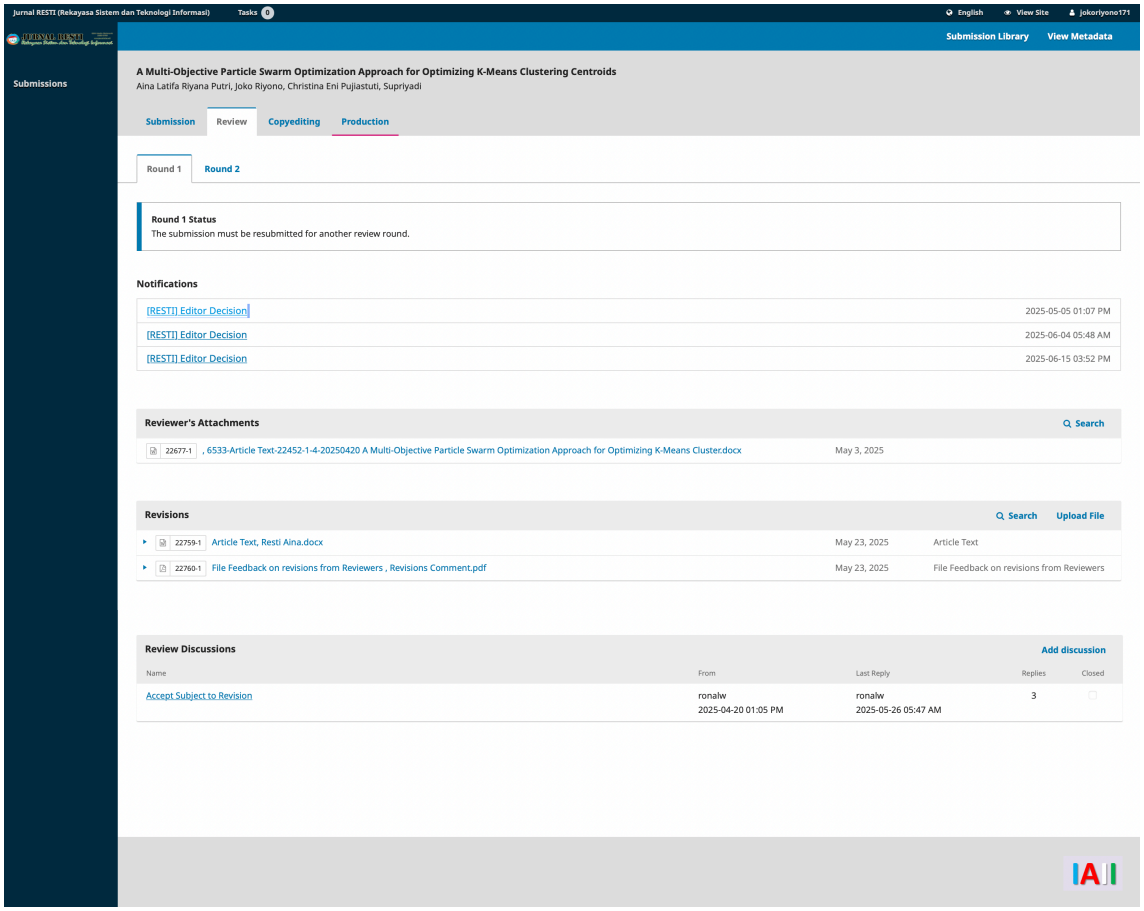
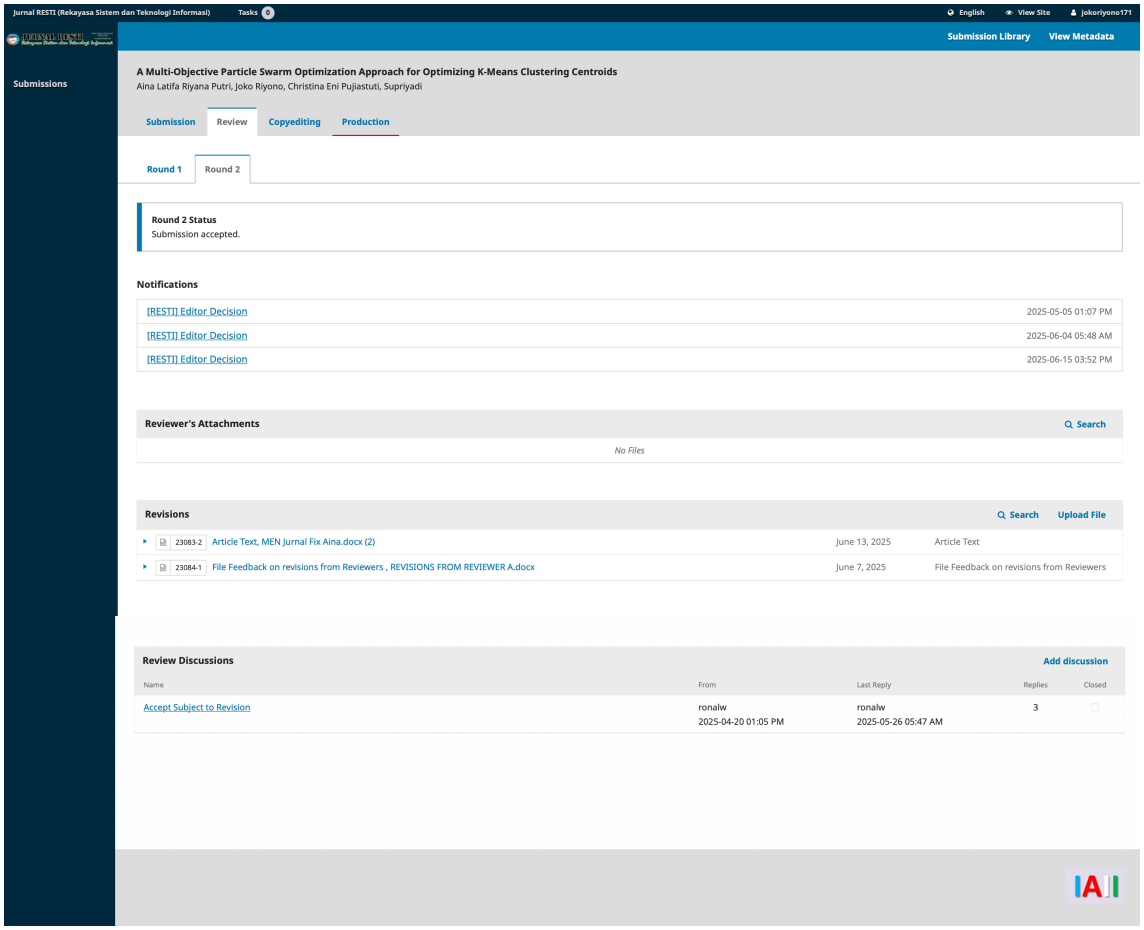

Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)
Tasks 0
English
View Site
jokoriyono171
Submission Library
View Metadata
Submissions
A Multi-Objective Particle Swarm Optimization Approach for Optimizing K-Means Clustering Centroids
Aina Latifa Riyana Putri, Joko Riyono, Christina Eni Pujiastuti, Supriyadi
Submission
Review
Copyediting
Production
Round 1
Round 2
Round 1 Status
The submission must be resubmitted for another review round.
Notifications
[RESTI] Editor Decision
2025-05-05 01:07 PM
[RESTI] Editor Decision
2025-06-04 05:48 AM
[RESTI] Editor Decision
2025-06-15 03:52 PM
Reviewer's Attachments
Search
22677-1
, 6533-Article Text-22452-1-4-20250420 A Multi-Objective Particle Swarm Optimization Approach for Optimizing K-Means Cluster.docx
May 3, 2025
Revisions
Search
Upload File
22759-1
Article Text, Resti Aina.docx
May 23, 2025
Article Text
22760-1
File Feedback on revisions from Reviewers , Revisions Comment.pdf
May 23, 2025
File Feedback on revisions from Reviewers
Review Discussions
Add discussion
Name
From
Last Reply
Replies
Closed
Accept Subject to Revision
ronalw
2025-04-20 01:05 PM
ronalw
2025-05-26 05:47 AM
3


Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)
Tasks 0
English
View Site
jokoriyono171
Submission Library
View Metadata
Submissions
A Multi-Objective Particle Swarm Optimization Approach for Optimizing K-Means Clustering Centroids
Aina Latifa Riyana Putri, Joko Riyono, Christina Eni Pujiastuti, Supriyadi
Submission
Review
Copyediting
Production
Round 1
Round 2
Round 2 Status
Submission accepted.
Notifications
[RESTI] Editor Decision 2025-05-05 01:07 PM
[RESTI] Editor Decision 2025-06-04 05:48 AM
[RESTI] Editor Decision 2025-06-15 03:52 PM
Reviewer's Attachments
Search
No Files
Revisions
Search
Upload File
23083-2 Article Text, MEN Jurnal Fix Aina.docx (2) June 13, 2025 Article Text
23084-1 File Feedback on revisions from Reviewers , REVISIONS FROM REVIEWER A.docx June 7, 2025 File Feedback on revisions from Reviewers
Review Discussions
Add discussion
Name
From
Last Reply
Replies
Closed
Accept Subject to Revision
ronalw
2025-04-20 01:05 PM
ronalw
2025-05-26 05:47 AM
3

Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 1 (2024) x - x e-ISSN: 2580-0760

A Multi-Objective Particle Swarm Optimization Approach for Optimizing K-Means Clustering Centroids

(for the review comments, refer to the most-bottom page)

Abstract

The K-Means algorithm is a popular unsupervised learning method for data clustering. However, its performance heavily depends on centroid initialization and the distribution shape of the data, making it less effective for datasets with complex or non-linear cluster structures. This study evaluates the performance of the standard K-Means algorithm and proposes a Multiobjective Particle Swarm Optimization K-Means (MOPSO-KMeans) approach to improve clustering accuracy. The evaluation was conducted on five benchmark datasets: Atom, Chainlink, EngyTime, Target, and TwoDiamonds. Experimental results show that K-Means is effective only on datasets with clearly separated clusters, such as EngyTime and TwoDiamonds, achieving accuracies of 95.6% and 100%, respectively. In contrast, MOPSO-KMeans demonstrated improved performance on datasets with non-linear structures, such as Target and Chainlink, with the highest accuracy reaching 59.2%. The evaluation used metrics including Sum of Square Within (SSW), Sum of Square Between (SSB), best accuracy, and standard deviation. The results indicate that MOPSO-KMeans provides more stable and consistent clustering outcomes compared to conventional K-Means. These findings support the application of swarm-based optimization for clustering tasks on datasets with high complexity.

Keywords: *Multiobjective Particle Swarm Optimization; K-Means; Centroid; The Sum of Square Within; The Sum of Square Between*

How to Cite: [Caption completed by the editor]
DOI:

1. Introduction

Clustering is one of the techniques in data mining that aims to group data into several clusters based on the similarity of their characteristics. One of the most popular clustering methods is k-Means clustering [1], which is a non-hierarchical cluster analysis method. This algorithm works by randomly initializing a set number of centroids, then assigning objects to *k* clusters based on their distance to these centroids, and iteratively updating the centroid positions until convergence is reached.

In k-Means clustering, the main objective is to form optimal clusters in which the members of each cluster are highly similar to one another, while being significantly different from members of other clusters [2]. To achieve this goal, two primary metrics are commonly used: the Sum of Squares Within-cluster (SSW) and the Sum of Squares Between-cluster (SSB). SSW measures cluster compactness, indicating how closely data points within a cluster are grouped around their centroid [3]. A smaller SSW value implies greater similarity among data points within the same cluster. On the other hand, SSB measures the distance between cluster centroids, which reflects the separation between [4]. A larger SSB value indicates greater distance between clusters, thus making the clustering more effective in distinguishing different data groups.

The k-Means algorithm is widely recognized as an efficient and scalable method for processing large datasets [5]. Several previous studies, such as those by [6] and [7], have indicated that k-Means can produce more compact clusters compared to hierarchical clustering methods. However, k-Means has several limitations, particularly regarding the random selection of initial centroids, which can lead to varying clustering results each time the algorithm is executed [8]. Additionally, k-Means tends to get trapped in local optima, which may result in suboptimal cluster

Received: xx-xx-xxxx | Accepted: xx-xx-xxxx | Published Online: xx-xx-xxxx

assignments [9]. Its sensitivity to outliers is also a major concern, as extreme values can significantly shift the centroid positions [10]. Furthermore, the assumption that clusters are spherical [11] and of uniform size makes k-Means less effective when dealing with datasets containing complex-shaped clusters or varying densities.

To address these issues, this study proposes an optimization approach based on Multi-Objective Particle Swarm Optimization (MOPSO) to enhance the performance of k-Means in determining more optimal centroids. MOPSO is a variant of Particle Swarm Optimization (PSO) [12] designed to handle multi-objective optimization problems. In the context of clustering, MOPSO aims to simultaneously minimize the Sum of Squares Within-cluster (SSW) and maximize the Sum of Squares Between-cluster (SSB), thereby producing clusters that are well-balanced in terms of both homogeneity and separation.

This approach will be evaluated using several benchmark datasets commonly used in clustering studies [13], namely Atom, Chainlink, Engytime, Target, and Two Diamonds. Each dataset presents unique challenges that can test the effectiveness of the proposed method. The Atom dataset poses a challenge in separating closely located clusters, while Chainlink contains interwoven topological structures, which are difficult for centroid-based methods to handle. Engytime features uneven density distribution, which can complicate the identification of precise cluster boundaries. The Target dataset presents non-linear patterns that standard k-Means struggles to capture, whereas Two Diamonds involves closely situated clusters, making it challenging to determine optimal separation.

The performance evaluation will be conducted by comparing the clustering results against the ground truth labels using accuracy metrics, as well as by comparing the MOPSO-based approach with conventional clustering methods such as standard k-Means. With the MOPSO-based optimization, this method is expected to produce better clusters than conventional k-Means, particularly in terms of inter-cluster separation and intra-cluster uniformity. The findings of this study are expected to contribute to the development of more optimal clustering methods for various applications in the fields of data mining and machine learning.

1. Research Methods

The methodology section will sequentially present the analytical methods employed in this study on Figure 1.

2.1 Dataset Benchmark

The first step in this study is the selection of datasets to be used for testing the effectiveness of the proposed method. The datasets used in this research are as follows [13]:

1. Atom
   The Atom dataset [14] is one of the benchmark datasets commonly used to evaluate the performance of clustering algorithms under complex conditions. This dataset on Figure 1 exists in a three-dimensional space ($\mathbb{R}^3$) and consists of two main groups of data: the core cluster and the outer hull cluster. Geometrically, the core cluster is located at the center, while the outer hull completely surrounds it. This creates a condition known as an *overlapping convex hull*, where the convex shape of one cluster (the hull) entirely encloses the other cluster (the core). As a result, the two clusters overlap and cannot be linearly separated, meaning no straight line or plane can clearly divide them.

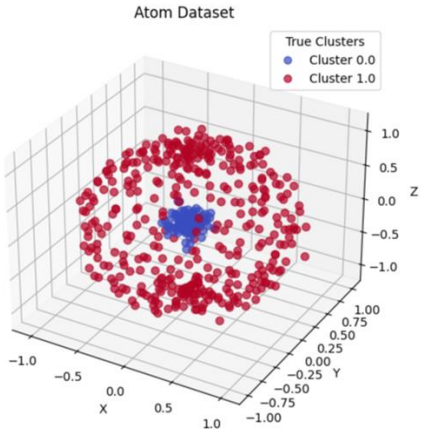


Figure 1. Atom Dataset

   The core cluster contains 100 data points, while the outer hull cluster contains 400 data points. The core cluster is much denser compared to the outer hull, meaning that the core data points are tightly packed and concentrated at the center, whereas the hull data points are more dispersed. This difference in density poses a particular challenge for algorithms such as k-Means, which rely on distance between data points to form clusters. In this case, the distance between the cluster centroids may be smaller than the spread within a single cluster, making separation more difficult.
   Therefore, the primary challenge of the Atom dataset lies in its spatial structure, where the clusters are entirely overlapped geometrically, making it very difficult to separate them effectively using centroid-based clustering algorithms such as k-Means.
2. ChainLink
   The Chainlink dataset [15];[16] is one of the benchmark datasets designed to evaluate the ability of clustering algorithms to handle complex, interrelated data structures. This dataset on Figure 2 consists of two clusters, each containing 300 data points, which together form a structure resembling interlinked chains in three-dimensional space ($\mathbb{R}^3$).

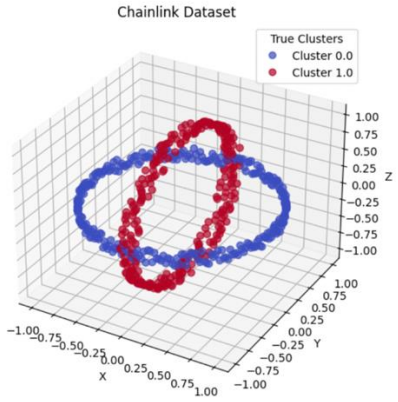


Figure 2. ChainLink Dataset

Each cluster in the Chainlink dataset has the shape of a ring, and the two rings are interlocked with one another, creating a structure known as a linear nonseparable entanglement. This refers to a condition where clusters cannot be separated linearly due to their intricate overlapping positions. Although the clusters appear globally distinct, many data points from one cluster are locally closer to points from the other cluster than to points within their own cluster. This creates a conflict between global separation and local proximity, posing a significant challenge for distance-based algorithms such as k-Means.

Moreover, both clusters have nearly identical average inter-point distances and densities, making it difficult to distinguish them based solely on size or distribution. The intertwined three-dimensional structure further complicates separation using linear boundaries.

3. EngyTime

The EngyTime dataset [17] is a benchmark dataset used to evaluate the capability of clustering algorithms in separating clusters that have different densities but are overlapping. This dataset consists of 2,000 data points divided into two clusters in a two-dimensional space ($\mathbb{R}^2$), with two main variables: “Engy” and “Time”.

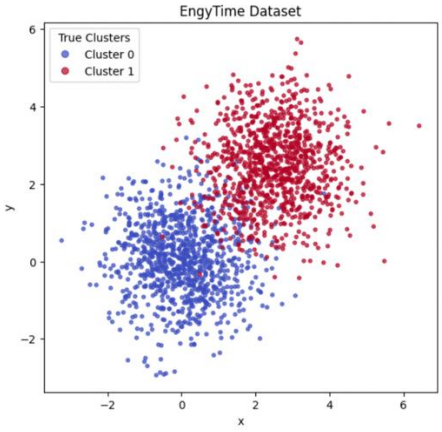


Figure 3. EngyTime Dataset

This dataset on Figure 3 represents a simplified form of a density-based problem, which frequently occurs in practice, such as in the analysis of unclassified high-dimensional flow cytometry data. EngyTime is constructed from a mixture of two-dimensional Gaussian distributions, commonly encountered in various applications, including sonar signal processing.

The main challenge of this dataset lies in the overlapping clusters, which are not separated by empty space. This means that the cluster boundaries cannot be clearly defined using only the position or distance between data points. Instead, it requires considering the density information of the data. Consequently, centroid-based algorithms like k-Means, which do not account for density variations, will struggle to accurately separate the clusters.

4. Target

The Target dataset is a benchmark dataset designed to evaluate the robustness of clustering algorithms in handling overlapping clusters and the presence of outliers [18]. It resides in a two-dimensional space ($\mathbb{R}^2$) and consists of 743 data points, divided into two main clusters and four outlier groups.

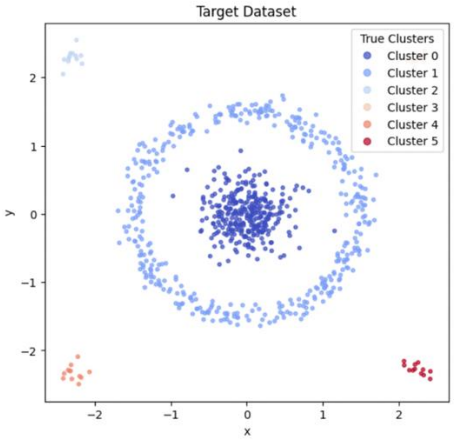


Figure 4. Target Dataset

The first cluster is a dense sphere initially containing 365 data points, while the second cluster forms a ring that surrounds the inner circle, consisting of 395 data points. These two clusters have overlapping convex hull structures, making them difficult to separate using only linear boundaries. Such geometric configuration presents a particular challenge for centroid-based algorithms like k-Means.

Additionally, the dataset on Figure 4 includes four small groups of outliers, each containing four points, located at the four corners of the space. The presence of these outliers increases the complexity of the clustering task, as they can interfere with the identification of cluster centroids or even be mistakenly interpreted as separate clusters by algorithms that are sensitive to noise.

5. TwoDiamonds

The TwoDiamonds dataset [19];[20] is a benchmark dataset designed to evaluate the performance of clustering algorithms in

recognizing weakly connected clusters, such as chain-like structures. This dataset on Figure 5 consists of two clusters, each containing 200 data points in a two-dimensional space ($\mathbb{R}^2$).

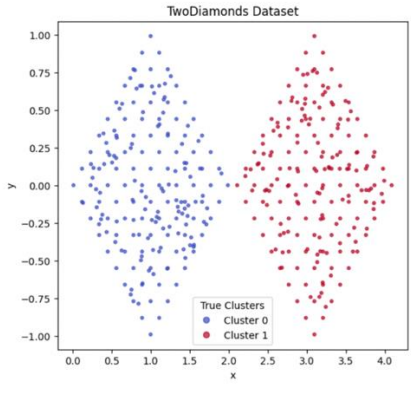


Figure 5. TwoDiamonds Dataset

Each cluster takes the shape of a diamond, with data points uniformly distributed across the area, resulting in an even spread within each cluster. Geometrically, the clusters are positioned in two adjacent square regions that nearly touch at one side, forming a structure resembling two diamonds placed close together.

The main challenge posed by this dataset is the presence of a "weak connection" area, where the two clusters nearly intersect. For clustering algorithms that rely solely on point-to-point distance, such as k-Means, this structure makes it difficult to determine whether the two areas represent a single large cluster or two distinct ones. Due to the chain-like connection between the clusters, identifying an appropriate boundary requires consideration of the overall spatial structure rather than just local proximity.

2.2 Standard K-Means Implementation

The next step is to run the standard K-Means algorithm on each dataset. K-Means works by randomly initializing centroids and then iteratively grouping data based on Euclidean distance and updating the centroids until convergence is reached. In this process, the number of clusters (k) is determined based on the number of known clusters in the ground truth dataset. K-Means forms clusters based on the proximity of data points to the centroids obtained during iteration. However, since the initial centroids are selected randomly, the clustering results may vary between runs. Therefore, it is important to evaluate the clustering quality using appropriate metrics.

The clustering results from the standard K-Means algorithm are evaluated using confusion matrix. The confusion matrix is used to compare the clustering results with the original dataset labels, which allows the calculation of clustering accuracy

2.3 Development of Multi-Objective PSO for K-Means Optimization

To improve the quality of clustering results, this study implements the Multi-Objective Particle Swarm Optimization (MOPSO) algorithm to optimize the selection of centroids in the K-Means algorithm. This approach simultaneously considers two objectives: minimizing the Sum of Squared Within-Cluster (SSW) and maximizing the Sum of Squared Between-Cluster (SSB).

1. The first objective function aims to minimize the Sum of Squared Within-Cluster (SSW):

$$\mathbf{f_1 = min}\left(\sum_{\mathbf{j=1}}^{\mathbf{k}}\sum_{\mathbf{x_i \in C_j}}\left\|\mathbf{x_i - \mu_j}\right\|^{\mathbf{2}}\right) \quad (2)$$

Where k is the number of clusters; $\boldsymbol{x_i}$ is the i-th data point; $\boldsymbol{\mu_j}$ is the centroid of cluster $\boldsymbol{C_j}$; $\left\|\boldsymbol{x_i - \mu_j}\right\|^{\mathbf{2}}$ is the squared Euclidean distance between the data point and the cluster centroid.

2. The second objective function aims to maximize the Sum of Squared Between-Cluster (SSB), which is expressed as the minimization of its negative:

$$\boldsymbol{f_2 = -min}\left(-\sum_{j=1}^{k}\boldsymbol{n_j}\left\|\boldsymbol{\mu_j - \mu}\right\|^{\mathbf{2}}\right) \quad (3)$$

Where n_j is the number of data points in cluster j; μ is the global centroid of the entire dataset; $\left\|\boldsymbol{\mu_j - \mu}\right\|^{\mathbf{2}}$ is the squared distance between the cluster centroid and the global centroid.

The goal of MOPSO is to find a set of optimal solutions (centroids) based on both objective functions simultaneously. The Pareto optimality approach is used, where the best solutions are selected based on dominance (i.e., no other solution is better in all objectives). Particles in the swarm are updated based on their personal best positions and global best positions from the Pareto archive.

Using this approach, MOPSO generates a set of candidate centroids that offer an optimal trade-off between cluster compactness (minimizing SSW) and cluster separation (maximizing SSB). The selected centroids from this solution set are then used to initialize K-Means, aiming for better clustering performance.

After MOPSO identifies the optimal centroids, the K-Means algorithm is run again using these optimized centroids. Thus, the clustering process no longer relies on random centroid initialization but instead uses optimized centroids, which are expected to yield better clustering results. The goal of this step is to determine whether the MOPSO-KMeans method can produce more stable and accurate clusters compared to standard K-Means.

To ensure the reliability of the results, both methods (standard K-Means and MOPSO-KMeans) are executed

30 times independently. This is done to observe the stability and variability of the clustering outcomes for each method. Each run produces values for SSW, SSB, and accuracy, which are then analyzed statistically. By conducting repeated independent tests, a clearer picture of the average performance and stability of the proposed method versus standard K-Means can be obtained.

3. Results and Discussions

In this section, an analysis is conducted on the clustering results obtained from the implementation of the K-Means algorithm on each benchmark dataset.

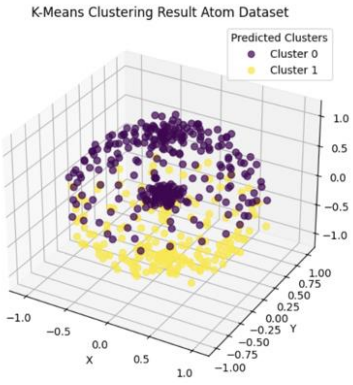


Figure 6. Atom Dataset With K-Means Clustering

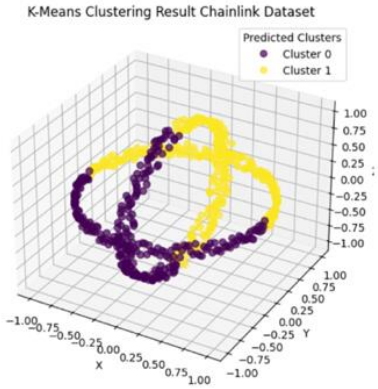


Figure 7. ChainLink Dataset With K-Means Clustering

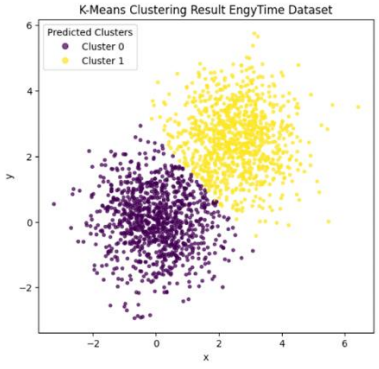


Figure 8. EngyTime Dataset With K-Means Clustering

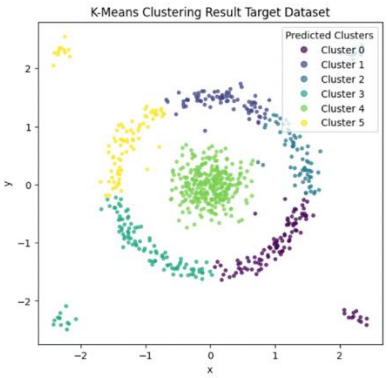


Figure 9. Target Dataset With K-Means Clustering

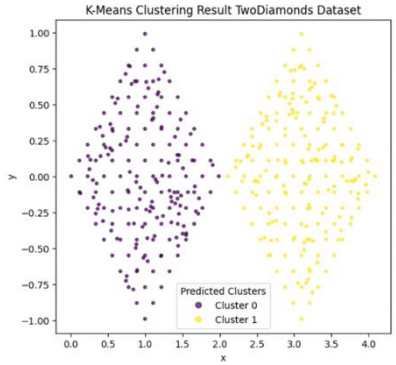


Figure 10. TwoDiamonds Dataset With K-Means Clustering

The clustering visualization results on the Atom dataset indicate that the K-Means algorithm was not able to group the data effectively. In Figure 6, which shows the clustering result using the K-Means algorithm, it is clear that the grouping does not align with the original structure. K-Means clusters the data based on the distance to the cluster centroids, resulting in two groups that appear to be split from top to bottom, rather than from center outward. As a result, many data points from the core and shell regions are incorrectly grouped.

The Chainlink dataset on Figure 7 is a synthetic dataset consisting of two interlinked rings in three-dimensional space. The K-Means algorithm was applied to cluster the data into two groups, corresponding to the actual number of clusters. K-Means begins by randomly selecting cluster centroids and then iteratively assigns data points based on their proximity to these centroids. However, due to the non-linear and complex shape of the Chainlink dataset, K-Means struggles to accurately cluster the data. This is clearly shown in the predicted clustering visualization, where the data points are incorrectly split across the two rings, rather than along their natural separation.

In the EngyTime dataset, based on the predefined ground truth labels, the two clusters appear clearly separated. Figure 8 shows the clustering result produced by the K-Means algorithm. Although K-Means is an unsupervised algorithm, the result shows that it performs fairly well on this dataset, producing two clusters that visually resemble the ground truth. The

purple and yellow points in the visualization represent a consistent mapping to the original data structure, with only a few points near the boundary areas that may have been misclassified.

For the Target dataset on Figure 9, the clustering result from the K-Means algorithm is visualized with data points colored according to their predicted clusters. A significant discrepancy can be observed when compared to the true cluster structure. The outer cluster is split into several segments, and smaller groups are not identified accurately. This indicates that the K-Means algorithm is unable to capture the complex clustering pattern in the Target dataset.

In the TwoDiamonds dataset on Figure 10, the clustering result using the K-Means algorithm is shown with points colored according to the predicted cluster labels. Although the colors do not match the original labels, the clustering pattern appears identical, demonstrating that K-Means is able to successfully identify the two-cluster structure in this dataset.

Table 1. Accuracy K-Means Clustering

Dataset	Accuracy K-means
Atom	54.4%
ChainLink	50%
EngyTime	95.6%
Target	0.2692%
TwoDiamonds	100%

Based on Table 1, it is evident that the performance of the K-Means algorithm is highly dependent on the shape and characteristics of each dataset. For datasets with simple and linearly separable cluster structures, such as TwoDiamonds and EngyTime, K-Means performs very well, achieving high accuracy—up to 100%. However, for datasets with more complex or non-linear structures, such as Atom, Chainlink, and Target, K-Means fails to cluster the data accurately. This is reflected in the low accuracy scores and clustering visualizations that do not match the true data structure. The main weaknesses of K-Means lie in two critical aspects: its reliance on random initialization of cluster centroids and its assumption that clusters are convex and linearly separable. Because K-Means depends solely on Euclidean distance to the cluster centroids, it is unable to capture circular, complex, or asymmetrical cluster patterns. Furthermore, suboptimal initial centroid selection can lead the algorithm to converge to local optima, resulting in inaccurate cluster assignments.

To address these limitations, this study proposes the use of Multi-Objective Particle Swarm Optimization (MOPSO) as an alternative approach to improve the effectiveness of data clustering. The experimental improvement was observed on the Target dataset. K-Means achieved only 26% accuracy, while MOPSO-K-Means improved the accuracy to 59.2%. This demonstrates that MOPSO-K-Means is more capable of handling datasets with complex or non-linearly separable cluster structures. Lastly, on the TwoDiamonds dataset, both K-Means and MOPSO-settings in this study were defined as follows: swarm size N = 40, and each test function was executed 30 times independently, with each run consisting of 100 iterations. All PSO algorithms were terminated upon reaching the predefined maximum number of iterations.

The performance of MOPSO-K-Means was evaluated using commonly used optimization metrics, namely the average solution and standard deviation. These metrics were used to assess the effectiveness of MOPSO-K-Means in solving the benchmark clustering tasks.

Table 2. Clustering With MOPSO-K-Means

Dataset	Item	SSW	SSB	Best Accuracy MOPSO-K-Means
Atom	Avg.	1191.22	1414.52	52.8%
	Std.	230.85	238.22	
ChainLink	Avg.	1531.74	1711.26	50.2%
	Std.	167.04	168.519	
EngyTime	Avg.	49122.71	85124.33	95.7%
	Std.	2951.153	3913.554	
Target	Avg.	4126.614	7658.074	59.2%
	Std.	892.147	1006.303	
TwoDiamonds	Avg.	1323.322	2863.340	100%
	Std.	42.822	44.191	

Table 2 presents the performance evaluation results of the MOPSO-K-Means algorithm on five benchmark datasets: Atom, ChainLink, EngyTime, Target, and TwoDiamonds. The evaluation was carried out using commonly used optimization metrics, namely the average solution and standard deviation of the SSW (Sum of Squares Within) and SSB (Sum of Squares Between), along with the best accuracy achieved for each dataset. The objective of this evaluation is to assess the effectiveness of the MOPSO-K-Means algorithm in producing optimal cluster partitions.

Compared to the conventional K-Means algorithm, the results indicate that MOPSO-K-Means generally performs better on most datasets. On the Atom dataset, K-Means achieved an accuracy of 54.4%, while MOPSO-K-Means recorded an accuracy of 52.8%. Although there was a slight decrease, the SSW and SSB values obtained by MOPSO-K-Means still reflect a good and stable cluster distribution, with relatively low standard deviations. For the ChainLink dataset, K-Means achieved 50% accuracy, while MOPSO-K-Means achieved 50.2%, suggesting a slightly better performance in separating the clusters.

Next, on the EngyTime dataset, K-Means reached an accuracy of 95.60%, while MOPSO-K-Means achieved 95.7%. The difference is very small, indicating that both algorithms are equally effective in clustering data with clear cluster structures. However, the most significant

K-Means achieved perfect accuracy (100%), indicating that this dataset has a very clear structure that can be easily separated by both algorithms.

Overall, the evaluation results show that MOPSO-K-Means has advantages in terms of flexibility and effectiveness in identifying complex cluster

structures that conventional K-Means struggles to handle. The relatively small standard deviations across most datasets also indicate that this algorithm can produce stable and consistent solutions in each optimization run. Therefore, MOPSO-K-Means can be considered a more reliable alternative for clustering tasks involving datasets with diverse characteristics.

4. Conclusions

Based on the analysis and evaluation of five benchmark datasets, it can be concluded that the performance of the K-Means algorithm is highly dependent on the shape and structural characteristics of the clusters in the data. On datasets with simple and linearly separable structures, such as TwoDiamonds and EngyTime, K-Means performs very well, achieving high accuracy—up to 100%. However, on datasets with non-linear or complex structures, such as Atom, ChainLink, and Target, the algorithm fails to properly separate clusters, resulting in low accuracy and poor alignment with the ground truth.

To address these limitations, the MOPSO-K-Means approach was introduced as an alternative solution. Based on the experimental results, this algorithm shows significant performance improvement on datasets with complex structures—most notably on the Target dataset, where the accuracy increased from 26% (K-Means) to 59.2% (MOPSO-K-Means). In addition, the obtained SSW and SSB values, along with relatively low standard deviations, indicate that MOPSO-K-Means is capable of producing stable and consistent clustering solutions.

Overall, MOPSO-K-Means has proven to be more flexible and reliable in handling various types of cluster structures, making it a more suitable choice for clustering tasks involving non-convex or non-linearly separable data distributions.

Acknowledgements

The author declare no conflict interest. This research received no spesific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] T. M. Ghazal *et al.*, "Performances of K-Means Clustering Algorithm with Different Distance Metrics," *Intelligent Automation & Soft Computing*, vol. 30, no. 2, pp. 735–742, Aug. 2021, doi: 10.32604/IASC.2021.019067.

[2] R. Gustriansyah, N. Suhandi, and F. Antony, "Clustering optimization in RFM analysis Based on k-Means," *IJEECS*, vol. 18, no. 1, pp. 470–477, Apr. 2020, doi: 10.11591/ijeecs.v18.i1.pp470-477.

[3] R. Richard, H. Cao, and M. Wachowicz, "An Automated Clustering Process for Helping Practitioners to Identify Similar EV Charging Patterns across Multiple Temporal Granularities," *International Conference on Smart Cities and Green ICT Systems*, pp. 67–77, 2021, doi: 10.5220/0010485000670077.

[4] M. T. Guerreiro *et al.*, "Anomaly Detection in Automotive Industry Using Clustering Methods—A Case Study," *Applied Sciences 2021, Vol. 11, Page 9868*, vol. 11, no. 21, p. 9868, Oct. 2021, doi: 10.3390/APP11219868.

[5] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Electronics 2020, Vol. 9, Page 1295*, vol. 9, no. 8, p. 1295, Aug. 2020, doi: 10.3390/ELECTRONICS9081295.

[6] ABDULHAFEDH, Azad. Incorporating k-means, hierarchical clustering and pca in customer segmentation. *Journal of City and Development*, 2021, 3.1: 12-30, doi: 10.12691/jcd-3-1-3.

[7] PATEL, Punyaban; SIVAIAH, Borra; PATEL, Riyam. Approaches for finding optimal number of clusters using k-means and agglomerative hierarchical clustering techniques. In: *2022 international conference on intelligent controller and computing for smart power (ICICCSP)*. IEEE, 2022. p. 1-6, doi: 10.1109/ICICCSP53532.2022.9862439

[8] A. Vouros, S. Langdell, M. Croucher, and E. Vasilaki, "An empirical comparison between stochastic and deterministic centroid initialisation for K-means variations," *Mach Learn*, vol. 110, no. 8, pp. 1975–2003, Aug. 2021, doi: 10.1007/S10994-021-06021-7/FIGURES/8.

[9] QTAISH, Amjad, et al. Optimization of K-means clustering method using hybrid capuchin search algorithm. *The Journal of Supercomputing*, 2024, 80.2: 1728-1787, doi: 0.1007/s11227-023-05540-5

[10] Z. Zhang, Q. Feng, J. Huang, Y. Guo, J. Xu, and J. Wang, "A local search algorithm for k-means with outliers," *Neurocomputing*, vol. 450, pp. 230–241, Aug. 2021, doi: 10.1016/J.NEUCOM.2021.04.028.

[11] A. A. Wani, "Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions," *PeerJ Comput Sci*, vol. 10, pp. 1–45, Aug. 2024, doi: 10.7717/PEERJ-CS.2286/FIG-14.

[12] M. Jain, V. Saihjpal, N. Singh, and S. B. Singh, "An Overview of Variants and

Advancements of PSO Algorithm," *MDPI Applied Sciences*, 2022, doi: 10.3390/app12178392.

[13] M. C. Thrun and A. Ultsch, "Clustering benchmark datasets exploiting the fundamental clustering problems," *Data Brief*, vol. 30, p. 105501, Jun. 2020, doi: 10.1016/J.DIB.2020.105501.

[14] A. Ultsch, "Strategies for an Artificial Life System to cluster high dimensional Data," 2004, Accessed: Apr. 14, 2025. [Online]. Available: https://www.researchgate.net/publication/228932819

[15] A. Ultsch, G. Guimaraes, D. Korus, and H. Li, "Knowledge Extraction from Artificial Neural Networks and Applications," *Parallele Datenverarbeitung mit dem Transputer*, pp. 148–162, 1994, doi: 10.1007/978-3-642-78901-4_11.

[16] P. Mangiameli, S. K. Chen, and D. West, "A comparison of SOM neural network and hierarchical clustering methods," *Eur J Oper Res*, vol. 93, no. 2, pp. 402–417, Sep. 1996, doi: 10.1016/0377-2217(96)00038-0.

[17] S. P. Chatzis and D. I. Kosmopoulos, "A variational Bayesian methodology for hidden Markov models utilizing Student's-t mixtures," *Pattern Recognit*, vol. 44, no. 2, pp. 295–306, Feb. 2011, doi: 10.1016/J.PATCOG.2010.09.001.

[18] J. Poelmans, M. M. Van Hulle, S. Viaene, P. Elzinga, and G. Dedene, "Text mining with emergent self organizing maps and multi-dimensional scaling: A comparative study on domestic violence," *Appl Soft Comput*, vol. 11, no. 4, pp. 3870–3876, Jun. 2011, doi: 10.1016/J.ASOC.2011.02.026.

[19] A. Ultsch, "U *-Matrix : a Tool to visualize Clusters in high dimensional Data," 2004.

[20] A. Ultsch, "Density Estimation and Visualization for Data Containing Clusters of Unknown Structure," *Studies in Classification, Data Analysis, and Knowledge Organization*, pp. 232–239, 2005, doi: 10.1007/3-540-28084-7_25.

The manuscript presents a promising hybrid approach that integrates Multi-Objective Particle Swarm Optimization (MOPSO) with K-Means for centroid optimization. The topic is relevant and falls well within the journal's scope, and the proposed method shows originality in concept.

However, the submission falls short in several critical aspects that prevent it from being accepted in its current form. These include an overly narrow experimental scope, limited comparative analysis, and underdeveloped discussion. While the method is promising, the current presentation lacks the rigor and depth required for publication. If the authors address these concerns through substantial revision and stronger validation, this manuscript, in my opinion, could be reconsidered in the next review cycle.

Here some feedback to address:

1. While the manuscript generally maintains an academic tone, several sections, particularly the abstract and introduction contain verbose or repetitive sentences. A language revision is needed to ensure clarity and brevity.
2. The manuscript lacks a meaningful comparison with existing baseline approaches. It is essential to evaluate the proposed MOPSO-KMeans method against conventional K-Means, single-objective PSO, or other clustering metaheuristics (e.g., Genetic Algorithms, Ant Colony Optimization, etc.) to contextualize the advantages of the proposed solution.
3. The experiment is conducted solely on the Iris dataset, which is small and well-structured. This limits the generalizability of the findings. The authors, if possible, should validate their method on multiple and more challenging datasets to support broader claims.
4. Visualizations such as scatter plots of cluster outputs (before and after optimization) would significantly enhance the readability and intuitive grasp of the method's performance.
5. The discussion focuses mainly on presenting numeric outcomes. A deeper exploration of why and how the method performs as it does, its potential weaknesses, and implications for future research is necessary.
6. The absence of detailed pseudocode or parameter settings limits reproducibility. Providing these elements is crucial for validation by other researchers.
7. While the title suggests a novel multi-objective formulation, the specific objectives and how they are balanced in the optimization process are not clearly defined. This aspect should be elaborated to clarify the true novelty of the work.