Vol. 9 No. 3 June 2025

e-ISSN: 2580-0760

JURNAL RESTI

REKAYASA SISTEM DAN TEKNOLOGI INFORMASI

jokoriyono171 0

IAII

JURNAL RESTI

Rekayasa Sistem dan Teknologi Informas

ISSN Media Elektro
2580-0760

www.jurnal.iaii.or.i

Home Current Issue Tutorial About Us Announcements Archives Dashboard 0 Search

Logout

Home / Editorial Team

Editor-in-Chief:

Yuhefizar

Politeknik Negeri Padang, Padang-Indonesia

Scopus | Google Scholar | Orcid

Managing Editor:

Ronal Watrianthos

Politeknik Negeri Padang, Padang-Indonesia

ResearcherID | Scopus | Google Scholar | Orcid

Rita Komalasari

Politeknik LP31, Bandung-Indonesia

ResearcherID | Scopus | Google Scholar | Orcid

Budi Sunaryo

Universitas Bung Hatta, Padang-Indonesia

ResearcherID | Scopus | Google Scholar | Orcid

Associates Editor:

Ikhwan Arief

.:MAIN MENU:.

About Journal

Editorial Team

Peer Review Process

Focus & Scope

Ambassador to Indonesia - Directory of Open Access Journals (DOAJ), Roskilde, **Denmark**

Research areas: Information Systems, Engineering, Manufacturing

Scopus | Orcid

Dini Oktarina Dwi Handayani

International Islamic University Malaysia, Kuala Lumpur-**Malaysia**

Research areas: Artificial Intelligence

Scopus | Orcid

Windu Gata

Universitas Nusa Mandiri, Jakarta-Indonesia

Research areas: IOT; Data Science

Scopus | Orcid

Shoffan Saifullah

AGH University of Krakow, **Poland**

Research areas: image processing; medical image analysis

Scopus | Orcid

Teddy Mantoro

Universitas Nusa Putra, Sukabumi-Indonesia

Research areas: Artificial Intelligence, Deep Learning, Information Security, Mobile Computing

Scopus | Orcid

Editorial Board Members:

Media Anugerah Ayu

Sampoerna University, Indonesia

Scopus | Orcid

Zairi Ismael Rizman

Universiti Teknologi MARA, Shah Alam, **Malaysia**

Scopus | Orcid

Diki Arisandi

Universitas Muhammadiyah Riau, Indonesia

Scopus | Orcid

Hendrick

Politeknik Negeri Padang, Indonesia

Author's Guide
Authors Fee
Publication Ethics
Open Access Statement
Plagiarism Policy
Archive Policy
Copyright and License
Author's Statement
F.A.Q
Download RESTI Templates
..::VISITOR STATISTICS::..
..::INFORMATION::..
For Reader
For Author
..::Find Us::..
Google Scholar
Garuda
SINTA
CrossRef
Dimensions

Click here for more details

RESTI (Rekayasa Sistem dan Teknologi Informasi) Journal indexed by:

Published by Professional Organization Ikatan Ahli Informatika Indonesia (IAII)/Indonesian Informatics Experts Association
Website Publisher : www.iaii.or.id - Website Journal : www.jurnal.iaii.or.id/index.php/RESTI - Email : jurnal.resti@gmail.com
Halaman Facebook : http://fb.me/jurnalRESTI

UNIVERSITAS TRISAKTI

J. Riyono <jokoriyono@trisakti.ac.id>

[RESTI] Editor Decision

1 message

chief <jurnal@iaii.or.id> Wed, Jun 4, 2025 at 12:48 PM
To: Aina Latifa Riyana Putri <ainaqp@telkomuniversity.ac.id>, Joko Riyono <jokoriyono@trisakti.ac.id>, Christina Eni Pujiastuti <christina.eni@trisakti.ac.id>

Aina Latifa Riyana Putri, Joko Riyono, Christina Eni Pujiastuti:

We have reached a decision regarding your submission to Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), "A Multi-Objective Particle Swarm Optimization Approach for Optimizing K-Means Clustering Centroids".

Our decision is to: Accept Submission with minor revision

Please follow : http://jurnal.iaii.or.id/index.php/reSTI/accepted for requirement publish.

chief
Ikatan Ahli Informatika Indonesia (IAII) Nusantara
jurnal@iaii.or.id

Reviewer A:
Recommendation: Accept Submission

Is the title appropriate and aligned with the journal's scope?

Yes

Does the abstract concisely inform readers and summarize the research?

Yes

Are relevant prior studies adequately contextualized, especially in the background?

Yes

Does the author provide in-depth analysis, particularly in the discussion and results sections?

Yes

Does the author demonstrate sufficient scientific reasoning, argumentation and interpretation?

Yes

Are descriptions and explanations presented clearly and understandably?

Yes

Are figures, tables, and numbers of adequate quality and easy to interpret?

Yes

Are visuals like diagrams and images used appropriately and clearly?

Yes

Does the manuscript present an original contribution?

Yes

Is the writing style suitably academic with clear and proper language?

Yes

Give constructive feedback:

1. Enhance the abstract by including more specific numerical results, particularly the most significant improvements (e.g., the dataset with the greatest accuracy gain or deviation reduction).
2. Ensure all figures and diagrams are included in the final manuscript submission. Several visual references (e.g., Figure 1A, 1B, etc.) are mentioned in the text,

but the actual images are missing from the document reviewed.
3. Add a brief discussion on computational cost or time complexity, especially since MOPSO-based methods may involve additional overhead compared to standard K-Means.
4. Consider expanding the conclusion to include reflections on potential real-world applications of the proposed method (e.g., biomedical clustering, IoT sensor grouping, or document clustering), which could increase the relevance and appeal of the study.
5. Optional: Include a comparison with other metaheuristic optimization algorithms (e.g., Genetic Algorithm, Differential Evolution) either in the discussion or related works section. Even a brief comparative remark based on the literature or prior studies would strengthen the contribution.

Reviewer B:
Recommendation: Accept Submission

Is the title appropriate and aligned with the journal's scope?

Yes

Does the abstract concisely inform readers and summarize the research?

Yes

Are relevant prior studies adequately contextualized, especially in the background?

Yes

Does the author provide in-depth analysis, particularly in the discussion and results sections?

Yes

Does the author demonstrate sufficient scientific reasoning, argumentation and interpretation?

Yes

Are descriptions and explanations presented clearly and understandably?

Yes

Are figures, tables, and numbers of adequate quality and easy to interpret?

Yes

Are visuals like diagrams and images used appropriately and clearly?

Yes

Does the manuscript present an original contribution?

Yes

Is the writing style suitably academic with clear and proper language?

Yes

Give constructive feedback:

The authors have fulfilled all required revisions; the article can now be accepted and processed for publication.

Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)

Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI

(Rekayasa Sistem dan Teknologi Informasi)

Vol. 9 No. 3 (2025) 542 - 550 e-ISSN: 2580-0760

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

Aina Latifa Riyana Putri[1*], Joko Riyono[2], Christina Eni Pujiastuti[3], Supriyadi[4]
[1]Data Science, Telkom University, Purwokerto, Indonesia
[2,3,4] Universitas Trisakti, Jakarta, Indonesia

[1]ainaqp@telkomuniversity.ac.id, [2]jokoriyono@trisakti.ac.id, [3]christina.eni@trisakti.ac.id, [4]supri@trisakti.ac.id

Abstract

The K-Means algorithm is a popular unsupervised learning method used 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+K-Means) 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+K-Means achieved a substantial accuracy improvement on the complex Target dataset, increasing from 0.26% to 59.2%. The TwoDiamonds dataset achieved the most desirable trade-off: it had the lowest SSW (1323.32), relatively high SSB (2863.34), and lowest standard deviation values, indicating compact clusters, good separation, and high consistency across runs. These findings highlight the potential of swarm-based optimization to achieve consistent and accurate clustering results on datasets with varying structural complexity.

Keywords: centroid; k-means; multiobjective particle swarm optimization; the sum of square within; the sum of square between

How to Cite: A. Latifa Riyana Putri, J. Riyono, and C. Eni Pujiastuti, “A Multi-Objective Particle Swarm Optimization Approach for Optimizing K-Means Clustering Centroids ”, *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 3, pp. 542 - 550, Jun. 2025.
*Permalink/*DOI: https://doi.org/10.29207/resti.v9i3.6533

Received: April 11, 2025
Accepted: June 4, 2025
Available Online: June 21, 2025

1. Introduction

Clustering is one of the data mining techniques aimed at grouping data into several clusters based on similarities in their characteristics. One of the most popular clustering methods is k-Means clustering [1], which is a non-hierarchical clustering algorithm that works by initializing centroids randomly. The objects are then grouped into k clusters based on their distances to the centroids, and the centroids' positions are iteratively updated until convergence is reached.

The advantages and disadvantages of this algorithm have been widely discussed in various studies. The k-Means algorithm is well-known for being efficient and scalable in processing large datasets [2]. Previous research, such as that conducted by [3] and [4], has shown that k-Means can produce more compact clusters compared to hierarchical clustering methods. However, k-Means has some limitations, particularly related to the random initialization of centroids, which can cause the clustering results to vary each time the algorithm is run [5]. Additionally, k-Means tends to get stuck in local optima, leading to suboptimal cluster assignments [6]. Its sensitivity to outliers is also a major concern, as extreme values can significantly shift the centroids' positions [7]. Furthermore, the assumption that clusters are spherical [8] and of uniform size makes k-Means less effective when dealing with datasets that have complex cluster shapes or varying densities.

To understand the quality of clustering results generated by k-Means, it is important to review the objectives and evaluation metrics. In k-Means clustering, the main goal is to form optimal clusters, where members of each cluster are highly like one another but significantly different from members of other clusters [9]. To achieve this, two primary metrics commonly used are the Sum of Squares Within-cluster (SSW) and the Sum of Squares Between-cluster (SSB). SSW measures the density of the cluster, indicating how tightly the data points within a cluster are grouped around the centroid [10]. Smaller SSW values indicate greater similarity between data points within the same cluster. On the other hand, SSB measures the distance between centroids, reflecting the separation between clusters [11]. Larger SSB values indicate greater distance between clusters, making the clustering more effective at distinguishing between different groups of data.

To address these issues, this study proposes an optimization approach based on Multi-Objective Particle Swarm Optimization (MOPSO) to determine more optimal centroids. MOPSO is a variant of Particle Swarm Optimization (PSO) [12] developed to solve problems with multiple objectives, making it suitable for clustering tasks that involve balancing two criteria simultaneously: minimizing the Sum of Squares Within-cluster (SSW) and maximizing the Sum of Squares Between-cluster (SSB), thereby producing clusters that are balanced in terms of both homogeneity and separation. The main advantage of MOPSO over other optimization algorithms lies in its convergence speed, efficient global exploration capabilities, and ease of implementation. Unlike evolutionary algorithms that use complex selection and mutation processes, MOPSO relies on a simple particle interaction mechanism in the search space. Additionally, MOPSO uses an external archive to store the best non-dominated solutions (Pareto optimal), facilitating decision-making based on trade-offs between cluster 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. Unlike previous studies that often use classic datasets such as Iris, this study specifically selects these five datasets because they present more complex challenges in the data clustering process. The Atom dataset challenges the algorithm in separating very close clusters, while Chainlink has a topological structure that is interconnected, which is difficult for centroid-based methods to handle. Engytime has an uneven density distribution, which may complicate the identification of proper cluster boundaries. The Target dataset presents a non-linear pattern that is hard for standard k-Means to capture, while Two Diamonds involves clusters that are very close together, making optimal separation difficult.

The performance evaluation was conducted by comparing the clustering results against the ground truth, which refers to the true labels that are known beforehand and used as a reference to assess the accuracy of the clustering performed by the algorithm. This comparison will be made using accuracy metrics and by testing the MOPSO-based approach against conventional clustering methods such as standard k-Means. With MOPSO-based optimization, this method is expected to be able to produce more separated and uniform clusters, thus outperforming data complexity. The findings of this study are anticipated to contribute significantly to the development of more optimal clustering methods for various applications in data mining and machine learning.

2. Methods

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

2.1 Experimental Testing and Simulation of the Proposed Muti-Objective Particle Swarm Optimization (PSO) Algorithm

As part of the empirical analysis, this study also evaluates the effectiveness of the proposed method using a set of benchmark datasets. These datasets are employed in the testing process to assess the capability of MOPSO in determining the optimal centroids for the K-Means algorithm, which is a critical step in enhancing clustering quality across various data scenarios. 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 Atom, Chainlink, Engytime, Target, and Two Diamonds.

The Atom dataset [14] is a commonly used benchmark for evaluating clustering algorithms under complex spatial conditions. It exists in a three-dimensional space ($\mathbb{R}^3$) and consists of two main clusters: a dense core cluster with 100 data points located at the center, and a larger, more dispersed outer hull cluster with 400 data points that geometrically encloses the core. This structure creates what is known as an overlapping convex hull, where the outer cluster fully surrounds the inner one. The significant difference in density and spatial arrangement poses challenges for traditional clustering algorithms, especially those that rely on distance measures such as K-Means.

The Chainlink dataset [15], [16] is a benchmark designed to evaluate the ability of clustering algorithms to handle complex and interrelated data structures. It consists of two clusters, each containing 300 data points, forming an interlocked chain-like structure in three-dimensional space ($\mathbb{R}^3$). Each cluster is shaped like a ring, and the two rings are intertwined, creating a configuration known as linear nonseparable entanglement. This refers to a condition where the clusters cannot be linearly separated due to their intertwined spatial arrangement. Although the clusters are globally distinct, many points from one cluster are locally closer to points from the other, which introduces a conflict between global separability and local proximity. Additionally, both clusters have nearly identical densities and inter-point distances, making it difficult to distinguish them based solely on size or distribution. This makes Chainlink particularly challenging for distance-based algorithms such as K-Means.

The EngyTime dataset is a benchmark used to evaluate the effectiveness of clustering algorithms in handling overlapping clusters with varying densities [17]. It contains 2,000 data points grouped into two clusters in a two-dimensional space ($\mathbb{R}^2$), based on two variables: “Engy” and “Time”. This dataset represents a simplified yet realistic density-based clustering problem, like those encountered in applications such as flow cytometry and sonar signal processing. EngyTime is generated from a mixture of two 2D Gaussian distributions, making it a suitable test case for evaluating how well clustering algorithms can distinguish overlapping groups. The clusters in this dataset are not separated by empty space, and they differ in density, which creates a significant challenge for traditional centroid-based methods like K-Means. These algorithms rely primarily on distance metrics and often ignore local density, which may lead to incorrect groupings when faced with overlapping, unevenly distributed data.

The Target dataset [18] is a benchmark designed to test the robustness of clustering algorithms when faced with overlapping clusters and the presence of outliers. This dataset exists in two-dimensional space ($\mathbb{R}^2$) and comprises 743 data points, divided into two main clusters and four small outlier groups. The first cluster is a dense spherical structure with 365 data points, while the second cluster forms an enclosing ring with 395 data points. This circular arrangement results in overlapping convex hulls, a geometric configuration that is particularly difficult to resolve using centroid-based algorithms like K-Means, which assume well-separated, linearly distinguishable clusters. The dataset also includes four corner-located outlier groups, each containing four data points. These outliers introduce additional complexity by potentially skewing the centroid calculations or being misclassified as independent clusters. The combination of dense core–ring overlap and peripheral noise makes the Target dataset a comprehensive benchmark for evaluating both the accuracy and stability of clustering methods.

The TwoDiamonds dataset [19], [20] is a benchmark commonly used to evaluate the ability of clustering algorithms to distinguish between weakly connected yet distinct cluster structures. It consists of 400 data points distributed evenly across two clusters, each shaped like a diamond, and located in a two-dimensional space ($\mathbb{R}^2$). The two clusters are positioned in adjacent square regions that almost touch at one side, forming a configuration that resembles two diamonds placed side by side. The primary challenge of this dataset lies in the weak connection between the two clusters. While the clusters are globally distinct, the narrow gap separating them can mislead clustering algorithms—particularly those based on local distance metrics like K-Means—into interpreting them as a single elongated cluster. Successfully separating the clusters in this dataset requires the algorithm to capture the overall geometric structure rather than relying purely on inter-point distances.

2.2 Standard K-Means Implementation as a baseline Comparison

As a baseline for performance comparison, the next step involves applying the standard K-Means clustering algorithm to each benchmark dataset. K-Means begins by randomly initializing centroids, followed by an iterative process of assigning data points to the nearest centroid based on Euclidean distance and updating the centroids until convergence. The number of clusters (k) is predetermined based on the ground truth of each dataset.

Due to the random nature of centroid initialization, K-Means may produce different clustering results in different runs. To address this, multiple independent runs are performed to assess consistency. The resulting cluster assignments are then evaluated using a confusion matrix, which enables the calculation of clustering accuracy by comparing the predicted clusters to the true class labels.

This baseline evaluation provides a reference point for comparing the clustering performance of the proposed MOPSO-KMeans method. By analyzing both standard K-Means and the optimized version under the same conditions and metrics, a more comprehensive assessment of the benefits and improvements introduced by the proposed approach can be achieved.

3. Results and Discussions

3.1 *Design and Workflow of the Proposed MOPSO Method*

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).

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

$$\mathrm{f}_1 = \min\left(\sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{x_i} \in \mathrm{C_j}} \left\|\mathrm{x_i} - \mu_\mathrm{j}\right\|^2\right) \tag{1}$$

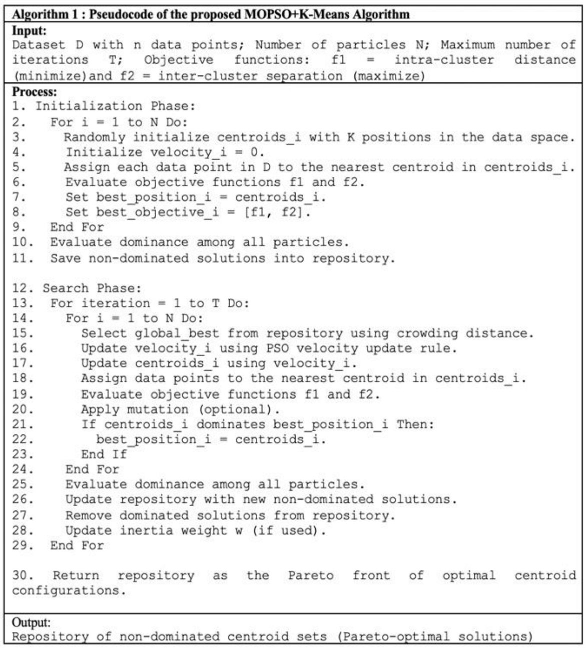
k is the number of clusters; x_i is the i-th data point; μ_j is the centroid of cluster C_j; $\left\|x_i - \mu_j\right\|^2$ is the squared Euclidean distance between the data point and the cluster centroid.

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

$$\mathrm{f}_2 = -\min\left(-\sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{n_j} \left\|\mu_\mathrm{j} - \mu\right\|^2\right) \tag{2}$$

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

The complete procedure of the proposed Multi-Objective Particle Swarm Optimization (PSO) algorithm can be summarized as follows, with its pseudocode presented on Figure 1.



Algorithm 1 : Pseudocode of the proposed MOPSO+K-Means Algorithm

Input:

```
Dataset D with n data points; Number of particles N; Maximum number of
iterations T; Objective functions: f1 = intra-cluster distance
(minimize)and f2 = inter-cluster separation (maximize)
```

Process:

```
Initialization Phase:
  For i = 1 to N Do:
    Randomly initialize centroids_i with K positions in the data space.
    Initialize velocity_i = 0.
    Assign each data point in D to the nearest centroid in centroids_i.
    Evaluate objective functions f1 and f2.
    Set best_position_i = centroids_i.
    Set best_objective_i = [f1, f2].
  End For
  Evaluate dominance among all particles.
  Save non-dominated solutions into repository.

Search Phase:
  For iteration = 1 to T Do:
    For i = 1 to N Do:
      Select global_best from repository using crowding distance.
      Update velocity_i using PSO velocity update rule.
      Update centroids_i using velocity_i.
      Assign data points to the nearest centroid in centroids_i.
      Evaluate objective functions f1 and f2.
      Apply mutation (optional).
      If centroids_i dominates best_position_i Then:
        best_position_i = centroids_i.
      End If
    End For
    Evaluate dominance among all particles.
    Update repository with new non-dominated solutions.
    Remove dominated solutions from repository.
    Update inertia weight w (if used).
  End For

Return repository as the Pareto front of optimal centroid
configurations.
```

Output:

```
Repository of non-dominated centroid sets (Pareto-optimal solutions)
```

Figure 1. Pseudocode of the Proposed MOPSO+K-Means Algorithm

The proposed MOPSO+K-Means algorithm integrates Multi-Objective Particle Swarm Optimization (MOPSO) with the traditional K-Means clustering method to overcome the limitations of random centroid initialization. As outlined in Figure 1, the process begins with the initialization phase, where each particle represents a potential solution in the form of a set of cluster centroids. Objective functions are defined to minimize intra-cluster distance (SSW) and maximize inter-cluster separation (SSB), enabling a balanced evaluation of cluster compactness and separation.

The fitness of each particle is assessed based on these two objectives, and the non-dominated solutions are stored in a Pareto-based repository. In the search phase, the particle positions are updated using both personal and global bests selected from the repository using the crowding distance. This iterative process continues until the stopping criteria are met, gradually refining the solutions toward the optimal trade-offs between the two objectives.

At the end of the MOPSO optimization, the repository contains a set of Pareto-optimal centroid configurations. From these, one or more candidate solutions can be selected to initialize the K-Means algorithm. This hybrid approach allows K-Means to begin with well-optimized centroid positions, potentially resulting in more stable and accurate clustering outcomes compared to traditional random initialization.

In addition to the Particle Swarm Optimization (PSO) algorithm, previous studies have also explored other metaheuristic approaches to improve clustering quality. Among the most widely used are Genetic Algorithm [21] (GA) and Differential Evolution [22] (DE). However, findings from several prior works suggest that PSO tends to offer advantages in terms of convergence speed and simplicity of parameters. PSO only requires a few parameters to be configured, such as inertia weight and learning coefficients. In contrast, GA and DE involve more complex parameter settings, including crossover rate, mutation rate, and selection strategies [23] which can significantly influence performance if not properly adjusted. Furthermore, PSO is known for its ability to maintain a good balance between exploration and exploitation during the search process [24], making it especially suitable for clustering problems with high complexity.

3.2 *Experimental Results and Analysis*

To evaluate the performance of the proposed Multi-Objective PSO, a series of experiments were conducted on a set of well-known benchmark datasets. These five datasets have been described in the Research Method.

The experimental setup in this study is defined as follows: the swarm size is set to N=40, and each benchmark dataset is tested independently 30 times, with each execution consisting of 100 iterations. All PSO-based algorithms are terminated upon reaching this maximum number of iterations. The performance of the proposed MOPSO+K-Means method is evaluated using standard clustering metrics, namely the best value, average value, and standard deviation of SSW, SSB, and accuracy. These metrics are used to assess the effectiveness and stability of the proposed method in comparison to standard K-Means across various benchmark datasets.

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 1. Clustering With MOPSO+K-Means

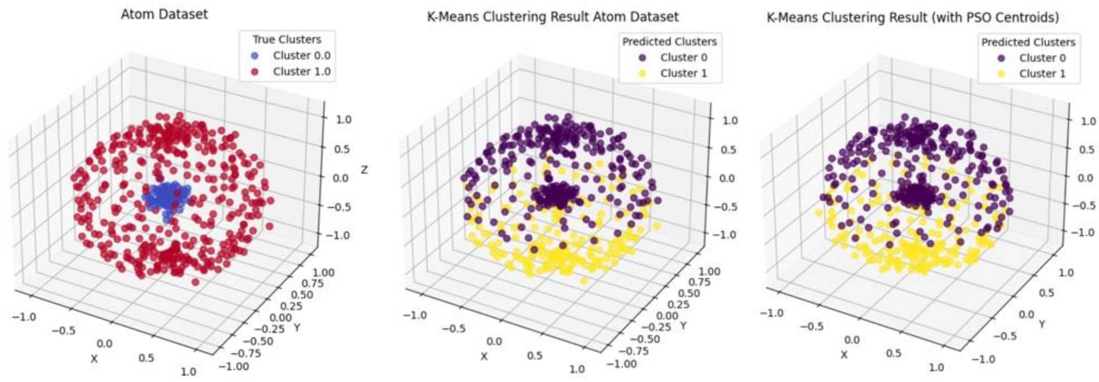
Dataset	Item	SSW	SSB
Atom	Avg.	1191.22	1414.52
	Std.	230.85	238.22
ChainLink	Avg.	1531.74	1711.26
	Std.	167.04	168.519
EngyTime	Avg.	49122.71	85124.33
	Std.	2951.153	3913.554
Target	Avg.	4126.614	7658.074
	Std.	892.147	1006.303
TwoDiamonds	Avg.	1323.322	2863.340
	Std.	42.822	44.191

Although the numerical results demonstrate that the proposed MOPSO+K-Means algorithm performs competitively across various datasets, a deeper analysis provides further insights into its behavior and performance dynamics. From a clustering quality perspective on Table 1, the ideal objective is to minimize the Sum of Squared Within-cluster distances (SSW) while maximizing the Sum of Squared Between-cluster distances (SSB). In this context, the EngyTime dataset exhibits the highest SSW (49122.71) and SSB (85124.33), indicating that the dataset likely has a large or widely spread structure. Despite this complexity, the algorithm successfully maintains strong inter-cluster separation. In contrast, the TwoDiamonds dataset achieves the most desirable trade-off: it has the lowest SSW (1323.32), relatively high SSB (2863.34), and the lowest standard deviation values, indicating compact clusters, good separation, and high consistency across runs.

For other datasets, such as Target and Atom, the SSW and SSB values fall in a mid-range category, but the relatively high standard deviations, particularly in Atom, highlight variability in the clustering results, suggesting sensitivity to initial conditions or swarm dynamics. Similarly, ChainLink yields results comparable to Atom but also shows considerable variability (Std SSW: 167.04, Std SSB: 168.52), likely reflecting the complexity or overlap in its cluster structure. These findings suggest that while MOPSO+K-Means can handle both simple and complex datasets, its stability may be challenged under certain data conditions.

Overall, TwoDiamonds emerges as the most stable dataset for this method, whereas EngyTime, despite strong average performance, reveals higher variability possibly due to noise or dispersed data points. These behavioral differences point to both strengths and limitations of the method. The ability of MOPSO+K-Means to find competitive clustering solutions is evident, but further enhancements could improve its robustness. Future research should explore adaptive parameter strategies, improved initialization techniques, or hybrid models to increase the method's reliability across diverse datasets. Additionally, expanding testing to high-dimensional or real-world datasets would help evaluate its scalability and broader applicability.

In this section, an analysis is conducted on the clustering outcomes derived from the implementation of the standard K-Means algorithm and the MOPSO-enhanced K-Means across a range of benchmark datasets. These results highlight the capabilities and limitations of both approaches when confronted with datasets exhibiting varying structural complexities.



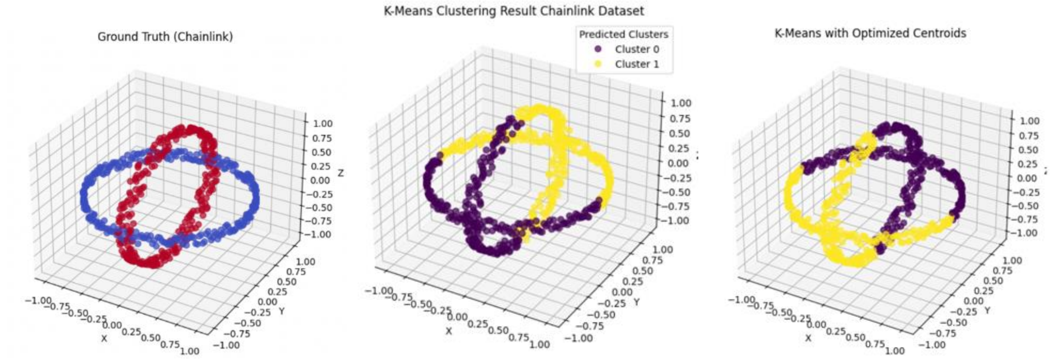
(a) (b) (c)

Figure 2. Clustering Result (a) Atom Dataset, (b) K-Means for Atom Dataset, (c) MOPSO+K-Means for Atom Dataset

The clustering results on the Atom dataset highlight the limitations of The Atom dataset illustrates the challenge of clustering data with a concentric structure, comprising a dense core surrounded by a shell. As depicted in Figure 2(a), the ground truth clearly reflects this core–shell configuration. However, the standard K-Means algorithm on Figure 2(b) produces a vertical partition, disregarding the radial nature of the data. This misalignment stems from K-Means assumption of spherical and convex clusters, which proves inadequate for capturing non-linear distributions. The integration of MOPSO+K-Means, as shown in Figure 2(c), results in a more accurate division that successfully distinguishes between the core and the surrounding shell. This demonstrates the ability of MOPSO to guide K-Means toward more structure-aware clustering outcomes in complex spatial configurations.

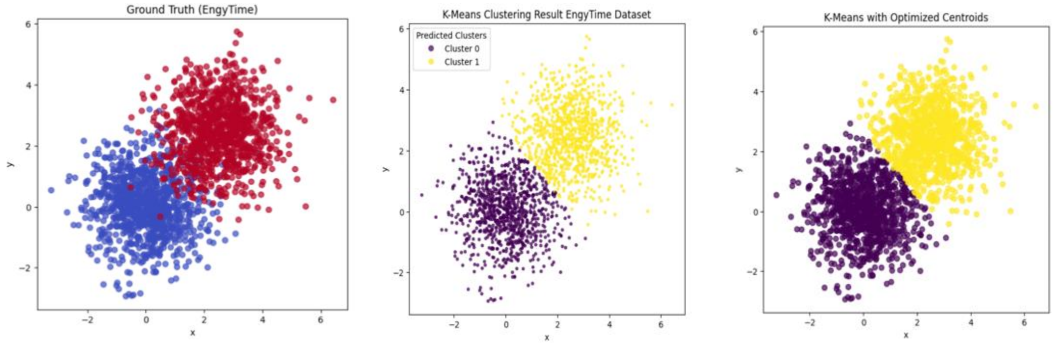
A similar pattern is observed in the ChainLink dataset, which features two intertwined, non-convex clusters on Figure 3(a). The standard K-Means algorithm on Figure 3(b) again fails to separate the data meaningfully, as it assigns points based on straight-line distances, ignoring the dataset's intricate shape. Conversely, the MOPSO+K-means on Figure 3(c) more effectively untangles the two chains, maintaining their topological distinction. This improved result underscores the role of MOPSO in adapting centroid placement to fit non-linear geometries that would otherwise confound traditional methods.

The EngyTime dataset, in contrast, presents a linearly separable structure, offering a more favorable scenario for K-Means. In Figure 4(a), the data exhibits two well-defined, adjacent clusters. The standard K-Means output on Figure 4(b) broadly captures the separation but misclassifies several points near the decision boundary. With optimized centroids on Figure 4(c), the clustering becomes cleaner, with reduced boundary ambiguity. This shows that even in simpler datasets, MOPSO+K-Means contributes by refining the clustering precision and minimizing convergence to suboptimal solutions.



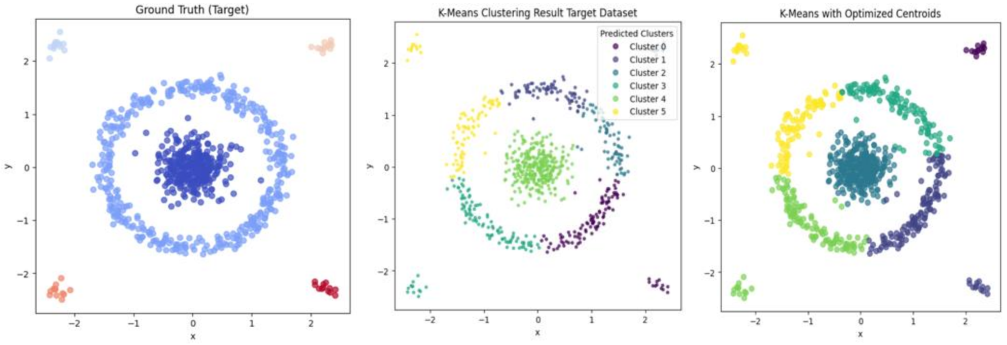
(a) (b) (c)

Figure 3. Clustering Result (a) Chainlink Dataset, (b) K-Means for Chainlink Dataset, (c) MOPSO+K-Means for Chainlink Dataset



(a) (b) (c)

Figure 4. Clustering Result (a) EngyTime Dataset, (b) K-Means for EngyTime Dataset, (c) MOPSO+K-Means for EngyTime Dataset



(a) (b) (c)

Figure 5. Clustering Results (a) Target Dataset, (b) K-Means for Target Dataset, (c) MOPSO+K-Means for Target Dataset

The analysis of the Target dataset provides further insight into the impact of complex geometries on clustering performance. This dataset contains concentric circular clusters and scattered peripheral groups on Figure 5(a). The clustering produced by the standard K-Means algorithm on Figure 5(b) results in fragmented groupings that poorly reflect the actual layout. Its tendency to impose spherical boundaries leads to significant structural mismatches. The MOPSO+K-Means on Figure 5(c) successfully aligns with the circular patterns and isolates the outlying groups more accurately, reinforcing the value of optimization in adapting to irregular spatial distributions.

Lastly, the TwoDiamonds dataset poses a moderate challenge due to its diamond-shaped, linearly separated clusters on Figure 6(a). While the standard K-Means on Figure 6(b) performs reasonably well, some inconsistencies are evident along the boundary, suggesting less-than-optimal centroid positioning. By contrast, the optimized version on Figure 6(c) yields a more symmetric and faithful clustering result, demonstrating how MOPSO can enhance K-Means even in datasets that are linearly separable but geometrically unconventional.

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.

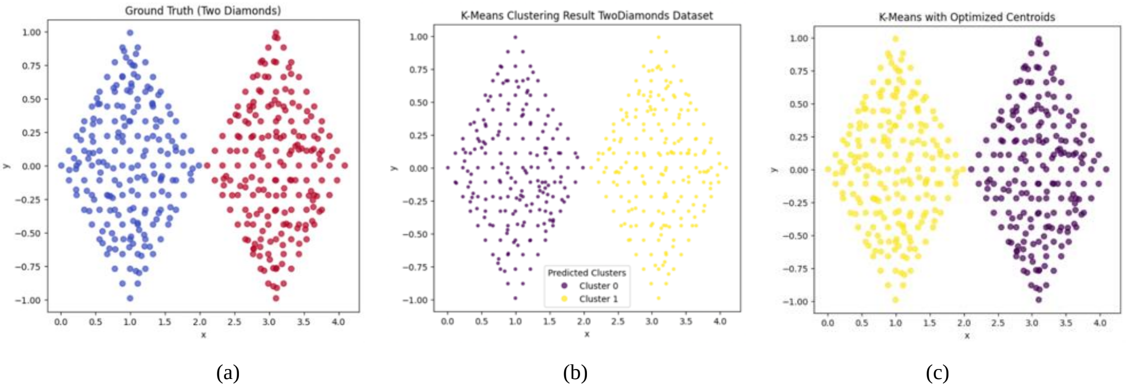


Figure 6. Clustering Result (a) TwoDiamonds Dataset, (b) K-Means Clustering Result TwoDiamonds Dataset, (c) MOPSO+K-Means Clustering Result TwoDiamonds Dataset

Table 2. Comparing Accuracy and Time Computational

Dataset	Accuracy K-means	Best Accuracy MOPSO+K-Means	Average Time Computational MOPSO+K-Means
Atom	54.4%	52.8%	5.1137 seconds
ChainLink	50%	50.2%	6.6889 seconds
EngyTime	95.6%	95.7%	7.617 seconds
Target	0.2692%	59.2%	8.539 seconds
TwoDiamonds	100%	100%	0.844 seconds

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 improvement was observed on the Target dataset. K Means achieved only 0.26% 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-K-Means achieved perfect accuracy (100%), indicating that this dataset has a very clear structure that can be easily separated by both algorithms.

The Average Time Computational *MOPSO+K-Means* column in Table 2 presents the average computational time required by the algorithm to complete one clustering process for each dataset. This time is measured from the beginning of the centroid optimization using the MOPSO algorithm to the final clustering result produced by K-Means. The values represent the average time taken across 30 independent trials. They indicate the efficiency of the algorithm in solving clustering tasks, which is influenced by the complexity of the data patterns and the algorithm's ability to converge toward optimal solutions.

For example, the TwoDiamonds dataset demonstrates the lowest average computational time (0.844 seconds), which can be attributed to its well-separated and clearly structured clusters. This allows MOPSO to converge quickly, requiring minimal exploration. In contrast, the Target dataset shows the highest computational time (8.539 seconds), suggesting that the algorithm needed significantly more effort to explore the solution space due to the highly irregular and overlapping nature of the data clusters. Similarly, EngyTime and Chainlink also required more computational time, which reflects the greater complexity in their data distributions and the increased difficulty in identifying distinct clusters. Meanwhile, Atom shows a moderate computational time, indicating that while the data is not entirely straightforward, it is still manageable for the algorithm to optimize efficiently.

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.

However, one key limitation of the current study is the assumption that the number of clusters (k) is known beforehand. Although this facilitates benchmarking against ground truth labels, it does not reflect the realities of unsupervised learning tasks, where k must be inferred from the data. In practice, k is a hyperparameter that must be estimated carefully. Common strategies include the elbow method, which identifies diminishing returns in intra-cluster variance as k increases, the silhouette score, which quantifies cluster cohesion and separation, and the gap statistic, which compares clustering performance against that of random reference distributions.

To overcome this limitation, future work could explore the extension of MOPSO to simultaneously optimize both the number of clusters and the centroid positions. This could be framed as a multi-objective optimization problem—balancing cluster compactness, separation, and model complexity—or as a constrained optimization task where *k* is bounded within a reasonable range. Such an approach would enhance the applicability of the method in real-world scenarios, allowing it to autonomously discover both the optimal clustering structure and its corresponding parameters without relying on prior knowledge.

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 can produce 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. Such characteristics are often found in real-world applications, including biomedical data analysis where patient subgroups may form irregular patterns, sensor grouping in IoT environments with overlapping signal zones, and document clustering tasks where semantic relationships are not linearly separable. These application domains can benefit from algorithms that offer both structural flexibility and solution stability, as demonstrated by MOPSO+K-Means in this study.

Acknowledgements

This work was supported by Telkom University.

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] 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.

[3] A. Abdulhafedh, “Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation,” *Journal of City and Development*, vol. 3, no. 1, 2021.

[4] P. Patel, B. Sivaiah, and R. Patel, “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 2022*, 2022. doi: 10.1109/ICICCSP53532.2022.9862439.

[5] 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.

[6] A. Qtaish, M. Braik, D. Albashish, M. T. Alshammari, A. Alreshidi, and E. J. Alreshidi, “Optimization of K-means clustering method using hybrid capuchin search algorithm,” *Journal of Supercomputing*, vol. 80, no. 2, 2024, doi: 10.1007/s11227-023-05540-5.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[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.

[21] B. Khusul Khotimah, F. Irhamni, and T. Sundarwati, "A GENETIC ALGORITHM FOR OPTIMIZED INITIAL CENTERS K-MEANS CLUSTERING IN SMEs," *J Theor Appl Inf Technol*, vol. 15, no. 1, 2016, Accessed: Jun. 13, 2025. [Online]. Available: www.jatit.org

[22] H. He, B. Sun, Y. Yang, and S. Liu, "An improved K-Means clustering based on differential evolution," *J Phys Conf Ser*, vol. 2595, no. 1, p. 012010, Sep. 2023, doi: 10.1088/1742-6596/2595/1/012010.

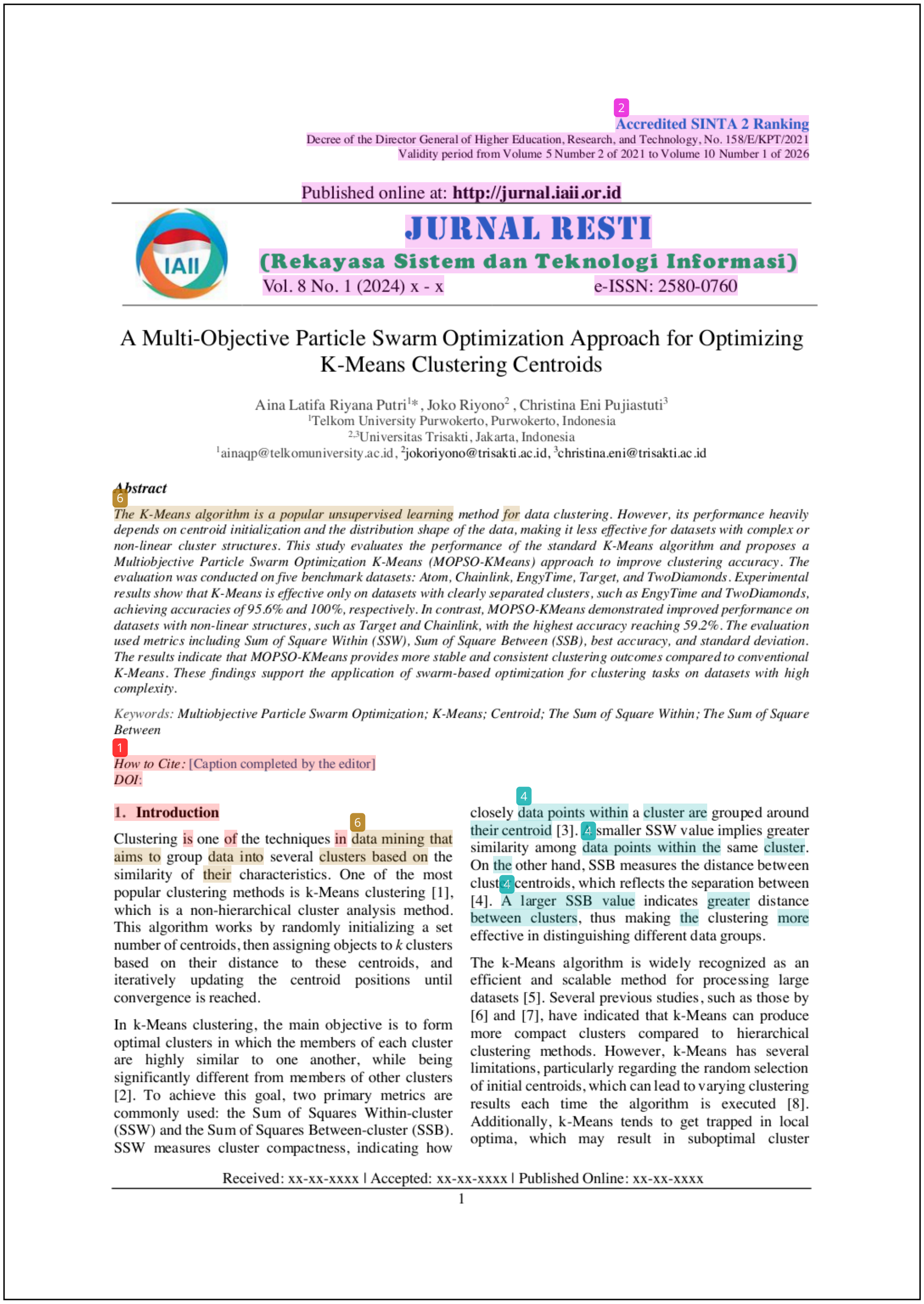
[23] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri, and V. B. S. Prasath, "Choosing Mutation and Crossover Ratios for Genetic Algorithms—A Review with a New Dynamic Approach," *Information 2019, Vol. 10, Page 390*, vol. 10, no. 12, p. 390, Dec. 2019, doi: 10.3390/INFO10120390.

[24] M. Jain, V. Saihjpal, N. Singh, and S. B. Singh, "An Overview of Variants and Advancements of PSO Algorithm," *Applied Sciences 2022, Vol. 12, Page 8392*, vol. 12, no. 17, p. 8392, Aug. 2022, doi: 10.3390/APP12178392.

naskah resti

by aina putri

Submission date: 10-Apr-2025 07:58AM (UTC+0700)
Submission ID: 2409637607
File name: Naskah_Resti2024.docx (2.74M)
Word count: 4336
Character count: 25638



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

Aina Latifa Riyana Putri[1]*, Joko Riyono[2], Christina Eni Pujiastuti[3]
[1]Telkom University Purwokerto, Purwokerto, Indonesia
[2,3]Universitas Trisakti, Jakarta, Indonesia
[1]ainaqp@telkomuniversity.ac.id, [2]jokoriyono@trisakti.ac.id, [3]christina.eni@trisakti.ac.id

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 exists in a three-dimensional space (ℝ³) 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.

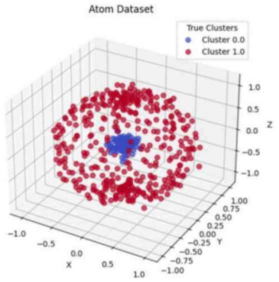
   

   Figure 1. Atom Dataset

   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 consists of two clusters, each containing 300 data points, which together form a structure resembling interlinked chains in three-dimensional space (ℝ³).

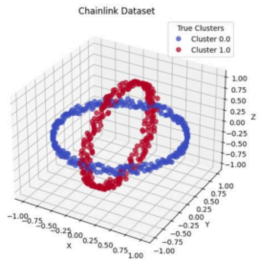


Figure 2. ChainLink Dataset

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”.

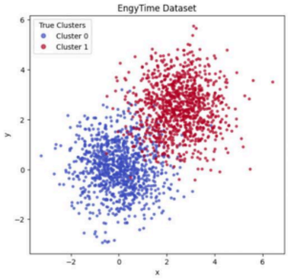


Figure 3. EngyTime Dataset

This dataset 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 [18] is a benchmark dataset designed to evaluate the robustness of clustering algorithms in handling overlapping clusters and the presence of outliers. 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.

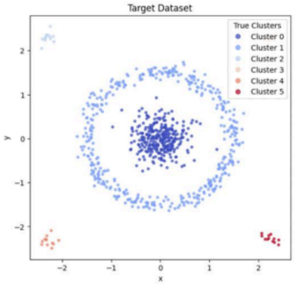


Figure 4. Target Dataset

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 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 consists of two clusters, each containing 200 data points in a two-dimensional space ($\mathbb{R}^2$).

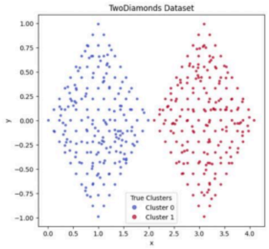


Figure 5. TwoDiamonds Dataset

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, described as follows:

1. True Positive (TP): The number of data points that are actually positive and are correctly predicted as positive.
2. True Negative (TN): The number of data points that are actually negative and are correctly predicted as negative.
3. False Positive (FP): The number of data points that are actually negative but are incorrectly predicted as positive.
4. False Negative (FN): The number of data points that are actually positive but are incorrectly predicted as negative.

Accuracy represents how accurately the model classifies data. In other words, it measures how close the predicted values are to the actual values. Accuracy is calculated as:

$$\text{Accuracy} = \frac{(\boldsymbol{TP+TN})}{(\boldsymbol{TP+FP+FN+TN})} \quad (1)$$

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):

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

   Where k is the number of clusters; $\boldsymbol{x_i}$ is the i-th data point; $\boldsymbol{\mu_j}$ is the centroid of cluster $\boldsymbol{C_j}$; $\|\boldsymbol{x_i} - \boldsymbol{\mu_j}\|^2$ 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:

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

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

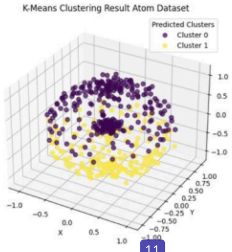


Figure 6. Atom Dataset With K-Means Clustering

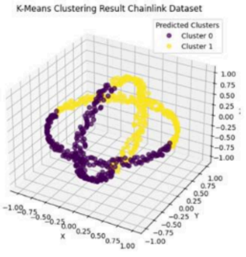


Figure 7. ChainLink Dataset With K-Means Clustering

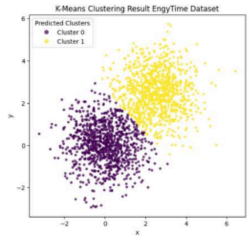


Figure 8. EngyTime Dataset With K-Means Clustering

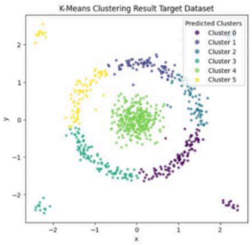


Figure 9. Target Dataset With K-Means Clustering

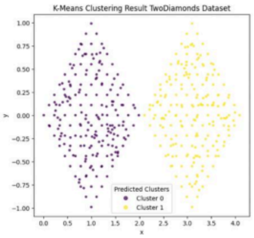


Figure 10. TwoDiamonds Dataset With K-Means Clustering

The clustering visualization results on the Atom dataset indicate that the K-Means algorithm was not able to group the data effectively. In Figure 1.A, 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 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 1.C 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, 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, 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 the results presented, 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 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	

The table above 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 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-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.

References

[1] Ghazal, T. M. (2021). Performances of k-means clustering algorithm with different distance metrics. *Intelligent Automation & Soft Computing*, *30*(2), 735-742.

[2] Gustriansyah, R., Suhandi, N., & Antony, F. (2020). Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 470-477.

[3] Richard, R., Cao, H., & Wachowicz, M. (2021, January). An Automated Clustering Process for Helping Practitioners to Identify Similar EV Charging Patterns across Multiple Temporal Granularities. In *SMARTGREENS* (pp. 67-77).

[4] Guerreiro, M. T., Guerreiro, E. M. A., Barchi, T. M., Biluca, J., Alves, T. A., de Souza Tadano, Y., ... & Siqueira, H. V. (2021). Anomaly detection in automotive industry using clustering methods—a case study. *Applied Sciences*, *11*(21), 9868.

[5] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.

[6] Kardi dalam Susanto, A.R., 2013, *Sistem Pendukung Keputusan Pengadaan Buku Perpustakaan STIKOM Surabaya Menggunakan Metode K-Means Clustering*. Makalah TA. Surabaya, STIKOM Surabaya.

[7] Helmiah, 2013, *Sisitem Pendukung Keputusan untuk Pengkatogorian IPK dan Llama Studi Alumni Menggunakan Metode K-Means*. *Skripsi*, Yogyakarta: UII Teknik Informatika

[8] Vouros, A., Langdell, S., Croucher, M., & Vasilaki, E. (2021). An empirical comparison between stochastic and deterministic centroid initialisation for K-means variations. *Machine Learning*, *110*, 1975-2003.

[9] Mussabayev, R., & Mussabayev, R. (2023). Comparative Analysis of Optimization Strategies for K-means Clustering in Big Data Contexts: A Review. *arXiv preprint arXiv:2310.09819*.

[10] Zhang, Z., Feng, Q., Huang, J., Guo, Y., Xu, J., & Wang, J. (2021). A local search algorithm for k-means with outliers. *Neurocomputing*, *450*, 230-241.

[11] Wani, A. A. (2024). Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions. *PeerJ Computer Science*, *10*, e2286.

[12] Jain, M., Saihjpal, V., Singh, N., & Singh, S. B. (2022). An overview of variants and advancements of PSO algorithm. *Applied Sciences*, *12*(17), 8392.

[13] Thrun, M. C., & Ultsch, A. (2020). Clustering benchmark datasets exploiting the fundamental clustering problems. *Data in brief*, *30*, 105501.

[14] Ultsch, A. (2004). Strategies for an artificial life system to cluster high dimensional data. *Abstracting and Synthesizing the Principles of Living Systems, GWAL-6*, 128-137.

[15] Ultsch, A., Guimaraes, G., Korus, D., & Li, H. (1994). Knowledge extraction from artificial neural networks and applications. In *Parallele Datenverarbeitung mit dem Transputer: 5. Transputer-Anwender-Treffen TAT'93, Aachen, 20.–22. September 1993* (pp. 148-162). Springer Berlin Heidelberg.

[16] Ultsch, A. (1995). Self organizing neural networks perform different from statistical k-means clustering. *Proc. GfKl, Basel, Swiss*.

[17] Baggenstoss, P. M. (2002). Statistical modeling using gaussian mixtures and hmms with matlab. *Naval Undersea Warfare Center, Newport RI*.

[18] Ultsch, A. (2005, October). U* C: Self-organized Clustering with Emergent Feature Maps. In *LWA* (pp. 240-244).

[19] Ultsch, A. (2003). U*-matrix: a tool to visualize clusters in high dimensional data.

[20] Ultsch, A. (2003). *Optimal density estimation in data containing clusters of unknown structure*. Univ.

naskah resti

ORIGINALITY REPORT

13%	9%	7%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Muslim Indonesia Student Paper	4%
2	jurnal.iaii.or.id Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	Rozzi Kesuma Dinata, Bustami, Sujacka Retno. "Optimizing the Evaluation of K-means Clustering Using the Weight Product", Revue d'Intelligence Artificielle, 2024 Publication	1%
5	www.uni-marburg.de Internet Source	1%
6	ph02.tci-thaijo.org Internet Source	1%
7	Submitted to University of Florida Student Paper	1%
8	"Advances in Neural Networks – ISNN 2009", Springer Science and Business Media LLC, 2009 Publication	1%
9	doi.org Internet Source	1%
10	Fakhrian Fadlia Adiwijaya, Arisza Zufar Fathurrahman, Chrismikha Hardyanto. "Application of Agglomerative Hierarchical	1%

Clustering (AHC) for Grouping Prospective Scholarship Recipients", 2023 International Conference on Informatics Engineering, Science & Technology (INCITEST), 2023
Publication

11	Submitted to University of Strathclyde Student Paper	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		

jokoriyono171 0

JURNAL RESTI
Rekayasa Sistem dan Teknologi Informas
ISSN Media Elektro
2580-0760
www.jurnal.iaii.or.i

Home Current Issue Tutorial About Us Announcements Archives Dashboard 0 Search

Logout

Home / Archives / Vol 9 No 3 (2025): June 2025

This issue features 28 scholarly articles from 32 distinguished institutions in Indonesia, two from Malaysia, and one from Australia. The editors thank all the researchers who chose the RESTI Journal as a platform to spread the fruits of their labor. Your contributions enrich not only our publication but also the broader scientific community, particularly in the ever-evolving field of information technology.

The cover, foreword, and table of contents can be downloaded here.

Indexed by Scopus

ELSEVIER

1.8 2024 CiteScore

33rd percentile

Powered by Scopus

sinta S2 Science and Technology Index

.:MAIN MENU:.

About Journal

Editorial Team

Peer Review Process

Focus & Scope

DOI: https://doi.org/10.29207/resti.v9i3

Published: 2025-05-24

Computer Science Applications

Author's Guide
Authors Fee
Publication Ethics
Open Access Statement
Plagiarism Policy
Archive Policy
Copyright and License
Author's Statement
F.A.Q
Download RESTI Templates
..::VISITOR STATISTICS::..
..::INFORMATION::..
For Reader
For Author
..::Find Us::..
Google Scholar
Garuda
SINTA
CrossRef
Dimensions
..::SUPERVISED BY::..

Artificial Intelligence

Computer Vision and Pattern Recognition

RESTI (Rekayasa Sistem dan Teknologi Informasi) Journal indexed by:

Published by Professional Organization Ikatan Ahli Informatika Indonesia (IAII)/Indonesian Informatics Experts Association
Website Publisher : www.iaii.or.id - Website Journal : www.jurnal.iaii.or.id/index.php/RESTI - Email : jurnal.resti@gmail.com
Halaman Facebook : http://fb.me/jurnalRESTI