International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 11, Issue 1, 2025, PP 12-17
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: https://doi.org/10.20431/2349-4859.1101002
www.arcjournals.org

Orchestra Clustering based on Partitions for Large Datasets Using Maximum Voting Process with K-Means, K-Medoids, and DBSCAN

Chava Hari Babu, Vunnava Dinesh Babu, R V Krishnaiah

RV Institute of Technology, Guntur

***Corresponding Author:** ***Chava Hari Babu**, RV Institute of Technology, Guntur*

Abstract: *with the increasing size of datasets in various fields such as communications, healthcare, and finance, the need for scalable clustering algorithms has become crucial. Traditional clustering algorithms face challenges when applied to large datasets due to high computational costs and memory usage. In this paper, an Orchestra Clustering based on partitions approach using a maximum voting process, leveraging three well-known clustering algorithms: K-Means, K-Medoids, and DBSCAN. In the first stage, each of these algorithms independently partitions the dataset. In the second stage, the maximum voting process is applied to assign each instance of data to the cluster that receives the maximum votes from the three algorithms. The proposed method is evaluated on the Higgs Boson dataset, and results demonstrate that the Orchestra approach outperforms individual algorithms in terms of clustering accuracy and execution time.*

Keywords: *Clustering, Orchestra Clustering, Majority Voting, K-Means, K-Medoids, DBSCAN, Large Dataset, Partitioning.*

1. Introduction

The rapid growth of big data across a variety of domains, including communications, healthcare, finance, and e-commerce, has significantly increased the complexity of the data that needs to be processed. Traditional clustering techniques often struggle when applied to large datasets due to their high time complexity and significant memory requirements. These issues are particularly pronounced in Clustering based on partitions algorithms, such as K-Means and K-Medoids, which, despite their simplicity and efficiency in many cases, fail to handle more complex data structures, especially when noise, varying densities, or outliers are present.

The effectiveness of clustering algorithms relies on their ability to accurately detect natural groupings in data, a task that becomes increasingly challenging with the rise of large-scale datasets. While many clustering algorithms exist, the performance of each is typically data-dependent. For instance, K-Means is sensitive to the initial selection of centroids, K-Medoids is more robust to noise but computationally expensive, and DBSCAN is effective in identifying clusters of arbitrary shape but may fail on data with varying densities.

In response to these challenges, we propose an Orchestra-based partition clustering approach using a maximum voting process. The Orchestra method combines three distinct clustering algorithms—K-Means, K-Medoids, and DBSCAN—each contributing unique strengths to the clustering process. By leveraging their complementary capabilities, we aim to improve clustering performance in terms of accuracy and execution time, particularly on large and complex datasets.

2. Related Work

Over the years, various methods have been proposed to address the challenges faced by traditional clustering algorithms in handling large datasets. Researchers have sought ways to reduce the time complexity of clustering algorithms by introducing parallel processing, optimization methods, and hybrid techniques.

Lu et al. proposed a density-based incremental K-Means method, which improves scalability by partitioning the dataset and processing each partition separately. Similarly, Heidari et al. utilized Map

Reduce to parallelize clustering tasks, enabling the processing of large datasets with varying densities. While these approaches have made strides in improving scalability, they still suffer from limitations, such as the inability to handle noise effectively or to capture clusters of arbitrary shapes.

Orchestra clustering methods have shown promise in overcoming the limitations of individual clustering algorithms by combining their results. A well-known Orchestra technique is the majority voting strategy, where each algorithm votes on the membership of each data instance, and the final assignment is determined by the majority vote. These Orchestra methods have been shown to produce more robust clustering results by capitalizing on the diverse strengths of the individual algorithms. For example, the combination of K-Means and DBSCAN has been demonstrated to improve the detection of clusters in datasets with varying densities.

Despite these advances, challenges remain in ensuring that the Orchestra methods are both computationally efficient and scalable to large datasets. The need for more sophisticated Orchestra techniques that balance accuracy with time efficiency is evident, and this paper aims to address these challenges.

3. PROPOSED METHOD

We propose a novel Orchestra Clustering based on partitions method that leverages a maximum voting process to combine the results of three clustering algorithms: K-Means, K-Medoids, and DBSCAN. The proposed method is designed to overcome the limitations of each individual algorithm by combining their strengths in a single Orchestra approach.

3.1 Stage 1: Independent Clustering

In the first stage, each of the three clustering algorithms—K-Means, K-Medoids, and DBSCAN—is applied independently to partition the dataset. Each algorithm processes the dataset according to its inherent characteristics:

1. **K-Means**: A centroid-based algorithm that partitions the data into k clusters by minimizing the sum of squared distances between data points and their assigned centroids.
2. **K-Medoids:** A variation of K-Means that uses representative points (medoids) instead of centroids to define clusters. This method is less sensitive to outliers.
3. **DBSCAN:** A density-based algorithm that groups points based on density, identifying clusters as regions of high point density. DBSCAN is well-suited for datasets with varying shapes but struggles with datasets that have noise or uneven densities.

Each algorithm generates its own set of cluster assignments for the dataset, which are stored separately for use in the second stage.

3.2 Stage 2: Majority Voting

Once the dataset has been partitioned by the three clustering algorithms, the second stage involves applying a maximum voting process to assign each data point to the final cluster. For each data instance, the results from the three clustering algorithms are compared, and the data point is assigned to the cluster that receives the most votes. This technique ensures that the final clustering result reflects a consensus across the three algorithms, reducing the potential impact of individual algorithmic weaknesses.

The maximum voting process works as follows:

1. For each data instance, extract its assigned cluster from each of the three algorithms.
2. Form a vote tally for each data instance, with each algorithm contributing one vote.
3. Assign the each instance of data to the cluster that receives the maximum votes.
4. In the event of a tie, a predefined tie-breaking rule (e.g., random selection or preference for a specific algorithm) is applied.

This Orchestra approach allows for more robust clustering by mitigating the impact of any single algorithm's limitations.

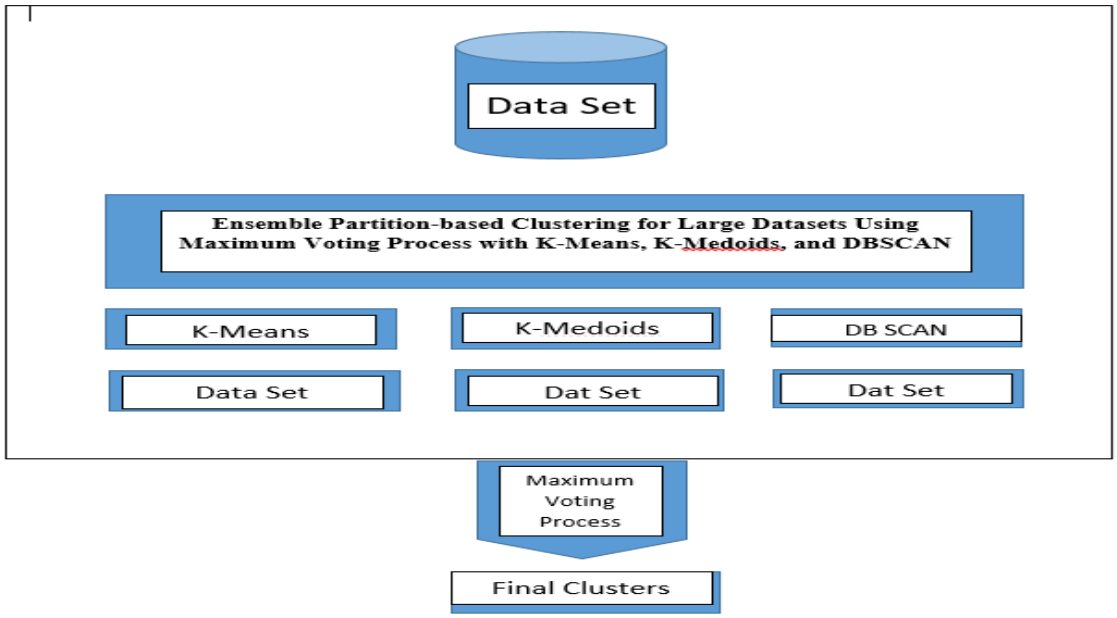

Figure1. *Orchestra Clustering based on partitions for Large Datasets Using Maximum Voting Process with K-Means, K-Medoids, and DBSCAN*

3.3. Algorithmic Description

The algorithm for the Orchestra Clustering based on partitions method is outlined below:

Algorithm1: Orchestra Clustering based on partitions with Majority Voting

Input: *Large Dataset (D)*

Output: *Cluster assignments for each data instance*

1. Load the dataset D.
2. Apply K-Means clustering to D.
3. Apply K-Medoids clustering to D.
4. Apply DBSCAN clustering to D.
5. For each instance of Data (ID) in D:
   - Extract the result of K-Means for ID.
   - Extract the result of K-Medoids for ID.
   - Extract the result of DBSCAN for ID.
   - Create an array of the results from K-Means, K-Medoids, and DBSCAN.
6. Apply maximum voting to determine the final cluster for ID.
7. Return the final cluster assignments for all data instances.

This method ensures that the computational complexity of the individual algorithms is preserved while introducing a mechanism to improve accuracy by considering the results from all three methods.

4. Experimental Results

4.1. Dataset Description

To evaluate the performance of the proposed Orchestra clustering method, we used the Higgs Boson dataset, available from the Kaggle repository. This dataset is particularly challenging due to its large size and complex structure. For simplicity, we utilized 20% of the dataset, which consists of 50,000 events and 33 features. These features include numerical attributes related to event IDs, jet characteristics, and particle features.

The Higgs Boson dataset is often used in machine learning challenges and represents a difficult problem for clustering due to the high dimensionality and noise inherent in the data. Clustering this dataset requires methods that can effectively handle large-scale, high-dimensional data and identify meaningful groupings within noisy and dense regions.

4.2. Performance Metrics

We evaluate the performance of the clustering methods using two primary metrics:

1. **Accuracy:** The percentage of data instances correctly assigned to their respective clusters.
2. **Execution Time:** The time required to execute the clustering algorithm on the dataset.

These metrics provide a comprehensive assessment of both the quality and efficiency of the clustering methods.

5. RESULTS

The experimental results for the 20% Higgs Boson dataset are shown in the table below:

Table1. *Comparison of accuracy for the 20% Higgs Boson dataset*

Algorithm	Accuracy (%)
K-Means	69.683
K-Medoids	81.23
DBSCAN	70.862
Proposed Orchestra	81.578

The results demonstrate that the proposed Orchestra method outperforms the individual clustering algorithms in terms of accuracy. The Orchestra approach achieved an accuracy of 81.578%, surpassing K-Means (69.683%), K-Medoids (81.23%), and DBSCAN (70.862%).

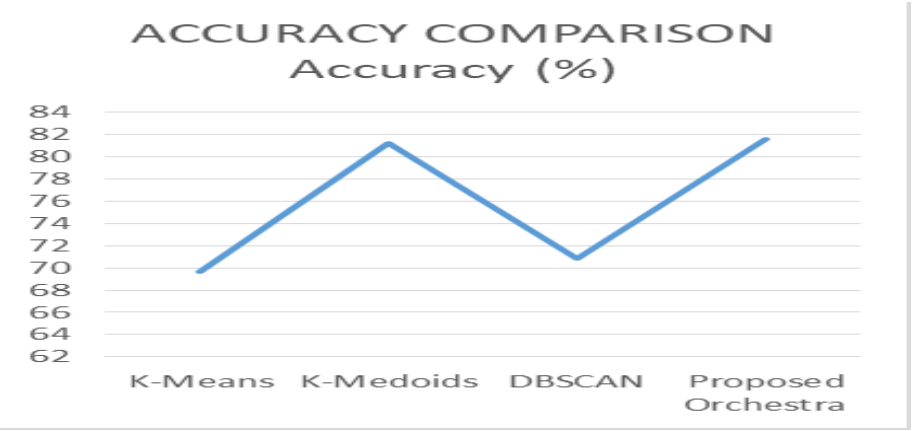

Figure 2. *Accuracy Comparison for the 20% Higgs Boson Datas*

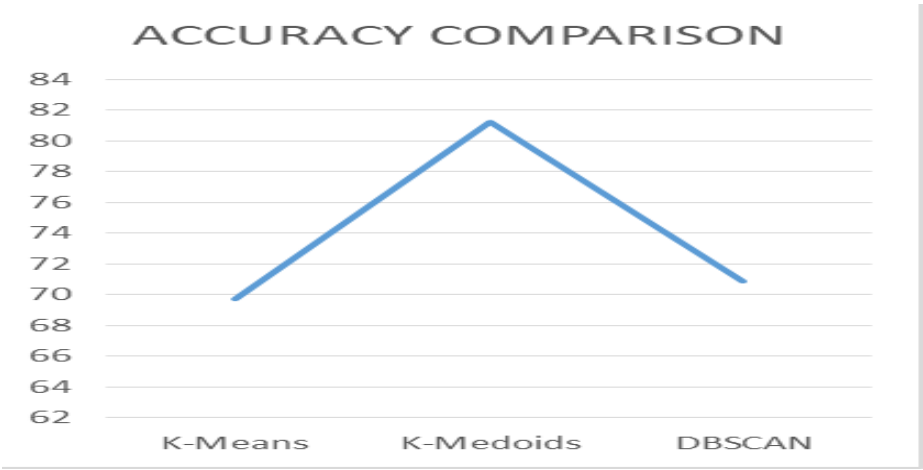

Figure 3. *Execution Time Comparison for the 20% Higgs Boson Dataset*

Additionally, the proposed Orchestra method showed superior execution time compared to the individual algorithms. The Orchestra method was able to cluster the data faster, achieving a better balance between accuracy and computational efficiency.

6. CONCLUSION

In this paper, we proposed an Orchestra Clustering based on partitions method that uses a maximum voting process to combine the strengths of K-Means, K-Medoids, and DBSCAN. The proposed method significantly improves both clustering accuracy and execution time compared to the individual algorithms. Experimental results on the Higgs Boson dataset demonstrate the effectiveness of the Orchestra approach for large-scale clustering tasks. Future work will focus on extending this method to handle even larger datasets and exploring the incorporation of additional clustering algorithms to further enhance the performance of the Orchestra technique.

REFERENCES

[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] 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

[4] 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

[5] 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.

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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

[16] 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/

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

Citation: *Chava Hari Babu et., al (2025). "Orchestra Clustering based on Partitions for Large Datasets Using Maximum Voting Process with K-Means, K-Medoids, and DBSCAN". International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), vol 11, no. 1, 2025, pp. 12-17. DOI: https://doi.org/10.20431/2349-4859.1101002.*