International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 8.141
Peer Reviewed Bi-Monthly
Vol. 6 No. 38 September - October - 2025

A Comparative Study Of K-Means And Parallel K-Means Clustering Algorithms For Efficient Data Analysis

Kudale Gautam Appasaheb[1], Shinde Monika[2] & Dr. Gupta Gaurav[3]
[1&2]*Research Students, Dr. A.P.J. Abdul Kalam University, Indore, M.P., India*
[3]*Research Guide, Dr. A.P.J. Abdul Kalam University, Indore, M.P., India*
Corresponding Author –M. Sadani Kudale Gautam Appasaheb
DOI - 10.5281/zenodo.17312746

Abstract:

Clustering group's unlabeled data into meaningful patterns. K-Means, a popular partition-based algorithm, offers simplicity and efficiency but struggles with large-scale, high-dimensional data due to scalability and initialization issues. Parallel K-Means addresses these limitations by utilizing parallel and distributed computing frameworks, enhancing performance, scalability, and computational efficiency in clustering tasks. This paper presents a comparative study of traditional K-Means and Parallel K-Means clustering algorithms. It reviews clustering techniques and algorithms, analyzes their methodologies, advantages, and limitations, and highlights the importance of parallel approaches. The study concludes by emphasizing parallelism's role in enhancing clustering efficiency and scalability for data-intensive applications.

Keywords: Machine Learning, Data mining, Clustering, K-Means Clustering, Parallel K-Means Clustering

Introduction:

Data mining involves extracting useful information from large databases, aiding organizations in retrieving valuable insights from data warehouses. Applicable to various database types, it is widely used in sectors like banking, insurance, and pharmaceuticals. As a branch of machine learning, data mining emphasizes exploratory data analysis and is key in predictive analytics. [16]

Machine learning (ML), a subset of artificial intelligence, enables computers to learn from data and make predictions without explicit programming. ML algorithms build models from training data for tasks like prediction and decision-making. It includes supervised, unsupervised, and reinforcement learning, with applications in education, pattern recognition, sports, and industry. [25].

Data clustering is the process of grouping a set of objects that objects is the same groups are more similar to each other than to those in other groups. [20] As datasets grow larger and more complex, efficient clustering techniques are essential for uncovering hidden patterns in high-dimensional data. Clustering aids applications like customer segmentation and genomic analysis. The increasing demand for data-driven insights has driven the widespread adoption of scalable, efficient clustering algorithms across diverse machine learning domains.

K-Means Clustering:

K-means is one of the easiest algorithms of unsupervised learning used for clustering [5]. K-means is one of the simplest

unsupervised learning algorithms used for clustering. [5, 7, 10, 20] K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. The K-Means method is numerical, unsupervised, non-deterministic and iterative [9]. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. [10, 20] The K-Means clustering technique is used to classify data in a crisp sense. [12] K-means is an old and widely used technique in clustering method [15]. K-means is the most popular clustering algorithm commonly used in all metric spaces [18]

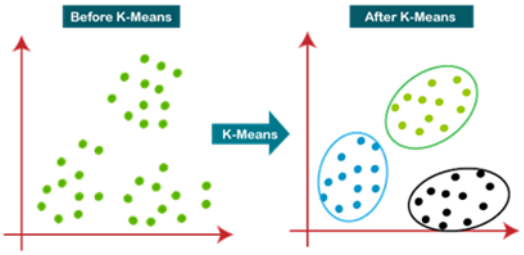

Fig. 1: Working of K-means Clustering Algorithm

A. Generalised Pseudocode of Traditional k-means [5, 8, 9, 15, 22, 24]

Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values

B. Flowchart

The below diagram explains the working of the K-means Clustering Algorithm:

Step 2: Initialize the first K clusters

- Take first k instances or
- Take Random sampling of k elements

Step 3: Calculate the arithmetic means of each cluster formed in the dataset.

Step 4: K-means assigns each record in the dataset to only one of the initial clusters - Each record is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).

Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

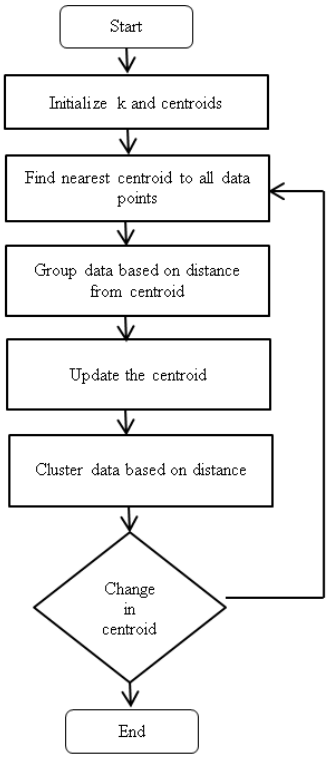

Fig. 2: Flow Chart of K-means Clustering Algorithm

C. Advantages of K-Means Clustering:

- Simple, easy to understand, and implement. [5,7,10,20]
- Fast convergence and low computational cost. [15, 25]
- Scalable to large datasets and adaptable to sparse data. [6]
- Assembles stable and tight clusters efficiently. [24]
- High efficiency and widely used across various fields due to its iterative, unsupervised, and non-deterministic nature. [28]

D. Disadvantages of K-Means Clustering:

- Requires predefined number of clusters (K).
- Sensitive to initial centroid selection, risking suboptimal results.
- Assumes spherical, equally sized clusters.
- Poor performance with non-linear or complex cluster shapes.
- Highly sensitive to outliers and noise.
- Not suitable for categorical data without preprocessing.
- Requires feature scaling for accurate results. [28]

E. Commonly used cluster evaluation metrics for K-Means Clustering:

- Inertia (WCSS): Measures compactness within clusters; lower is better.
- Silhouette Score: Evaluates cohesion and separation; ranges from -1 to 1.
- Calinski-Harabasz Index: Higher values indicate better-defined clusters.
- Davies-Bouldin Index: Lower values suggest better clustering.
- Dunn Index: Higher is better.
- Adjusted Rand Index (ARI) and Purity: Compare clustering to true labels.

F. Challenges of K-Means Clustering:

- Requires selecting the optimal number of clusters (K).
- Sensitive to initial centroid placement, risking local minima.
- Assumes spherical, similarly sized clusters.
- Ineffective for non-linear or complex cluster boundaries.
- Outliers and noise can distort clustering results.
- Scalability issues with large or high-dimensional datasets.
- Requires numerical data and proper feature scaling.

Parallel K-Means Clustering:

Parallel K-Means is an optimized version of the traditional K-Means algorithm designed to handle large-scale and high-dimensional datasets by leveraging parallel and distributed computing. It accelerates the computation by performing clustering operations simultaneously across multiple processors or nodes.

The Parallel K-Means Clustering Algorithm is an enhanced version of the traditional K-Means, optimized for parallel and distributed computing environments. It divides computation among multiple processors or nodes to efficiently cluster large-scale and high-dimensional datasets.

A. Generalised Pseudocode for Parallel K-Means

Initialize K centroids
Broadcast centroids to all processors
repeat
// Parallel Assignment Step
for each processor in parallel:
Assign local data points to nearest centroids

Compute partial sums and counts for each cluster

// Global Aggregation Step

Gather all partial sums and counts

Compute new centroids globally

Broadcast updated centroids to all processors until convergence criteria met

A. Flowchart

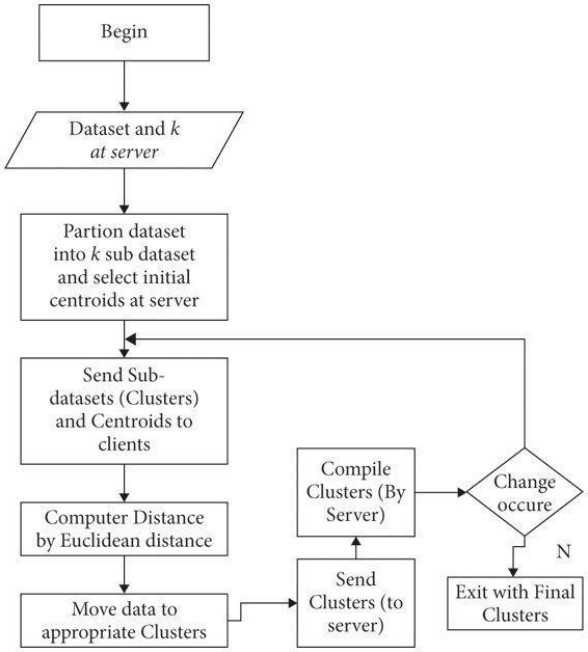

Fig. 3: Flow Chart of Parallel K-means Clustering Algorithm

B. Advantages of Parallel K-Means Clustering:

- Enhances scalability for large and high-dimensional datasets.
- Reduces computation time through parallel processing.
- Efficiently handles big data using distributed computing frameworks.
- Maintains clustering accuracy while improving performance.
- Suitable for real-time and data-intensive applications.
- Balances workload across multiple processors or nodes, ensuring faster convergence.

C. Disadvantages of Parallel K-Means Clustering:

- Requires complex parallel and distributed computing infrastructure.
- Communication overhead between processors can affect efficiency.
- Load balancing issues may arise in heterogeneous environments.
- Sensitive to initial centroid selection, similar to traditional K-Means.
- Scalability can be limited by hardware and network constraints.
- Increased implementation complexity compared to standard K-Means.

D. Commonly used cluster evaluation metrics for Parallel K-Means Clustering:

- Inertia (WCSS): Measures compactness within clusters; lower values are better.
- Silhouette Score: Evaluates cohesion and separation between clusters; higher is better.

- Calinski-Harabasz Index: Assesses cluster separation; higher values indicate well-defined clusters.
- Davies-Bouldin Index: Lower values reflect better clustering.
- Adjusted Rand Index (ARI): Compares clustering against ground truth labels.

E. Challenges/Limitations of Parallel K-Means Clustering:

- Requires complex parallel or distributed computing setup.
- Communication overhead between nodes can reduce efficiency.
- Sensitive to initial centroid selection, risking suboptimal clustering.
- Scalability may be constrained by hardware and network limitations.
- Load balancing issues in heterogeneous systems.
- Increased algorithmic and implementation complexity compared to traditional K-Means

IV. Conclusion:

This study presents a comprehensive comparative analysis of the traditional K-Means and Parallel K-Means clustering algorithms for efficient data analysis. The findings reveal that while K-Means offers simplicity and ease of implementation, it encounters limitations in handling large-scale and high-dimensional datasets due to scalability and computational inefficiencies. Parallel K-Means, leveraging parallel and distributed computing frameworks, significantly enhances clustering performance by improving scalability, reducing execution time, and handling data-intensive tasks more effectively. The study underscores the critical role of parallelism in modern clustering applications, providing a more robust and efficient approach for large-scale data analysis across diverse domains.

Future Work:

Future research will explore hybrid clustering models integrating deep learning techniques, such as autoencoders, with Parallel K-Means to enhance dimensionality reduction and clustering accuracy on complex, high-dimensional data. Scalability on heterogeneous computing environments will also be investigated.

References:

1. Data Mining Introductory and Advanced Topics, Margaret H. Dunhan, Pearson
2. Data Mining Practical Machine Learning Tools and Techniques, 3rd Edition, Ian H.witten, Eibe Frank, Mark A. Hall
3. Mining of Massive Datasets, 2nd Edition, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman
4. Data Mining, Concepts and Techniques, 3rd Edition, Jiawei Han, Micheline Kamber, Jian Pei
5. Prof. Prashant Sahai Saxena, Prof. M. C. Govil, “Prediction of Student’s Academic Performance using Clustering,” Special Conference Issue: National Conference on Cloud Computing & Big Data
6. Bindiya M Varghese, Jose Tomy J, Unnikrishnan A, Poulose Jacob K, “Clustering student data to characterize performance patterns,” (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence,
7. Md. Hedayetul Islam Shovon, Mahfuza Haque, “An Approach of Improving Student’s Academic Performance by using K-means clustering algorithm and Decision tree,” (IJACSA) International

Journal of Advanced Computer Science and Applications, Vol.3, No. 8, 2012

8. Oyelade, O. J, Oladipupo, O. O., Obagbuwa, I. C., "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance," (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010
9. Rakesh Kumar Arora, Dr. Dharmendra Badal, "Evaluating Student's Performance Using k-Means Clustering," International Journal of Computer Science And Technology, IJCST Vol. 4, Issue 2, April - June 2013, ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)
10. Sharmila, R.C Mishra, "Performance Evaluation of Clustering Algorithms," International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue7- July 2013, ISSN: 2231-5381
11. Patel, J. and Yadav, R.S. (2015) "Applications of Clustering Algorithms in Academic Performance Evaluation." Open Access Library Journal, 2: August 2015 | Volume 2 | e1623
12. Jyotirmay Patel, Ramjeet Singh Yadav, "Applications of clustering algorithms in academic performance evaluation"
13. E.Venkatesan, S.Selvaragini, "Prediction of students academic performance using classification and clustering algorithms," International Journal of Pure and Applied Mathematics Volume 116 No. 16 2017, 327-333 ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
14. Snehal Bhogan , Kedar Sawant , Purva Naik , Rubana Shaikh , Odelia Diukar , Saylee Dessai, "Predicting student performance based on clustering and classification," IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN:2278-8727, Volume 19, Issue 3, Ver. V (May-June 2017), PP 49-52
15. Mr. Shashikant Pradip Borgavakar, Mr. Amit Shrivastava, "Evaluating student's performance using k-means clustering," International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 6 Issue 05, May – 2017
16. Mrs .Mary vidya john, Akshata police patil, Anjali mishra, Bindhu reddy G, Jamuna N, "Clustering technique for student performance," International Research Journal of Computer Science (IRJCS), Issue 06, Volume 6 (June 2019), ISSN: 2393-9842
17. Noel Varela , Edgardo Sánchez Montero , Carmen Vásquez , Jesús García Guiliany , Carlos Vargas Mercado , Nataly Orellano Llinas , Karina Batista Zea , and Pablo Palencia, "Student performance assessment using clustering techniques," © Springer Nature Singapore Pte Ltd. 2019 Y. Tan and Y. Shi (Eds.): DMBD 2019, CCIS 1071, pp. 179–188, 2019. https://doi.org/10.1007/978-981-32-9563-6_19
18. N.Valarmathy, S.Krishnaveni, "Performance evaluation and comparison of clustering algorithms used in educational data mining," International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6S5, April 2019
19. Lubna Mahmoud Abu Zohair, "Prediction of Student's performance by modelling small dataset size," Abu Zohair International Journal of Educational Technology in Higher Education (2019) 16:27 https://doi.org/10.1186/s41239-019-0160-3

20. Mrs. Bhawna Janghel, Dr. Asha Ambhaikar, "Performance of student academics by k-mean clustering algorithm," International J. Technology. January – June, 2020; Vol. 10: Issue 1, ISSN 2231-3907 (Print), ISSN 2231-3915 (Online)
21. Marzieh Babaie, Mahdi Shevidi Noushabadi, "A review of the methods of predicting students' performance using machine learning algorithms," Archives of Pharmacy Practice ¦ Volume 11 ¦ Issue S1 ¦ January-March 2020
22. Dr. G. Rajitha Devi, "Prediction of student academic performance using clustering," International Journal of Current Research in Multidisciplinary (IJCRM) ISSN: 2456-0979 Vol. 5, No. 6, (June'20), pp. 01-05
23. Dewi Ayu Nur Wulandari; Riski Annisa; Lestari Yusuf, Titin Prihatin, "An educational data mining for student academic prediction using k-means clustering and naïve bayes classifier," journal Pilar Nusa Mandiri Vol 16, No 2 September 2020
24. Yann Ling Goh, Yeh Huann Goh, Chun-Chieh Yip, Chen Hunt Ting, Raymond Ling Leh Bin, Kah Pin Chen, "Prediction of students' academic performance by k-means clustering," Peer-review under responsibility of 4th Asia International Multidisciplinary Conference 2020 Scientific Committee
25. Revathi Vankayalapati, Kalyani Balaso Ghutugade, Rekha Vannapuram, Bejjanki Pooja Sree Prasanna, "K-means algorithm for clustering of learners performance levels using machine learning techniques," Revue d'Intelligence Artificielle Vol. 35, No. 1, February, 2021, pp. 99-104
26. Rina Harimurti, Ekohariadi, Munoto, I. G. P Asto Buditjahjanto, "Integrating k-means clustering into automatic programming assessment tool for student performance analysis," Indonesian Journal of Electrical Engineering and Computer Science Vol. 22, No. 3, June 2021, pp. 1389~1395 ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i3.pp1389-1395
27. Rui Shang , Balqees Ara, Islam Zada, Shah Nazir , Zaid Ullah, and Shafi Ullah Khan, "Analysis of simple k-mean and parallel k-mean clustering for software products and organizational performance using education sector dataset," Hindawi Scientific Programming Volume 2021, Article ID 9988318, 20 pages https://doi.org/10.1155/2021/9988318
28. Bao Chong, "K-means clustering algorithm: a brief review," Academic Journal of Computing & Information Science ISSN 2616-5775 Vol. 4, Issue 5: 37-40, DOI: 10.25236/AJCIS.2021.040506
29. Said Abubakar Sheikh Ahmed, "Evaluating students' performance of social work department using k-means and two-step cluster "a case study of mogadishu university"," Mogadishu University Journal, Issue 7, 2021, ISSN 2519-9781
30. Zhihui Wang, "Higher education management and student achievement assessment method based on clustering algorithm," Hindawi Computational Intelligence and Neuroscience Volume 2022, Article ID 4703975, 10 pages https://doi.org/10.1155/2022/4703975
31. Ahmad Fikri Mohamed Nafuri , Nor Samsiah Sani, Nur Fatin Aqilah Zainudin , Abdul Hadi Abd Rahman and Mohd Aliff, "Clustering analysis for classifying student academic performance in higher education," Appl. Sci. 2022, 12, 9467. https://doi.org/10.3390/app12199467

32. Othman, F., Abdullah, R., Rashid, N. A. A., & Salam, R. A. (2004, December). Parallel k-means clustering algorithm on DNA dataset. In *International Conference on Parallel and Distributed Computing: Applications and Technologies* (pp. 248-251). Berlin, Heidelberg: Springer Berlin Heidelberg.
33. Zhao, W., Ma, H., & He, Q. (2009). Parallel k-means clustering based on mapreduce. In *Cloud Computing: First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. Proceedings 1* (pp. 674-679). Springer Berlin Heidelberg.
34. Kerdprasop, K., & Kerdprasop, N. (2010). A lightweight method to parallel k-means clustering. *International Journal of Mathematics and Computers in Simulation*, *4*(4), 144-153.
35. Kumar, J., Mills, R. T., Hoffman, F. M., & Hargrove, W. W. (2011). Parallel k-means clustering for quantitative ecoregion delineation using large data sets. *Procedia Computer Science*, *4*, 1602-1611.
36. Jin, S., Cui, Y., & Yu, C. (2016). A new parallelization method for K-means. *arXiv preprint arXiv:1608.06347*.
37. Alguliyev, R. M., Aliguliyev, R. M., & Sukhostat, L. V. (2021). Parallel batch k-means for Big data clustering. *Computers & Industrial Engineering*, *152*, 107023.
38. Nigro, L. (2022). Performance of parallel K-means algorithms in Java. *Algorithms*, *15*(4), 117.