Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
January-2024 5(1), 1141–1150

ENHANCED UNSUPERVISED K-MEANS CLUSTERING ALGORITHM

Dr. Gowsic K[1], Mugunthan S[2], Sakthivel Logavaseekarapakther[3], Puviyarasu A[4], Mohammed Farook R[5]

[1] Associate professor, Department of Computer Science and Engineering, Mahendra Engineering College
[2, 3, 4, 5] UG students, Department of Computer Science and Engineering, Mahendra Engineering College

Corresponding Author

Dr. Gowsic K, kgowsic@gmail.com

DOI
10.29121/shodhkosh.v5.i1.2024.2867

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

K-Means clustering is an unsupervised learning algorithm for distinguishing data into separate groups called clusters based on similarity. However, the need to specify the cluster count (K) beforehand highly affects the effectiveness of the algorithm, which can be challenging in practice. In our manuscript, we introduce an improved iteration of the K-Means algorithm, which incorporates the elbow method to autonomously identify the required number of clusters during the clustering procedure. Our approach also incorporates optimization techniques to improve computational efficiency. The experimental findings substantiate the efficacy of our refined algorithm in automatically identifying the precise count of clusters while reducing computational overhead compared to traditional methods.

Keywords: Dynamic Clustering, Optimal Clusters, Clustering, K-Means Clustering, Algorithms, Computational Efficiency

1. INTRODUCTION

A fundamental problem that is wavering frequently in a various types of fields including machine learning and data mining, and pattern classification is clustering. Over the last ten years, the importance of data mining and machine learning has surged remarkably. In today's highly competitive market landscape, timely access to high-quality information is paramount for effective decision-making, particularly in policymaking. This phenomenon has garnered considerable attention not only within the information industry but also in broader society. There is very large amount of data availability in the real world and extracting the required information from this vast dataset proves to be quite challenging and provide the information to which it is needed within the specified time frame and in required pattern. Performing clustering analysis is a core component of the task in data mining and machine learning, aimed at grouping analogous data points into clusters relies on their similarities, ensuring that data with higher resemblance are grouped together within the same cluster, while those with lower resemblance are segregated.

Various types of data mining algorithms exist for clustering, encompassing density-based, hierarchical-based, partitioning-based, grid-based, and model-based approaches. In partition-based clustering, among them the most famous algorithms are K-Means clustering algorithm [6]. Partitioning algorithms endeavours to create a single partition

of a database X containing n objects, organizing them into a set of K clusters. Notably, among these algorithms, K-Means is distinguished for its simplicity and effectiveness. The objective is to identify K centroids in real d-dimensional space R, minimizing the Euclidean distance calculated for ever data point and its closest centroid.

Although K-Means clustering algorithm is an unsupervised algorithm it is partially unsupervised, as the need for users to pass the clusters count a priori, which can be subjective and challenging, particularly for real time large or high-dimensional datasets. In this paper, we address this limitation by proposing an enhanced unsupervised version of the K-Means algorithm that automatically determines the optimal and required clusters (K).

Our approach integrates the elbow method, a popular technique for identifying the ideal cluster quantity based on the within-cluster sum of squares (WCSS) curve. As using elbow method to find optimal cluster count (K) will increase time complexity, we introduce optimization techniques to enhance the computational efficiency of the algorithm, making it more appropriate for real-world applications. To enhance the computational efficiency of the algorithm, we incorporate several optimization techniques:

1.1 MANHATTAN DISTANCE

Instead of using the traditional Euclidean distance metric, we employ the Manhattan distance metric, which is computationally less intensive, particularly in high-dimensional spaces.

1.2 OPTIMIZING ITERATIONS

We optimize the calculation of distances between data points and cluster centroids by storing previous distances and only recalculating distances for data points likely to change clusters based on centroid adjustments. In classical K-Means, the algorithm calculates the distance between every pair of data points and every cluster centroid in each iteration, which can incur significant computational costs, particularly with large datasets.

1.3 TERMINATION CONDITION

We introduce a termination condition for the elbow method to stop the iterative process once the rate of change in WCSS falls below a certain threshold, thus reducing unnecessary computation.

The present paper is therefore organized as follows: in section 2 we discuss the related works done by others in K-Means clustering algorithm; in section 3 and 4 we discuss the basic details and workflow of both elbow method and K-Means clustering algorithm respectively; in section 5 and 6 we introduce the proposed algorithm; in section 7 we discuss the results, and in section 8 we finalize our findings.

2. RELATED WORK

Multiple researchers have enhanced both the effectiveness and performance of the K-Means clustering algorithm, encompassing advancements in cluster quality [1] and the algorithm's runtime [2][3].

Kristina P. Sinaga and Miin Shen Yang [1] introduced an innovative unsupervised learning framework for the k-means clustering algorithm, eliminating the need for initializations and parameter selection while concurrently determining an optimal clusters count. Their proposed method, termed unsupervised k-means (U k-means), utilizes the concept of entropy to automatically identify the clusters count without requiring any initialization or parameter selection.

Various distance metrics can be employed to calculate point-to-point distances. Taher M Ghazal, Muhammad Zahid Hussain [2] evaluated K-Means clustering with three different mathematical metrics regarding execution time with different datasets and different number of clusters.

Fahim A M, Salem A M, Torkey F A, Ramadan M A [3] proposed an efficient enhanced k-means algorithm by creating a rudimentary data structure to retain essential information within each iteration for subsequent use. Thus, reducing the unnecessary computation and increasing the efficiency of the algorithm.

3. ELBOW METHOD

Clustering constitutes a foundational task in data analysis, entailing the segmentation of data points into clusters according to their similarities. Establishing the ideal number of clusters is crucial for effective clustering, as it directly impacts the quality and interpretability of the results. The elbow method is a popular heuristic technique used to identify the optimal clusters in a dataset, primarily in conjunction with K-Means clustering [7]. In this section, we offer an exhaustive examination of the elbow method., including its principles, applications, limitations, and variations.

The term "elbow method" originates from its characteristic plot shape resembling an elbow. The elbow method, also known as the "scree" method, is a simple yet effective technique for identifying the optimal number of clusters in a dataset.

The method aims to pinpoint the juncture on the plot where the rate of decline in the within-cluster sum of squares (WCSS) slows down significantly. WCSS (Within-Cluster Sum of Squares) signifies the sum of squared distances between every data point and its respective cluster centroid. As the clusters increases, WCSS typically decreases, as clusters become more specific to the data points they contain. However, beyond a certain point, the rate of decrease in WCSS starts to diminish, resulting in a bending or "elbow" in the plot. The point at which this bend occurs is considered the optimal clusters.

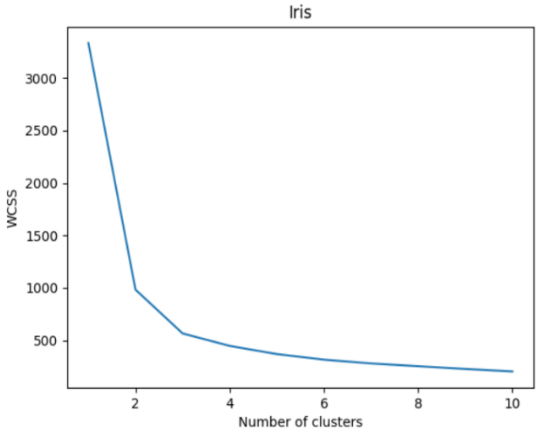

Fig.1 Elbow plot for iris dataset. Elbow plot for iris dataset is plotted and the required clusters is found to be 3.

The elbow method is widely used in various domains and applications to determine the clusters count in clustering analysis. It provides a simple yet effective means of selecting an appropriate value for K, thereby aiding in the interpretation and visualization of clustered data. Common applications of the elbow method include customer segmentation, market segmentation, image segmentation, and pattern recognition. By identifying the clusters count, the elbow method facilitates more accurate and meaningful clustering results, enabling insights and decision-making in diverse fields.

Although the elbow method serves as a valuable heuristic for identifying the optimal number of clusters, it does have its limitations. One major limitation is its subjective interpretation, as the elbow point may not always be clearly defined or intuitive, particularly in datasets with complex or overlapping structures.

Additionally, the effectiveness of the elbow method can be influenced by factors such as dataset size, dimensionality, and cluster shape. Using the elbow method in conjunction with other validation techniques and domain knowledge is essential to ensure robust and reliable clustering results.

Several variations and extensions of the elbow method have been proposed to address its limitations and enhance its utility. These include the use of alternative metrics for evaluating cluster quality, such as silhouette analysis, gap statistics, and Davies–Bouldin index. Additionally, hierarchical clustering techniques and model-based clustering algorithms offer alternative approaches to determining the effective number of clusters. By incorporating complementary validation techniques and exploring alternative clustering methodologies, researchers and practitioners can overcome the limitations of the elbow method and enhance the efficiency of clustering analysis.

In summary, the elbow method provides a simple yet powerful heuristic for determining the optimal clusters count in clustering analysis. By identifying the point of diminishing returns in the rate of decrease in within-cluster sum of squares, the elbow method enables data-driven selection of the optimal number of clusters, facilitating more accurate and interpretable clustering results.

While the elbow method has limitations and considerations, it remains a valuable tool in the clustering toolbox, particularly when used alongside other validation techniques and domain knowledge.

Continued research and development in this domain will further augment the effectiveness and applicability of the elbow method in clustering analysis.

4. K-MEANS CLUSTERING ALGORITHM

K-Means clustering is widely recognized as one of the most extensively used unsupervised machine learning algorithms for dividing data into distinct groups, or clusters, based on similarities among data points. It is a centroid-based

algorithm that It progressively allocates data points to clusters and adjusts cluster centroid to minimize the sum of squares within each cluster. In this section, we offer an in-depth overview of the K-Means clustering algorithm, including its principles, steps, applications, strengths, and limitations. At the core of K-Means clustering is the notion of similarity between data points, typically measured using a distance metric such as Euclidean distance. The goal of K-Means clustering is to partition the data into K clusters, where K is a user-defined parameter. The algorithm repeatedly assigns each data point to the nearest cluster centroid and updates the centroids based on the mean of the data points assigned to each cluster. This process never stops until convergence is reached, signifying that the cluster assignments have stabilized and are no longer changing significantly.

4.1 INITIALIZATION

The process commences with the random initialization of K cluster centroids. These centroids serve as the initial representatives of the clusters.

4.2 ASSIGNMENT

Subsequently, every data point gets allocated to the nearest cluster centroid, typically determined by a distance metric such as the Euclidean distance. The distance between a data point and a centroid is calculated, and the data point is allocated to the cluster corresponding to the nearest centroid.

4.3 UPDATE

After assigning all data points to clusters, the algorithm proceeds to update the cluster centroids. Each centroid is recalculated as the mean of all data points assigned to its corresponding cluster. This step aims to reposition the centroids to better represent the data points within each cluster.

4.4 CONVERGENCE

The assignment and update steps are reiterated iteratively until convergence is attained. Convergence occurs when the cluster assignments no longer change significantly, indicating that the algorithm has reached a stable solution. The primary goal of K-Means clustering is to minimize the within-cluster sum of squares (WCSS), alternatively known as inertia or distortion. The within-cluster sum of squares (WCSS) quantifies the total sum of squared distances between every data point and its respective cluster centroid. The algorithm aims to discover cluster centroids that minimize the overall WCSS across all clusters. One of the main hurdles in employing K-Means clustering is deciding the appropriate number of clusters (K) for a given dataset. Traditionally, users must specify the K value a priori, which can be subjective and challenging, particularly in unsupervised settings.

```
1 Import dataset
2  Initialise random centroids
3  While (Centroids not equal to Old centroids)
4    Old centroids = centroids
5    For i=1 to n
6      For j=1 to k
7        Compute Euclidean distance
8      Endfor
9      Assign datapoint to the closest centroid
10   Endfor
11   Centroids = updatecentroids
```

Fig.2 K-Means algorithm. Basic step by step approach for K-Means.

While K-Means clustering offers simplicity and efficiency, it is prone to being influenced by the initial placement of cluster centroids and may converge to local optima depending on the initialization. Additionally, the algorithm requires the number of clusters (K) to be predefined, which might not consistently correspond to the inherent structure of the data. In summary, K-Means clustering is a versatile algorithm used for partitioning data into clusters based on similarity. While it offers simplicity and efficiency, careful consideration must be given to initialization and the determination of the clusters.

In the subsequent sections, we present our enhanced version of the K-Means algorithm, which addresses a few of these limitations by automating the determination of K and optimizing computational efficiency.

5. INTERMEDIATE UNSUPERVISED K-MEANS CLUSTERING ALGORITHM

In our proposed "Enhanced Unsupervised K-Means Clustering Algorithm" we improve it upon the traditional approach by automating the determination of the expected cluster count and optimizing computational efficiency. This intermediary iteration of the unsupervised K-Means clustering algorithm serves as a stepping stone toward refining the proposed enhanced version, that only finding optimal clusters is implemented in this section. While optimising this version is made in the next section.

5.1 FINDING OPTIMAL CLUSTERS

To automate the selection of the optimal number of clusters, we incorporate the elbow method. Utilizing this technique, our algorithm efficiently identifies the elbow point in the within-cluster sum of squares (WCSS) plot, removing the necessity for users to predefine the number of clusters. This integration enhances the usability and flexibility of our algorithm, making it suitable for a variety of clustering tasks.

The elbow method's workflow involves iteratively applying the K-Means algorithm to the same dataset with varying values of K to determine the optimal clusters count, which is then plotted against the within-cluster sum of squares (WCSS). It would consume huge amount of time even for smaller datasets. Here comes the need for a termination condition to avoid unnecessary computations, which is been implemented in the next section.

5.2 MANHATTAN DISTANCE

The conventional K-Means clustering method utilizes the Euclidean distance formula to assess the similarities among data points. Transitioning from Euclidean distance to Manhattan distance entails adjusting the manner in which distances are computed between data points and cluster centroids within the K-Means clustering procedure. The Euclidean distance represents the straight-line distance between two points within a Euclidean space, while Manhattan distance, also known as city block distance or taxicab distance, measures the measurement of the space separating two points along perpendicular axes.

In the Enhanced Unsupervised K-Means clustering algorithm, this modification is implemented in the step where distances between data points and cluster centroids are calculated during the assignment phase. Instead of using the Euclidean distance formula, the Manhattan distance formula is used to compute the distances. Changing from Euclidean distance to Manhattan distance can have implications for the clustering process. The Manhattan distance is often less affected by outliers compared to the Euclidean distance. The main rationale for replacing Euclidean distance with Manhattan distance is its reduced computational complexity, particularly noticeable in high-dimensional scenarios.

5.3 OPTIMIZING ITERATIONS

Iteration optimization is a method employed to enhance the efficiency of the K-Means clustering algorithm by diminishing the number of distance calculations conducted in each iteration. In conventional K-Means, the algorithm computes the distance between every data point and every cluster centroid in each iteration, which can be computationally demanding, particularly for extensive datasets. Optimizing iterations aims to minimize unnecessary distance computations without compromising the clustering quality. In the classic K-Means algorithm, during the assignment step, the computation of distances between each data point and every cluster centroid is conducted employing a distance metric such as Euclidean distance. This process entails calculating the distances between each data point and each cluster centroid, incurring a substantial computational expense, especially noticeable with large datasets.

```
1  Import dataset
2  For k=1 to 10
3    Initialise random centroids
4    While (Centroids not equal to Old centroids)
5      Old centroids = centroids
6      For i=1 to n
7        Compute Manhattan distance to their centroid
8        If distance > previous distance
9          For j=1 to k
10             Compute Manhattan distance
11           Endfor
12           Assign datapoint to the closest centroid
13       Update previous distance
14     Endfor
15     Centroids = updatecentroids
16 Endfor
```

Fig.3 Intermediate unsupervised K-Means algorithm. Basic step by step approach for intermediate unsupervised K-Means.

To optimize distance calculations, a data structure is introduced to store the distances separating data points and cluster centroids from the previous iteration. Initially, all distances are calculated and stored in this data structure.

In successive iterations, instead of recalculating distances for all data points and cluster centroids, the algorithm contrasts the current distance between a data point and its allocated cluster centroid with the distance stored in the data structure from the prior iteration.

If the present distance exceeds the stored distance, it suggests that the data point may have shifted away from its initial cluster centroid owing to centroid adjustments. The algorithm recalculates the distance from the data point to all cluster centroids only if the current distance is greater than the stored distance. This conditional approach to distance calculation significantly diminishes the count of distance computations, focusing particularly on those data points likely to change their cluster assignments.

Optimizing iterations can lead to significant improvements in computational efficiency, especially for datasets with a large number of data points and clusters. By avoiding unnecessary distance calculations, the algorithm converges faster, resulting in reduced runtime and resource consumption.

6. ENHANCED UNSUPERVISED K-MEANS CLUSTERING ALGORITHM

In the previous section the problem with finding optimal clusters by implementing elbow method is done. Performing the elbow method involves iteratively running the K-Means algorithm on the identical dataset for various values of K and plotting the within-cluster sum of squares (WCSS). This process can be time-consuming, even for smaller datasets, because of the repetitive nature of the computation. Here comes the need for a termination condition to avoid unnecessary computations.

6.1 OPTIMISING

As the clusters count is found once the elbow shape in a plot is reached, it is unnecessary to iterate the process for other values of K. In order to find whether we had reached the elbow shape or not, we are considering the variation among the WCSS value for K=1 and K=2 and let it be D. Thus, if the elbow shape has obtained then the difference between the last performed K value and K-1 value would obviously less than or equal to D/10. Hence the termination condition is that to iterate the process only if difference between the WCSS value of K and K-1 is greater than D/10. Then at last performing clustering once again with K-1 clusters to get the dataset clustered into actual number of clusters.

```
Import dataset
For k=1 to 10
    Initialize random centroids
    While (Centroids not equal to Old centroids)
        Old centroids = Centroids
        For i=1 to n
            Compute Manhattan distance to their centroid
            If distance > previous distance
                For j=1 to k
                    Compute Manhattan distance
                Endfor
                Assign datapoint to the closest centroid
        Update previous distance
        Endfor
        Centroids = updatecentroids
        Compute inertia
        If iteration > 2
            X = inertia [0] – inertia [1]
            U = inertia[iteration – 2] – inertia[iteration -1]
             If U < (X/10)
                Break
```

Fig.3 Enhanced unsupervised K-Means algorithm. Basic step by step approach for enhanced unsupervised K-Means.

7. EXPERIMENTAL RESULTS

We assessed the proposed algorithm using multiple real datasets and compared its performance with the traditional K-Means algorithm, specifically in terms of total execution time. Our experimental findings were recorded on a PC equipped with a 2.69 GHz CPU and 16 GB RAM.

The datasets used in our algorithm evaluation are described in the table 1.

Dataset	No. of entries	No. of attributes
Iris dataset	150	3
Glass dataset	215	9
Wine dataset	1144	12

Table.1 Datasets used. Describes iris, glass and wine datasets with number of entries 150, 215, 1144 and the number of attributes 3, 9, 12 respectively.

7.1 IRIS DATASET

It contains information about various characteristics of iris flowers, including measurements of sepal length, sepal width, petal length, and petal width, as well as the species of iris. The dataset is composed of 150 samples, with each sample categorized into one of three species: Setosa, Versicolor, or Virginica. The Iris dataset is widely recognized and utilized for classification and clustering tasks, serving as a foundational benchmark for evaluating the efficacy of various machine learning algorithms.

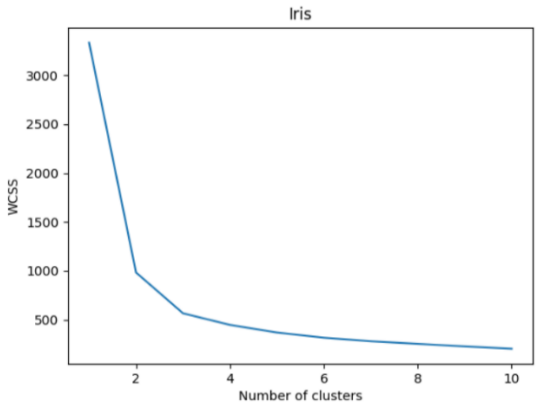

Fig. 5 Elbow plot for iris dataset. Elbow plot for iris dataset is plotted and the optimal clusters is found to be 3.

Fig. 5 shows the elbow plot for the iris dataset, in which the clusters is found to be 3 (K=3).

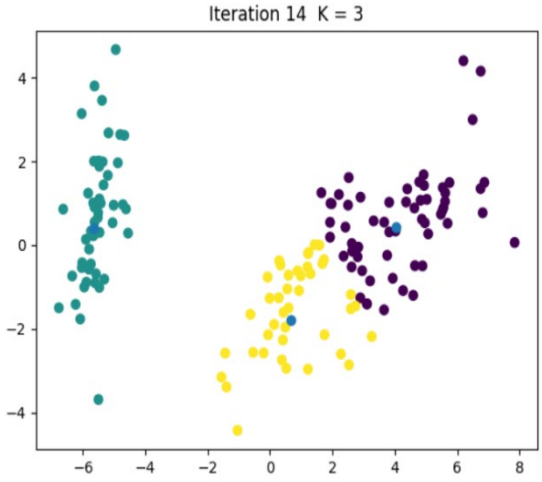

Fig. 6 Iris clustering. A graph plot representing the iris dataset classified into 3 distinct clusters.

Fig. 6 shows that enhanced unsupervised K-Means clustering algorithm finds the optimal count of clusters precisely.

7.2 GLASS DATASET

The Glass dataset is a well-known benchmark dataset frequently used in machine learning and statistical analysis. It contains information on various types of glass, each labelled with its respective class.

This dataset is often employed for classification tasks, where the objective is to predict the type of glass based on its attributes. Attributes typically include measurements such as refractive index and concentrations of different chemical elements.

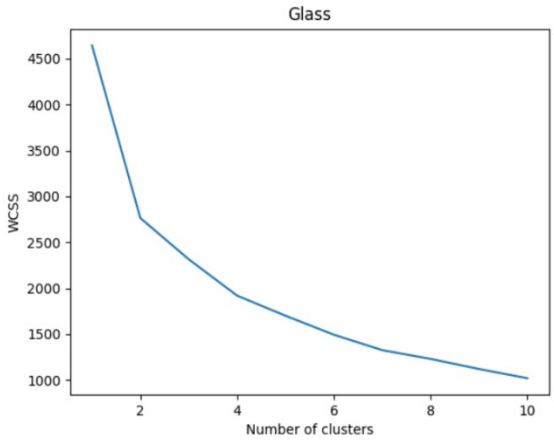

Fig. 7 Elbow plot for glass dataset. Elbow plot for glass dataset is plotted and the optimal clusters is found to be 4.

Fig. 7 shows the elbow plot for the glass dataset, in which the clusters is found to be 4 (K=4).

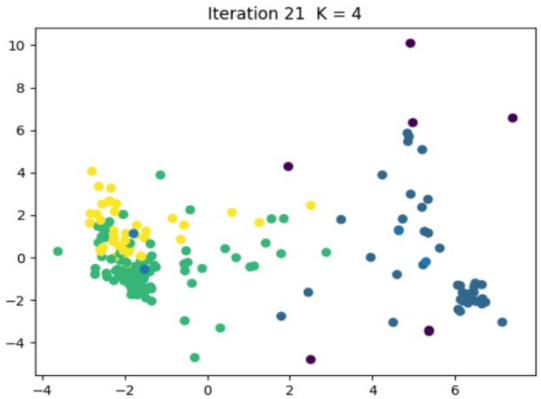

Fig. 8 Glass clustering. A graph plot representing the glass dataset classified into 4 distinct clusters.

Fig. 8 shows that enhanced unsupervised K-Means clustering algorithm finds the optimal count of clusters precisely.

7.3 WINE DATASET

It comprises measurements of various constituents found in wines, such as alcohol content, malic acid concentration, and ash content, along with the wine's class label, which represents one of three different cultivars.

The dataset contains a total of 178 samples, indeed, the structured nature of the wine dataset makes it highly suitable for supervised learning tasks, where the goal is to predict an outcome based on labelled input data like classification and regression.

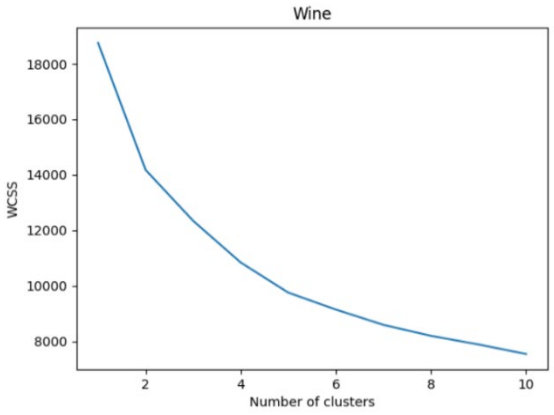

Fig. 9 Elbow plot for wine dataset. Elbow plot for wine dataset is plotted and the required clusters is found to be 5.

Fig. 9 shows the elbow plot for the wine dataset, in which the clusters is found to be 5 (K=5).

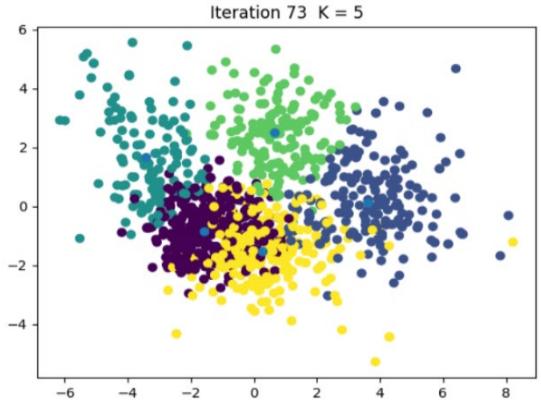

Fig. 10 Wine clustering. A graph plot representing the wine dataset classified into 5 distinct clusters.

Fig. 10 shows that enhanced unsupervised K-Means clustering algorithm finds the optimal count of clusters precisely.

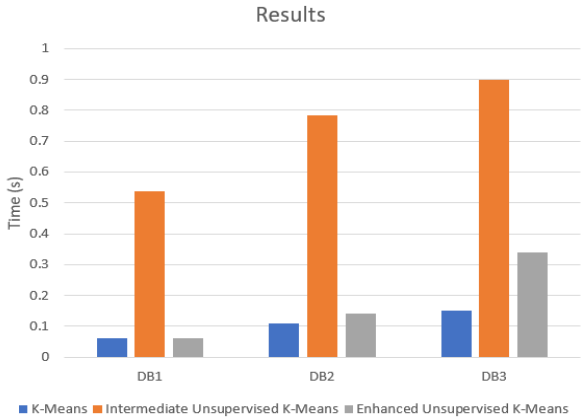

Fig. 11 Experimental results. A bar graph representing the time complexity comparison between K-Means, intermediate K-Means and enhanced K-Means algorithm, in which intermediate K-Means performs worst as expected while K-Means and enhanced K-Means performs similar to each other.

Fig 11 illustrates that the execution time of the enhanced unsupervised K-Means clustering algorithm is comparable to that of the traditional K-Means algorithm. However, there is a significant reduction in execution time compared to the intermediate version of the unsupervised K-Means algorithm. In Fig. 11, DB1 represents the Iris dataset, DB2 represents the Glass dataset and DB3 represents the Wine dataset.

8. CONCLUSION

In this paper, we presented an enhanced unsupervised K-Means clustering algorithm designed to automate the determination of the optimal clusters count. By incorporating the elbow method and employing optimization techniques, our algorithm presents a more streamlined and user-friendly approach to clustering analysis. Experimental findings illustrate the efficacy of our approach in terms of both clustering performance and computational efficiency.
Potential avenues for future research could encompass further optimization of the algorithm by considering the initial positions of centroids [4], exploring parallel processing implementations [5], and applying the algorithm to specific domains and real-world datasets. Additionally, investigating alternative clustering techniques and exploring integration with other machine learning algorithms could broaden the capabilities and applicability of our approach.

DATAAVAILABILITY STATEMENT

The iris dataset that supports the findings of this study are openly available in UCI Machine Learning repository at https://doi.org/10.24432/C56C76
The glass dataset that supports the findings of this study are openly available in UCI Machine Learning repository at https://doi.org/10.24432/C5WW2P
The wine dataset that supports the findings of this study are openly available in UCI Machine Learning repository at https://doi.org/10.24432/C56S3T

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Kristina P Sinaga, Miin Shen Yang. "Unsupervised k-means clustering algorithm." IEEE Access (2020).

Taher M Ghazal, Muhammad Zahid Hussain. "Performances of k-means clustering algorithm with different distant metrics." IASC (2021).

Fahim A M, Salem A M, Torkey F A, Ramadan M A. "An efficient enhanced k-means clustering algorithm." J Zhejian Univ SCIENCE A (2006).

Jianpeng Qi, Yanwei Yu, Lihong Wang, Jinglei Liu. "K*-means: An effective and efficient k-means clustering algorithm." IEEE International conferences on BDCloud, SocialCom, SustainCom (2016).

Giuliano Laccetti, Marco Lapegna, et al. "Performance enhancement of a dynamic k-means clustering algorithm through a parallel adaptive strategy on multicore CPUs." Journal of Parallel and Distributed Computing (2020).

A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil et al., "A survey of clustering algorithms for big data: taxonomy and empirical analysis." IEEE Transactions on Emerging Topics in Computing. (2014).

Hastie, T., Tibshirani, R., & Friedman, J. "The elements of statistical learning: data mining, inference, and prediction" (2nd ed.). Springer.(2009).