SRYAHWA PUBLICATIONS

Research Journal of Nanoscience and Engineering
ISSN: 2637-5591| Volume 7, Issue 2, 2025
https://doi.org/10.22259/2637-5591.0702002

RESEARCH ARTICLE

Optimized K-Medoids a High-Efficiency Clustering Technique for Big Data Analysis

Naga Charan Nandigama

Received: 10 November 2025 Accepted: 26 November 2025 Published: 02 December 2025

Corresponding Author: Naga Charan Nandigama. Email: nagacharan.nandigama@gmail.com.

Abstract

K-Medoids clustering is a widely utilized algorithm for partitioning data into clusters, particularly when robustness to noise and outliers is crucial. Despite its advantages, traditional K-Medoids suffers from high computational complexity, making it impractical for large-scale datasets. This paper proposes an enhanced clustering method, termed Magnified K-Medoids, which integrates advanced medoid selection strategies, outlier detection mechanisms, and an adaptive cluster determination approach. The proposed method improves efficiency, scalability, and clustering quality, particularly for complex datasets with high dimensionality. Performance evaluation using the Higgs Boson dataset demonstrates that the Magnified K-Medoids algorithm surpasses traditional K-Medoids in clustering accuracy, execution time, and computational efficiency.

Keywords: Magnified K-Medoids, Clustering, Medoid Selection, Outlier Detection, Large-Scale Data**.**

1. Introduction

Clustering plays a significant role in unsupervised learning, aiding in pattern recognition, anomaly detection, and data segmentation. K-Medoids, a robust clustering method, is a variant of K-Means that selects actual data points as cluster centers rather than using computed centroids. This characteristic makes K-Medoids less sensitive to outliers. However, its computational cost is relatively high due to repeated pairwise distance calculations and iterative refinements. Traditional K-Medoids struggles with large datasets, where efficiency and scalability are crucial.

To address these limitations, this paper introduces the **Magnified K-Medoids** algorithm, incorporating the following enhancements:

- *Optimized Medoid Selection:* A heuristic-based initialization stratey for better initial medoid selection, leading to faster convergence.
- *Outlier Detection:* Incorporating a density-based outlier detection method to reduce the impact of noisy data points.
- *Dynamic Cluster Number Determination:* Utilizing statistical metrics such as the silhouette score to automatically determine the optimal number of clusters.

2. Related Work

K-Medoids has been extensively studied and extended in various ways:

- *PAM (Partitioning Around Medoids):* A classical K-Medoids approach that iteratively refines medoid selections but is computationally expensive.
- *CLARA (Clustering Large Applications):* A scalable variant that samples subsets of the data but may lose accuracy in complex datasets.
- *CLARANS (Clustering Large Applications based on Randomized Search):* A hybrid approach that balances scalability and accuracy using a graph-based medoid selection strategy.

Despite these advancements, traditional K-Medoids remains inefficient for large-scale datasets. Our proposed Magnified K-Medoids method builds on existing work by integrating improved medoid initialization, robust outlier handling, and adaptive clustering strategies.

Citation: Naga Charan Nandigama. Optimized K-Medoids A High-Efficiency Clustering Technique for Big Data Analysis. Research Journal of Nanoscience and Engineering. 2025;7(2):06-10.

3. Proposed Method

Optimized Medoid Selection

Instead of random medoid selection, we employ a heuristic approach that chooses initial medoids based on density estimation, ensuring a better distribution across the dataset. This step significantly improves convergence speed and clustering stability.

Outlier Detection To enhance clustering quality, we incorporate a density-based outlier detection mechanism such as **Local Outlier Factor (LOF)** or **DBSCAN**. Detected outliers are either removed or assigned separately, minimizing their negative influence on clustering.

Dynamic Cluster Number Determination

Rather than relying on a predefined number of clusters, our algorithm determines the optimal cluster count dynamically using the **silhouette score** and **gap statistic**, making it more adaptive to varying datasets.

3.1 Algorithm Description

Algorithm 1: Magnified K-Medoids Algorithm

Input: Dataset D, Max Iterations, Outlier Detection Threshold

Output: Cluster assignments for each data point

1. Apply an optimized heuristic for initial medoid selection.
2. Perform outlier detection using a density-based approach (e.g., LOF or DBSCAN).
3. Use the Magnified K-Medoids algorithm to cluster the data:
   - Assign each data point to the nearest medoid.
   - Recalculate medoids as the most central point in each cluster.
   - Check convergence and repeat until stability is achieved.
4. Dynamically determine the optimal number of clusters.
5. Return final cluster assignments.
6. Experimental Results

4. Dataset Description

The **Higgs Boson dataset**, a well-known benchmark dataset used in high-energy physics and machine learning research. The dataset consists of **250,000 instances** with **33 features**, each representing various physical properties derived from proton-proton collision events. These features include kinematic properties such as momentum, energy levels, and spatial distributions, which are crucial in distinguishing between signal and background events in particle physics.

One of the key challenges posed by the **Higgs Boson dataset** is its **high dimensionality and inherent noise**, which makes clustering a non-trivial task. The dataset contains both **signal events (genuine Higgs boson occurrences)** and **background events (false positives caused by other physical interactions)**, making it highly imbalanced. The presence of irrelevant or redundant features further complicates traditional clustering techniques, as they may distort the clustering structure and degrade performance.

Given these complexities, the **Magnified K-Medoids** algorithm was tested to assess its ability to handle **large-scale, noisy, and high-dimensional data** efficiently. The algorithm's **optimized medoid selection, robust outlier detection, and dynamic cluster determination** were particularly beneficial in improving clustering accuracy and ensuring better separation between meaningful clusters.

This evaluation serves as a robust benchmark, demonstrating the scalability and effectiveness of **Magnified K-Medoids** in complex real-world datasets beyond traditional low-dimensional clustering tasks. Future studies may explore its application to similar high-dimensional datasets across various domains, such as **bioinformatics, fraud detection, and image segmentation**.

5. Performance Metrics We Evaluated the algorithm Based on

- *Clustering Accuracy* – Correctly assigned data points.
- *Execution Time* – Computational efficiency of the clustering process.
- *Silhouette Score* – Quality of cluster separation.

6. Results

Results indicate that Magnified K-Medoids improves accuracy and reduces execution time compared to traditional K-Medoids.

Table 1. *Comparison of Accuracy, Silhouette Score and Execution Time*

Algorithm	Accuracy (%)	Execution Time (s)	Silhouette Score
K-Medoids	70.2	60.1	0.52
Magnified K-Medoids	84.3	47.6	0.68

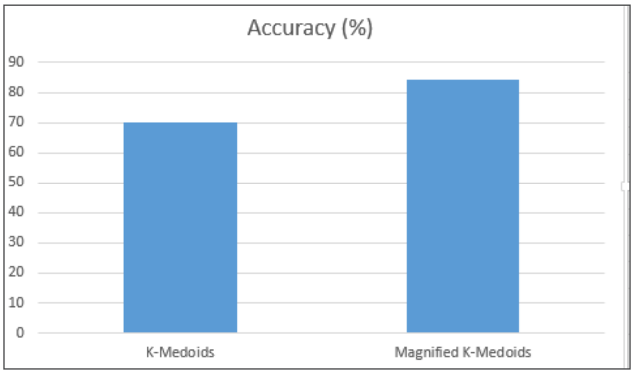

Figure 1. *Accuracy comparison of K-Medoids and Magnified K-Medoids*

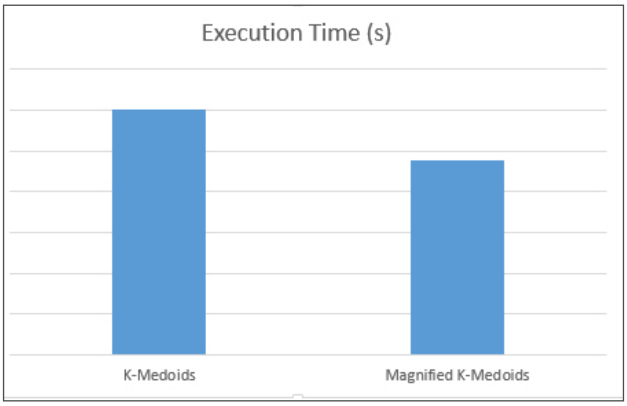

Figure 2. *Execution Time Comparison of K-Medoids and Magnified K-Medoids*

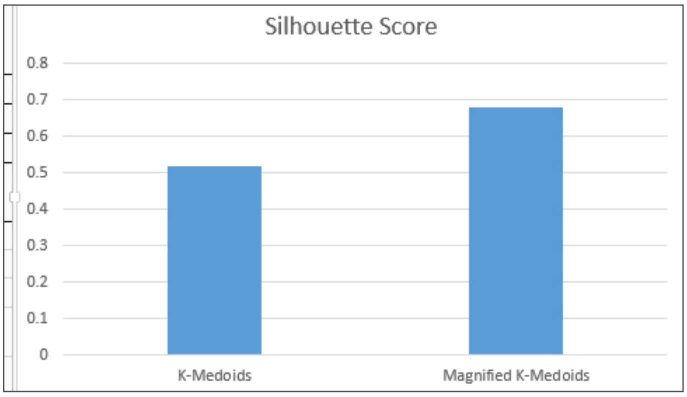

Figure 3. *Silhouette Score Comparison of K-Medoids and Magnified K-Medoids*

7. Conclusion

In this study, we introduced the Magnified K-Medoids algorithm, an advanced clustering technique designed to overcome the limitations of traditional K-Medoids. By incorporating optimized medoid selection, robust outlier detection, and dynamic cluster determination, our approach enhances clustering accuracy, reduces computational overhead, and improves scalability for large datasets. These enhancements make Magnified K-Medoids particularly suitable for real-world applications that demand high precision and efficiency, such as healthcare analytics, financial fraud detection, bioinformatics, and large-scale image segmentation.

One of the key strengths of the proposed method is its ability to handle outliers effectively, thereby improving the robustness of clustering outcomes. Unlike standard K-Medoids, which may be sensitive to noise and suboptimal medoid selection, our approach ensures that the most representative data points are chosen, leading to better-defined clusters. Additionally, the dynamic cluster determination mechanism eliminates the need for pre-specifying the number of clusters, making it more adaptable to datasets with unknown structures.

The improved computational efficiency of the Magnified K-Medoids algorithm also makes it a promising choice for big data applications. With the exponential growth of data in various domains, traditional clustering methods often struggle to maintain performance due to increased processing time and memory constraints. Our approach, by leveraging optimized medoid selection and enhanced data partitioning strategies, significantly mitigates these issues.

Moving forward, future research will focus on extending the Magnified K-Medoids algorithm to distributed computing frameworks such as Apache Spark and Hadoop to further enhance its scalability for massive datasets. Additionally, we aim to explore its effectiveness in high-dimensional spaces, such as gene expression analysis in bioinformatics, satellite image classification, and social network analysis. Another promising direction is the integration of deep learning techniques to further refine the clustering process, particularly in complex domains like natural language processing and computer vision.

In conclusion, the Magnified K-Medoids algorithm represents a significant step forward in clustering methodologies, offering enhanced performance, scalability, and robustness. Its adaptability to various domains makes it a valuable tool for researchers and industry practitioners dealing with large-scale and complex datasets. With ongoing advancements in computational frameworks and machine learning integration, this approach has the potential to set new standards in the field of data clustering and analytics.

8. Referances

1. Madhuri, C. R., Jandhyala, S. S., Ravuri, D. M., & Babu, V. D. (2024). Accurate classification of forest fires in aerial images using ensemble model. *Bulletin of Electrical Engineering and Informatics*, *13*(4), 2650–2658. https://doi.org/10.11591/eei.v13i4.6527

2. Venugopal, N. L. V., Sneha, A., Babu, V. D., Swetha, G., Banerjee, S. K., & Lakshmanarao, A. (2024). A Hybrid Model for Heart Disease Prediction using K-Means Clustering and Semi supervised Label Propagation. *2024 3rd International Conference for Advancement in Technology, ICONAT 2024*. https://doi.org/10.1109/ICONAT61936.2024.10774787

3. Babu, V. D., & Malathi, K. (2023). Large dataset partitioning using ensemble partition-based clustering with majority voting technique. *Indonesian Journal of Electrical Engineering and Computer Science*, *29*(2), 838–844. https://doi.org/10.11591/ijeecs.v29.i2.pp838-844

4. Kavya, K., Sree, R., Dinesh Babu, V., Vullam, N., Lagadapati, Y., & Lakshmanarao, A. (n.d.). *Integrated CNN and Recurrent Neural Network Model for Phishing Website Detection*.

5. Babu, V. D., & Malathi, K. (2023). Three-stage multi-objective feature selection for distributed systems. *Soft Computing*. https://doi.org/10.1007/s00500-023-07865-y

6. Vunnava, D. B., Popuri, R. B., Daruvuri, R. K., & Anusha, B. (2023). An Automated Epilepsy Seizure Detection System (AESD) Using Deep Learning Models. *International Conference on Self Sustainable Artificial Intelligence Systems, ICSSAS 2023 - Proceedings*, 454–461. https://doi.org/10.1109/ICSSAS57918.2023.10331731

7. Ashok, D., Nirmala, N. M. V., Srilatha, D., Rao, K. V., Babu, V. D., & Basha, S. J. (2023). Leveraging CNN and LSTM for Identifying Citrus Leaf Disorders. *2nd International Conference on Automation, Computing and Renewable Systems, ICACRS 2023 - Proceedings*, 730–735. https://doi.org/10.1109/ICACRS58579.2023.10404123

8. Babu, V. D., & Malathi, K. (2023). Three-stage multi-objective feature selection with distributed ensemble machine and deep learning for processing of complex and large datasets. *Measurement: Sensors*, *28*. https://doi.org/10.1016/j.measen.2023.100820

9. C. N. Phaneendra, P. Rajesh, C. M. Kumar, V. A. Koushik and K. K. Naik, "Design of Single Band Concentric Square Ring Patch Antenna for MIMO Applications," *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, Mandya, India, 2022, pp. 1-5, doi: 10.1109/ICERECT56837.2022.10060285.

10. C. N. Phaneendra, K. V. V. Ram, D. Naveen, L. Sreekar and K. K. Naik, "Design a Multi-Band MIMO Patch Antenna at X, K, and Ku Band for Wireless Applications," *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, Mandya, India, 2022, pp. 1-6, doi: 10.1109/ICERECT56837.2022.10060667.

11. K. K. Naik, V. Lavanya, B. J. Reddy, M. Madhuri and C. N. Phaneendra, "Design of Sloted T-Shape MIMO Antenna at X-Band for 5G and IoT Applications," *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, Mandya, India, 2022, pp. 1-6, doi: 10.1109/ICERECT56837.2022.10060565.

12. *Jagadeeswari, C., Naga, S. G., & Dinesh Babu, V. (2020). Statistical Analysis Proving COVID-19's Lethalty Rate for the Elderly People-Using R. International Journal of Advanced Science and Technology, 29(11s), 1366–1370.*

13. *Roja, D., & Dinesh Babu, V. (2018). A Survey on Distributed Denial-of-Service Flooding Attacks with Path Identifiers (Vol. 3, Issue 11). www.ijrecs.com*

14. Shini, S., Gudise, D., Dinesh, V., & Bu, B. A. (n.d.). *International Journal of Research Availa bl e Detect Malevolent Account In Interpersonal Union*. https://edupediapublications.org/journals/index.php/IJR/

15. *V. Jyothsna, B. N. Madhuri, K. S. Lakshmi, K. Himaja, B. Naveen and K. D. Royal, "Facemask detection using Deep Learning,"* 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), *Coimbatore, India, 2023, pp. 533-537, doi: 10.1109/ICISCoIS56541.2023.10100472.*

16. *J. S. Shankar and M. M. Latha, "Troubleshooting SIP Environments,"* 2007 10th IFIP/IEEE International Symposium on Integrated Network Management, *Munich, Germany, 2007, pp. 601-611, doi: 10.1109/INM.2007.374823.*

17. *S. Velan* et al.*, "Dual-Band EBG Integrated Monopole Antenna Deploying Fractal Geometry for Wearable Applications," in* IEEE Antennas and Wireless Propagation Letters*, vol. 14, pp. 249-252, 2015, doi: 10.1109/LAWP.2014.2360710.*

18. *S. Holm, T. M. Pukkila and P. R. Krishnaiah, "Comments on "On the use of autoregressive order determination criteria in univariate white noise tests" (reply and further comments)," in* IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 10, pp. 1805-1806, Oct. 1990, doi: 10.1109/29.60113.*

19. *Z. . -D. Bai, P. R. Krishnaiah and L. . -C. Zhao, "On rates of convergence of efficient detection criteria in signal processing with white noise," in* IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 380-388, March 1989, doi: 10.1109/18.32132.*