illustrates a visual representation of the aforementioned stages.

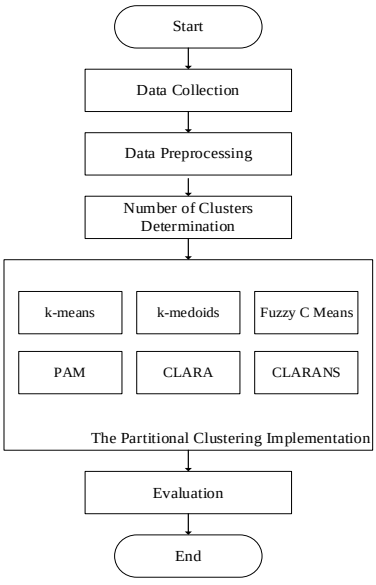

Figure 1 Research Stages

Data Collection

Data collection is the initial stage to be performed. the dataset is collected from the

online repository https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales. The data pertains to the historical sales of a supermarket company over a period of three months. The dataset comprises 1,000 rows, encompassing a diverse range of attributes such as product line, unit price, quantity, rating, and payment method. A representative sample of the data outline is presented in Table 1.

Table 1. Sample of Datasets

No.	Product line	Unit price	Quantity	Payment	Rating
1	Health and beauty	74.69	7	E-wallet	9.1
2	Electronic accessories	15.28	5	Cash	9.6
3	Home and lifestyle	46.33	7	Credit card	7.4
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
1000	Fashion accessories	88.34	7	Cash	6.6

Data Preprocessing

This stage encompasses a range of tasks, which include the elimination of redundant data, rectification of inaccuracies in the data, addressing instances of missing data, and other related activities. Moreover, data transformation is also conducted. The categorical data such as product line and payment are transformed into numerical form through the utilization of Label Encoding. The purpose of this is to make it suitable for mathematical algorithms and models that require numerical data as input. (Houssein et al., 2022). Table 2 shows the numerical form of those attributes.

Table 2. Transformation of Categorical Attributes into Numerical

Attribute	Categorical Form	Numerical Form
Product Line	Health and beauty	1
	Electronic accessories	2
	Home and lifestyle	3
	Sports and travel	4
	Food and beverages	5
	Fashion accessories	6
Payment	E-wallet	1
	Cash	2
	Credit card	3

Number of Clusters Determination

The third stage of this research aims to determine the optimal value for clustering. This is essential to identify clusters that reflect the intrinsic groupings present in the data, and obtain the acquisition of noteworthy observation and the derivation of pragmatic conclusions (Ezugwu et al., 2022). To ascertain the optimal number of clusters, the study employed the Elbow method, which compares the error of each proposed cluster number using the Sum of Squared Error (SSE) that forms several points (Jollyta et al., 2023). The SSE value can be obtained by using Equation (1). Optimal clustering results can be achieved by selecting a cluster value that is as small as possible and as large as necessary, as indicated by the distinct flattening of the curve at a particular cluster value, commonly referred to as the elbow (Weißer et al., 2020).

$$SSE = \sum_{K=1}^{K} \sum_{x_i \epsilon S_K} \|x_i - c_k\|_2^2 \quad (1)$$

The Partitional Clustering Implementation

There are five algorithms to be implemented on the transactional datasets, namely k-means, k-medoids, Fuzzy C Means, CLARA, and CLARANS in sequence. The k-means algorithm is a widely used clustering technique that partitions data into k distinct

clusters, with each data point assigned to the cluster with the nearest mean (Ahmed et al., 2020). Another clustering algorithm, k-medoids, is similar to k-means but employs medoids (actual data points) as representatives of each cluster instead of the mean (Lund & Ma, 2021).

On the other hand, Fuzzy C Means is a clustering algorithm that assigns a degree of membership to each data point for each cluster, allowing for soft clustering and handling of ambiguous data (Arora et al., 2019). Furthermore, CLARA is an algorithm which designed for clustering large datasets, utilizes a random sample to create subsets of the data, performs k-medoids on each subset, and then refines the best medoids found. On the other hand, CLARANS, another clustering algorithm suitable for large datasets, employs a randomized search to identify the optimal medoids configuration (Schubert & Rousseeuw, 2021). These algorithms offer diverse approaches to clustering transactional datasets, providing flexibility and options based on the specific requirements and characteristics of the data.

Evaluation

In the process of evaluating clustering algorithms, it is imperative to take into account various intrinsic measures. Among the commonly employed metrics are the Calinski-Harabasz Index, the Davies-Bouldin Index, and the Silhouette Index (Jin et al., 2021). The Calinski-Harabasz Index is a metric that quantifies the proportion of variance between clusters to that within clusters, whereby larger values denote superior differentiation among clusters. This measure is valuable in assessing algorithms that strive to optimize the differentiation between clusters (Wang & Xu, 2019).

On the other hand, The Davies-Bouldin Index measures the average similarity between each cluster and its most similar cluster, where similarity is based on the ratio of the within-cluster distances to between-cluster distances. Lower values indicate better clustering results. This can be useful for algorithms that aim to minimize intra-cluster similarity and maximize inter-cluster similarity (Mughnyanti et al., 2020). The Silhouette Index measures the separation distance between clusters and how well each data point fits into its assigned cluster. The index ranges from -1 to 1, where higher values indicate better clustering results. A score of 0 indicates overlapping clusters and negative scores indicate misclassification (Dudek, 2020).

RESULT

Determination of Cluster Numbers

The number of clusters used for each algorithm is determined by using the Elbow method. In order to obtain the “elbow” in a line plot, SSE for each cluster number is calculated by using Equation 1. In this study, we proposed a range of 2 to 11 clusters. The results of the Elbow method are illustrated in Table 3 and Figure 2.

Table 3. The Value of SSE from cluster 2 to 11

k	SSE
2	4214.35
3	3760.41
4	3370.39
5	3052.73
6	2793.26
7	2662.25
8	2520.41
9	2331.91
10	2226.29
11	2128.08

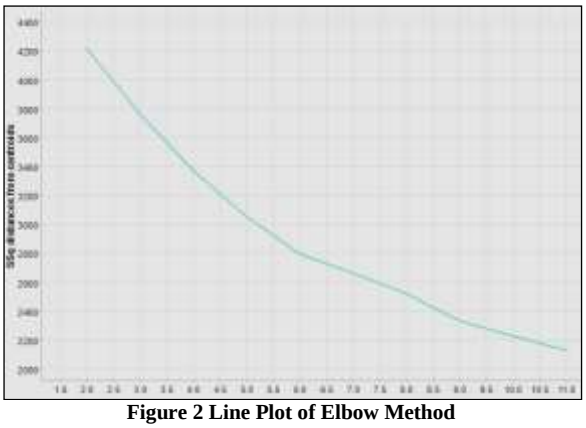

Figure 2 Line Plot of Elbow Method

Based on Figure 2, the plot indicates a sharp decline in the SSE value at the coordinate point of the number of clusters ranging from 2 to 5. Subsequently, there is a gradual decrease in the SSE value from the coordinate point of the number of clusters 6 until the last cluster number. Based on this analysis, it can be inferred that the optimal number of clusters is 6, as the elbow point in the line plot is observed at the coordinate point.

The Algorithms Implementation

This study aims to evaluate the suitability of five partitional clustering algorithms for handling transactional data. The Elbow method was employed to determine the optimal number of clusters, resulting in 6 clusters that correspond to the demand level of product sales. These clusters are categorized as "Very highly demanded" (C1), "Highly demanded" (C2), "Slightly demanded" (C3), "Moderately demanded" (C4), "Low demanded" (C5), and "Very low demanded" (C6). Each algorithm was designed to cluster the demand level of product sales using its unique set of rules and procedures. The clustering outcomes of each algorithm are presented in Table 4.

Table 4. Clustering Results from each Algorithms

Cluster	Total Data				
	k-means	k-medoids	Fuzzy C Means	CLARA	CLARANS
C1	133	176	136	140	148
C2	167	147	188	181	173
C3	138	177	147	163	180
C4	187	143	181	184	154
C5	207	194	192	155	170
C6	168	163	156	177	175

Based on Table 4, the clustering algorithms have generated various of total data. The k-means algorithm has produced the highest data from the cluster of C5 which having a total of 207. Similarly, the k-medoids algorithm has also generated C5 as the most populated cluster, with 194 data points. The Fuzzy C Means clustering algorithm has also identified C5 as the prominent cluster among others. On the other hand, the CLARA algorithm has identified C4 as having the highest value in accordance with the clustering result. Additionally, the CLARANS clustering algorithm has indicated that C3 has the highest amount of data. Figure 3 shows the scatter plots of each clustering algorithm in accordance with the result.

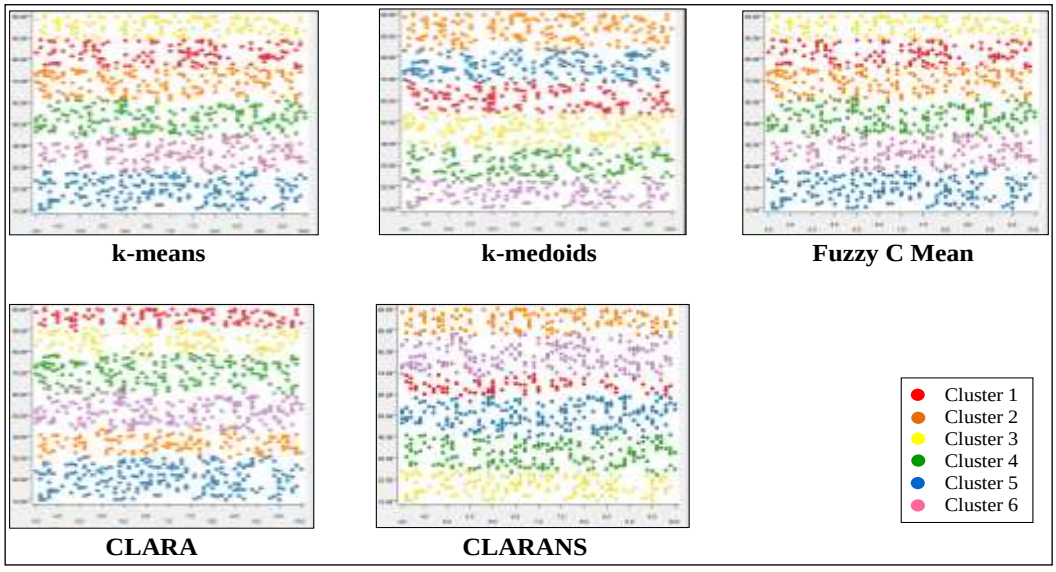

Figure 3 Scatter Plot of each Partitional Clustering Algorithm

Evaluation

There are three aforementioned metrics that used as the evaluation for the clustering quality for each algorithm, namely Calinski-Harabasz Index, Davies-Bouldin Index, and Silhouette Index. The metrics assess the efficacy of the five clustering algorithms and determining most appropriate one for transactional dataset. The results of the metrics evaluation are shown in Table 5.

Table 5. Result of Clustering Evaluation of each Algorithm

Partitional Clustering Algorithm	Calinski-Harabasz Index	Davies-Bouldin Index	Silhouette Index
k-means	324.706	0.317	0.404
k-medoids	360.588	0.345	0.418
Fuzzy C Mean	316.471	0.325	0.254
CLARA	298.235	0.373	0.377
CLARANS	244.118	0.323	0.384

As shown in Table 5, all the partitional clustering algorithms have various metric scores. The Calinski-Harabasz Index metric indicates that k-medoids has the highest score among the algorithms, followed by k-means and Fuzzy C Mean. k-medoids algorithm excels in this regard with the value of 360.588. Hence, it resulting the highest score among all the algorithms under scrutiny. Following closely behind are the k-means and Fuzzy C Mean algorithms. Furthermore, the Davies-Bouldin Index metric shows that all algorithms have similar scores. Nevertheless, k-means has the lowest score as 0. 317, indicating its ability to produce a better clustering result. However, k-medoids again appears as the highest score of Silhouette Index metrics with the value of 0.418. These findings suggest that k-medoids is a promising algorithm for clustering tasks, while k-means may be suitable for producing high-quality results in certain scenarios.

In light of these findings, it can be asserted that the k-medoids algorithm displays significant potential in clustering transactional data. Its ability to produce distinct clusters and optimize inter-cluster variance renders it a formidable candidate to be used in wide range of applications.

CONCLUSION

This article provides a comprehensive review of partitional clustering algorithms for managing transactional data. By means of rigorous analysis and evaluation, we have emphasized the utmost significance of method selection, highlighting the crucial nature of aligning the clustering algorithm choice with the distinctive attributes of the dataset in question and the intended objective. Our exhaustive examination encompassed prominent partitional clustering techniques, including k-means, k-medoids, Fuzzy C Means approach, CLARA, and CLARANS methods based on the nature of the data and desired outcome. Based on the findings, k-medoids algorithm pointed as the algorithm that has standout performer in the domain of clustering transactional data. Its remarkable ability to effectively navigating intricate transactional datasets, discerning unique patterns, and establishing clearly defined clusters has established it as a compelling option for analysing data such business. This is due to the result of clustering metrics evaluation highlighted that k-medoids show a better performance rather that other algorithms

Declaration by Authors

Acknowledgement: None

Source of Funding: None

Conflict of Interest: The authors declare no conflict of interest.

REFERENCES

1. Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.
2. Ajah, I. A., & Nweke, H. F. (2019). Big data and business analytics: Trends, platforms, success factors and applications. *Big Data and Cognitive Computing*, *3*(2), 32.
3. Ariza Colpas, P., Vicario, E., De-La-Hoz-Franco, E., Pineres-Melo, M., Oviedo-Carrascal, A., & Patara, F. (2020).

Unsupervised human activity recognition using the clustering approach: A review. *Sensors*, *20*(9), 2702.
4. Arora, J., Khatter, K., & Tushir, M. (2019). Fuzzy c-means clustering strategies: A review of distance measures. *Software Engineering: Proceedings of CSI 2015*, 153–162.
5. Azcoitia, S. A., & Laoutaris, N. (2022). A survey of data marketplaces and their business models. *ACM SIGMOD Record*, *51*(3), 18–29.
6. Cioffi, R., Travaglioni, M., Piscitelli, G., Petrillo, A., & De Felice, F. (2020). Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions. *Sustainability*, *12*(2), 492.
7. Dudek, A. (2020). Silhouette index as clustering evaluation tool. *Studies in Classification, Data Analysis, and Knowledge Organization*, 19–33.
8. Esra, M., & Sevilen, Ç. (2021). Factors influencing EFL students' motivation in online learning: A qualitative case study. *Journal of Educational Technology and Online Learning*, *4*(1), 11–22.
9. Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743.
10. Ghosal, A., Nandy, A., Das, A. K., Goswami, S., & Panday, M. (2020). A short review on different clustering techniques and their applications. *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, 69–83.
11. Guidotti, R., Monreale, A., Nanni, M., Giannotti, F., & Pedreschi, D. (2017). Clustering individual transactional data for masses of users. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 195–204.
12. Houssein, E. H., Abohashima, Z., Elhoseny, M., & Mohamed, W. M. (2022). Machine learning in the quantum realm: The state-of-the-art, challenges, and future vision. *Expert Systems with Applications*, *194*, 116512.
13. Ismi, D. P., & Murinto, M. (2020). Clustering based feature selection using Partitioning Around Medoids (PAM). *Jurnal Informatika Ahmad Dahlan*, *14*(2), 50–57.
14. Jin, J., Heimann, M., Jin, D., & Koutra, D. (2021). Toward understanding and evaluating structural node embeddings. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *16*(3), 1–32.
15. Jollyta, D., Efendi, S., Zarlis, M., & Mawengkang, H. (2023). Analysis of an optimal cluster approach: a review paper. *Journal of Physics: Conference Series*, *2421*(1), 12015.
16. Lipovetsky, S. (2022). Multivariate statistical methods: A brief review on their modifications and applications. *Model Assisted Statistics and Applications*, *17*(2), 145–147.
17. Liu, T., Yu, H., & Blair, R. H. (2022). Stability estimation for unsupervised clustering: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(6), e1575.
18. Lund, B., & Ma, J. (2021). A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering. *Performance Measurement and Metrics*, *22*(3), 161–173.
19. Mahdi, M. A., Hosny, K. M., & Elhenawy, I. (2021). Scalable clustering algorithms for big data: A review. *IEEE Access*, *9*, 80015–80027.
20. Mughnyanti, M., Efendi, S., & Zarlis, M. (2020). Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation. *IOP Conference Series: Materials Science and Engineering*, *725*(1), 12128.
21. Schubert, E., & Rousseeuw, P. J. (2021). Fast and eager k-medoids clustering: O (k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Information Systems*, *101*, 101804.
22. Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering*, *569*(5), 52024.
23. Weißer, T., Saßmannshausen, T., Ohrndorf, D., Burggräf, P., & Wagner, J. (2020). A clustering approach for topic filtering within systematic literature reviews. *MethodsX*, *7*, 100831.

24. Zada, I., Ali, S., Khan, I., Hadjouni, M., Elmannai, H., Zeeshan, M., Serat, A. M., & Jameel, A. (2022). Performance evaluation of simple K-mean and parallel K-mean clustering algorithms: big data business process management concept. *Mobile Information Systems*, *2022*, 1–15.

How to cite this article: Andre Hasudungan Lubis, Elysa Ramayana. A review on appropriateness of partitional clustering algorithms in handling transactional data. *International Journal of Research and Review*. 2023; 10(9): 162-169.
DOI: *https://doi.org/10.52403/ijrr.20230918*

******