Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.102. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# A NEW METHOD FOR GETTING THE OPTIMAL NUMBER OF CLUSTERS BY K-MEANS USING THE WEIGHTED BARYCENTER

**JEDDIN SARA, BENTALEB YOUSSEF**

Engineering Sciences Laboratory of National School of Applied Sciences Ibn Tofail University, Morocco

Engineering Sciences Laboratory of National School of Applied Sciences Ibn Tofail University,Morocco
E-mail: sara.jeddin@uit.ac.ma, youssef.Bentaleb@uit.ac.ma

**ABSTRACT**

Clustering is a popular unsupervised algorithm in data science used to group similar data points together. One of the major challenges of using clustering algorithms is to determine the optimal number of clusters. To achieve this step, the Elbow method is a commonly used technique to identify the optimal number of clusters and often used in conjunction with K-means algorithm. However this method has some limitations and disadvantages, it is based on minimizing the sum of squared distances between each data point and the centroid of its assigned cluster that's why it provides information about the homogeneity inside clusters but it can't provide information about how is the distance between clusters. This paper suggests an enhanced Elbow algorithm that utilizes the concept of weighted barycenter to address the issue of group separation. The improved method is based on calculating the distance between the barycenter of the clusters identified by the K-means.

**Keywords:** *K-means, Elbow, Barycenter.*

## 1. INTRODUCTION

In the field of unsupervised algorithm, determining the optimal number of clusters (k) has always been a critical challenge in clustering algorithm [1, 2]. This question using the K-means algorithm has been addressed by several researchers from the early days of cluster analysis with many solutions that have been proposed, S. Davies and D. Bouldin (1979) proposed Davies-Bouldin validity test which can be used to measure the quality of clusters in K-Means. and help to determine the optimal number of clusters, Jörg Sander(1987) contributed to research on the K-Means algorithm and its aspects related to determining the optimal number of clusters, Anil K. Jain and Richard C. Dubes (1988) introduces a variety of techniques and methodologies for identifying the most suitable number of clusters in K-Means including methods based on internal and external indices [1], Tibor Cserháti (2000) proposed a method based on the ratio between inter-cluster and intra-cluster distances to determine the ideal number of clusters for the K-means algorithm.

While these methods have contributed significantly to the understanding of determining the optimal number of clusters using K-Means, there are notable limitations, the choice of a method should consider the specific characteristics of the dataset, The Davies-Bouldin Validity Test(1979) offers a quantitative measure of cluster quality in K-Means but is highly sensitive to noise and outliers, and it is subject to the interpretation of the potentially ambiguous "optimal" index value. Tibor Cserháti's Method (2000) focuses on the ratio between inter-cluster and intra-cluster distances, providing an intuitive approach, but it is sensitive to outliers and may not perform well with non-globular clusters, and assumes clusters have comparable shapes and sizes.

Anil K. Jain and Richard C. Dubes (1988) introduce a comprehensive range of techniques, including both internal and external indices, while acknowledging that some methods within this framework may be sensitive to data distribution, posing challenges in selecting the most suitable index for a given dataset.

The critiques of these methods include potential subjectivity in interpretation, sensitivity to noise and outliers, and absence insights into the homogeneity within clusters, and the absence of a universally accepted method, highlighting the need for a more adaptable and standardized approach in determining the optimal number of clusters.

The K-means algorithm represents a widely adopted technique in the field of machine learning that requires the specification of the number of clusters (k) to be generated for a given dataset [2, 4, 5]. The main objective in defining clusters is to minimize the total intra-cluster variation, often referred to the total within-cluster sum of squares (WSS) [2, 4, 5]. The Elbow method is a commonly used technique to determine the optimal number of clusters in K-means algorithm for a given dataset [4, 5, 6]. While it is a useful method, it has also some limitations and disadvantages. The fundamental concept underlying the Elbow method is to create a plot that illustrates the relationship between the within-cluster sum of squares (WCSS) and the number of clusters [6, 7]. The WCSS quantifies the dispersion of data points within each cluster and is computed as the sum of the squared distances between each data point and the centroid of the cluster to which it belongs [12, 13]. This approach offers insights into the homogeneity within clusters, but it fails to provide information about the distances between clusters. Within this paper, we suggest an enhancement to the Elbow method for determining the optimal number of clusters in K- means using the notion of the weighted barycenter approach.

The proposed method introduces the concept of weighting or ponderation which reflects the contribution of the data point "i" in relation to the cluster "k". As a result, by taking into account the notion of weighting, we are able to obtain a more nuanced understanding of the role and significance of individual data points in shaping the composition and structure of their respective clusters. Thus, if a point has a low contribution within a cluster, it means that removing this point from the cluster would not result in a significant displacement of the cluster's barycenter. In this paper, we will present practical examples to showcase the effectiveness of the proposed algorithm, the purpose of these examples is to demonstrate and assess the effectiveness of the method we have introduced. We will utilize public benchmark dataset and a simulated dataset with predefined and known class numbers (K) to compare our method with the improved elbow method.

## 2. THE UNSUPERVISED K-MEANS CLUSTERING ALGORITHM

### 2.1 Related conception

K-means algorithm is an iterative method wish is usually used in data mining and clustering problems [1]. This algorithm is an unsupervised method that separate a given data into K disjoint and predefined subgroups (clusters) and where each data point belongs only to one group [1, 2]. The K-means clustering algorithm is based on calculating in each iteration the distance between each data point and all the data centers called later by centroid. [1, 2]. In this section, we will define the distance function and the notion of centroid.

#### 2.1.1 General definition of a distance

We consider E a finite set with n individuals. For each of them, we have p values that we will note $\mathbf{x} = (x_{i1}, \ldots, x_{ip})$.

We define the distance formally as a metric or function [8]:

$$d(i,j) = d(j,i) = 0$$

if only i and j are not disjoint

$$d(i,j) \leq d(i,i') + d(i',j)$$

#### 2.1.2 Weighted euclidean distance

$$d^2(i,j) = \sum_{k=1}^{p} w_k (x_{ik} - x_{jk})^2 \quad (1)$$

where $w_k$ are predetermined weights.

### 2.1.3 Concept of the centroid

The centroid is a real or fictitious individual that represents a center of gravity around which several other individuals gather. It catches elements that resemble it, that is, individuals sharing similar characteristics [9].

Generally, the centroid of a population of individuals or a fraction of individuals is determined by aggregating the coordinates for each individual according to a specific aggregation function [9]. One of the most common aggregation functions used to calculate a centroid is the arithmetic mean.

Let E a set of individuals i of a size n, for each of them, we have p values that we will note $x = (x_{i1}, \ldots, x_{ip})$. The centroid or mean individual of this set of individuals is defined by the coordinates $c = (c_{i1}, \ldots, c_{ip})$ as:

$$c_{ik} = \frac{1}{n}\sum_{i=1}^{n} x_{ik} \quad k \in \{1, \ldots, p\} \quad (2)$$

### 2.1.4 Distance between an individual and a centroid

Suppose that R is a subgroup of E. When implementing the classification method K-means, it is necessary to be able to measure the resemblance of an individual i and a centroid c from the group of individuals R

The centroid c relative to the group R and a point j ∈ E are the most similar if their distance is the smallest.

### 2.2 Algorithm Description [5, 12]

Throughout the algorithm's progression, it's essential to update the elements of the E into their appropriate K-classes.

- **Step 0**

We specify a number of clusters K

- **Step 1**

We have n individuals to classify in K partition, we choose randomly K individuals representing the initial centroids.

- **Step 2**

We calculate the matrix of distances between the initial centroid K and all the n individuals. We aggregate each individual among the n − k elements with its nearest centroid, It results a partition of the set E in K classes of individuals.

- **Step 3**

The matrix of distances between the obtained centroid K and all n individuals is calculated and every individual data point is assigned to the cluster with the nearest centroid. This results in a new updated partition of the E set in K classes of individuals.

- **Step 4**

We keep iterating until there is no change to the centroids and the assignment of data points to clusters isn't changing.

- **Final step**

We finally get the K-classes when they become constant from one iteration to another.

### 2.3 The K-means objective function [5, 12]

The objective function of K-means algorithm is:

$$\sum_{k=1}^{K}\sum_{i \in k} \|x_{i,k} - C_k\|^2 \quad (3)$$

where: $\|x_{i,k} - C_k\|^2$ is the euclidean distance between the data point xi and the cluster (centroid) $C_k$. The K-means algorithm is iterated in order to minimize the K-means objective function. An essential step for the K-means algorithm is to determine the optimal number of clusters for data clustering [4], and the Elbow Method is a popular tool for determining this optimal value of k [6, 7].

## 3. NEW METHOD FOR GETTING THE OPTIMAL NUMBER OF CLUSTERS BY K-MEANS USING WEIGHTED BARYCENTER

### 3.1 Elbow curve method

It is necessary to identify the optimal number of clusters in any unsupervised algorithm into which the data may be clustered. The Elbow Method is a widely accepted approach for determining the optimal value of k [6, 13].We showcase its application using the K-Means clustering.

### 3.1.1 Method description

Elbow Curve Method runs k-means clustering on the dataset for a range of k values [13], typically from 1 to 10.

- **Step 0**

A number of clusters K is specified.

- **Step 1**

We execute K-means clustering for each of these different K values.

- **Step 2**

For each K value, we compute the average distances between data points and their respective centroids.

- **Step 3**

We choose the optimal value of K when the average distances value falls suddenly or we plot these data points and identify the "Elbow" point where the average distance from the centroid experiences a sudden decrease.

### 3.1.2 Criticize Elbow method

There is, however a difficult problem with the Elbow method which is based on calculating the distance between each data point and its closest centroid [7]. That's why it provides information about the homogeneity inside clusters but it can't provide information about how is the distance between clusters as shown in Fig.1 wish shows that the classes are well separated using the classic method. In practice, we not only want a separated elements in the same cluster but also we need an information about (If the clusters are very well separated or not). To address the limitations of the Elbow method, we introduce a new approach wish determines the optimal cluster number taking into consideration the separation between clusters. This method, unlike the existing method consider that the clusters in Fig.1 are not very well separated.

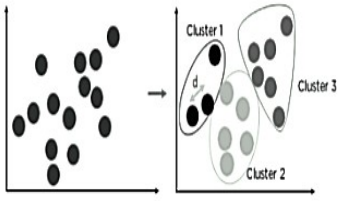


*Figure 1: Experimental results of clustered data set with 3 clusters*

### 3.2 Proposed method

### 3.2.1 Principle

We are proposing an enhancement to the Elbow method that aims to address the challenge of group separation. This improvement involves the introduction of a weighted barycenter concept. To implement this enhancement, we must first calculate the distances between two groups identified by the K-means method.

We start with a dataset consisting of n individuals:

$\Omega = \{w_1, \ldots, w_n\}$ randomly selected from a population.

Each individual is associated with p values for p variables, denoted as $X_1, \ldots, X_p$.

Assuming a partition into K groups with K centers of gravity $(u_1, \ldots, u_k)$, we can decompose the total variability of the data points as follows [11]:

$$I_T = \sum_{k=1}^{n} \left\{ \sum_{i \in k} d^2(x_i), m_k \right\} + \sum_{k=1}^{K} n_k \, d^2(M_k, x_G) \quad (4)$$

$$I_T = I_W + I_B \quad (5)$$

With:

"• $I_I$ (total variance) signifying the total variability of the point cloud: which remains constant for a given dataset.

• $I_W$ (within variance) signifying the dispersion of points around their respective centers.

• $I_b$ (between variance) indicating the separability of groups, a factor we aim to maximize.

• d is a predefined distance metric."[12]

Groups A and B constitute subsets of observations within the dataset.

Our aim is to minimize the increase in intra-group Variance during group merging, as defined by the provided formula [11]:

$$I_W(A, B) = \sum_{i \in A} d^2(i, \mu_A) + \sum_{i \in B} d^2(i, \mu_B) \quad (6)$$

$$I_W(A \cup B) = \sum_{i \in A \cup B} d^2(i, \mu_{AB}) \quad (7)$$

For the K-means algorithm the objective function of algorithm is:

$$\sum_{k=1}^{K} \sum_{i \in k} \left\| x_{i,k} - C_k \right\|^2 \quad (8)$$

Where: $\left\|x_{i,k}-C_k\right\|^2$ is the euclidean distance between the data point $x_i$ and the cluster (centroid) $C_k$. The K-means algorithm is iterated in order to minimize the K-means objective function.

The new method proposes to calculate the distance between clusters instead of calculating the distance between each point and it closest centroid.

• $(p_{j,k})$ represents the weight assigned to variable Xj in association with observation k relating to an observation k;

• The point $G_A$ represents the weighted center of gravity within a specific class, denoted as A

•Let $\overline{X_j}$ represent the weighted centroid of the coordinate set $(\overline{X_1},\ldots\ldots,\overline{X_p})$:

$$\overline{x_j}=\frac{1}{n}\frac{\sum_{k=1}^{p}p_{j,k}x_{j,k}}{\sum_{k=1}^{p}p_{j,k}} \quad (9)$$

$$with: \sum_{k=1}^{p}p_{j,k}\neq 0$$

### 3.2.2 Measuring the distance between classes A and B

"Measuring the distance between two classes, A and B, is accomplished by calculating the distances between the weighted barycenter of clusters, outlined as follows"[12]:

$$d(A,B)=d(G_A,G_B)=d(\bar{x}_A,\bar{x}_B) \quad (10)$$

Where:

$$\bar{x}_A=\frac{1}{n_A}\frac{\sum_{1_{j\in A}}p_{j,k}x_j}{\sum_{j\in A}p_{j,k}} \quad (11)$$

and

$$\bar{x}_B=\frac{1}{n_B}\frac{\sum_{1_{j\in B}}p_{j,k'}x_j}{\sum_{j\in B}p_{j,k'}} \quad (12)$$

### 3.2.3 Euclidean distance case

Let's denote A and B as two classes or elements of a given score.

In this context, we assume that the data is represented as an (n x n) matrix of Euclidean distances between pairs of individuals.

• Include $(p_{j,k})$ and $(p_{j,k'})$ the weights for the two classes A and B

• $d_{i,j}$ the distance between any two individuals i and j.

"The square of the distance between two barycenter is computed based on the matrix of distances between individuals pairwise."[12]

$$d^2(g_A,g_B)=\frac{1}{\sum_{i\in A,j\in B}p_{j,k}p_{j,k'}}\sum_{i\in A,j\in B}p_{j,k}p_{j,k'}d_{i,j}^2 \quad (13)$$

We define $p_{j,k}$ as predetermined weights and $d_{i,c_k}^2$ is the distance between i and its closest centroid.

$p_{j,k}$ is the contribution of the point i in a given cluster k where:

$$p_{i,k}=\frac{1}{1+d_{i,c_k}^2} \quad (*)$$

* To avoid dividing by 0 when a point i is the barycenter of the cluster k, so in this case the distance between them is 0.

Then the new-inertia is:

$$I_B=\sum_{k=1}^{K}\sum_{l=1}^{K}\frac{1}{\sum_{j\in k}\sum_{j\in l}p_{i,k}p_{j,l}}\sum_{j\in k}\sum_{i\in l}p_{i,k}p_{j,l}d_{i,j}^2 \quad (15)$$

• where $d_{i,j}^2$ is the squared distance between i and j

• K is the total number of clusters

• $I_B$ (between variance) indicates the degree of separability among groups, with the goal of maximizing it

Unlike the existing method, our problem here is a maximization problem and thus, we need to modify our approach.

Then we look for optimal value of k which maximize the new formula of inertia.

$p_{i,k}$ is the contribution of the point i in a given cluster k, a low value of $p_{i,k}$ means a low contribution of a point i in cluster k.

When we have a data point with a low contribution in a cluster, deleting it from this cluster will not move the barycenter significantly. This can only achieve if the distance between the point i and the barycenter of the cluster k is high.

### 3.3 Results and discussion

In this section, we provide illustrative examples using numerical and real datasets to showcase the effectiveness of the proposed algorithm in action. To demonstrate and confirm the performance of our proposed method in this section, we provide illustrative examples using public and real benchmark datasets and a simulated datasets with predetermined and known class numbers (K) in

order to compare this method with the classic elbow method.

We will implement the proposed method using python, Anaconda, and the sklearn on 10 datasets. The complete source code of our algorithm is available on GitHub at the following link: Link-github-Sara-Jeddin.

This GitHub repository contains all the necessary resources to reproduce our results and run our algorithm on other datasets.

### 3.3.1 Example 1

We apply K-means algorithm in iris datasets which contains 3 groups.
Let us determine the optimal number using the proposed method.

To identify the optimal value of k we choose the point at the"elbow" point, which is where the inertia starts increasing linearly. Therefore, based on the given data.

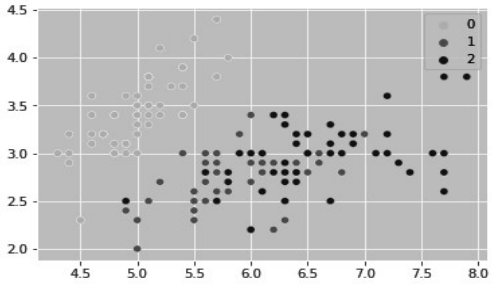


*Figure 2: Iris data*

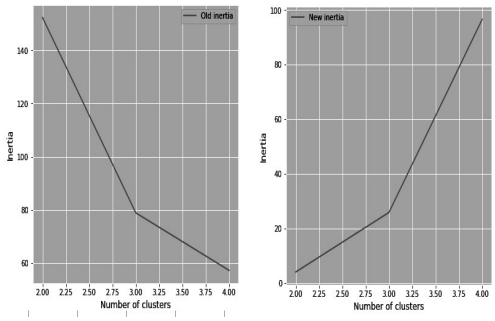


*Figure 3: The classic elbow method and the new method using the two inertia formulas*

We can conclude that the optimal number of clusters for the data is 3.
By obtaining the same predetermined number of classes, we can verify the credibility of the proposed method. Furthermore, we aim to go beyond and determine an optimal number of clusters that takes into consideration the separation between the clusters. This approach ensures that the clustering solution not only aligns with the predefined class labels but also maximizes the distinctiveness and separability among the identified clusters.

### 3.3.2 Example 2

We create a new dataset using make blobs function from
scikit-learn which have 2 columns and the points are separated into 3 groups

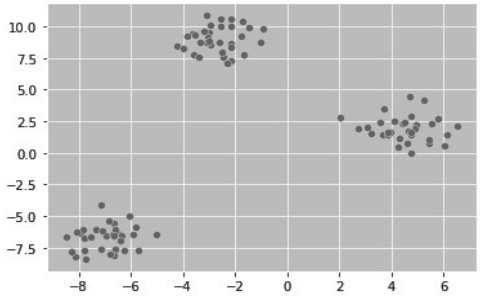

*Figure 4: A created dataset with 3 groups*

Using the same approach, we will determine the optimal number of clusters using both methods.

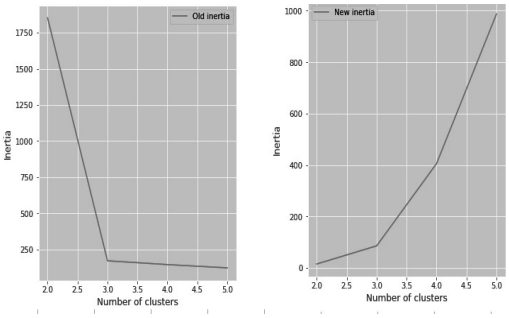


*Figure 5: The classic elbow method and the new method using the two inertia Formulas*

Both methods (using the old and new formula for inertia) suggest that k = 3 is the optimal number of clusters for K-means.

### 3.3.3 Example 3

We generate a new dataset with 3 columns and 3 groups using make blobs function from scikitlearn

Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.102. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

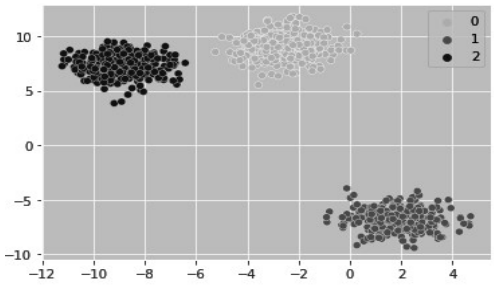


*Figure 6: A created dataset with 3 groups*

Using the same approach, we will determine the optimal number of clusters using both methods. Both methods

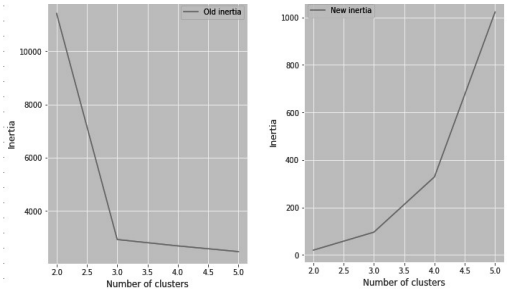


*Figure 7: The classic elbow method and the new method using the two inertia formulas*

(using the old and new formula for inertia) suggest that k = 3 is the optimal number of clusters for K-means.

### 3.3.4 Example 4

We generate a new dataset with 3 columns and 5 groups. we can note that there are 3 groups clusets to each other.

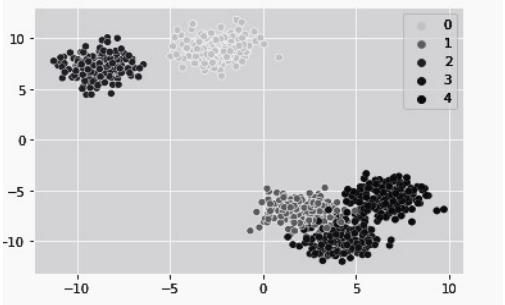


*Figure 8: Created dataset with 5 groups*

Utilizing the identical approach, we will determine the optimal number of clusters by employing both methods.

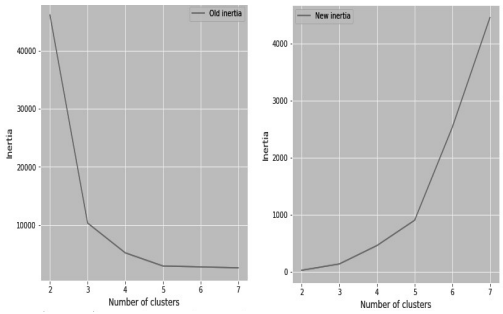


*Figure 9: The classic elbow method and the new method using the two inertia formulas.*

k = 3 is the optimal number of clusters using the classic elbow but in the other hand 5 is the optimal based on the new formula.

### 3.3.5 Example 5

We generate a new dataset with 3 columns and 6 groups. we note there are 3 groups clusters to each other.

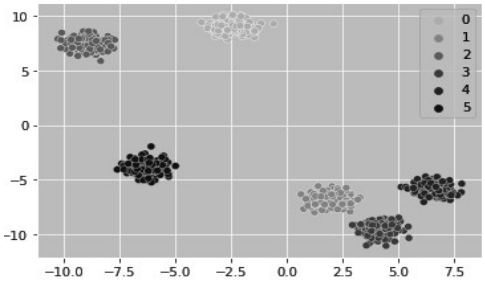


*Figure 10: Generated dataset with 3 columns and 6 groups*

By the same approach, we will determine the optimal number of clusters by employing both methods.

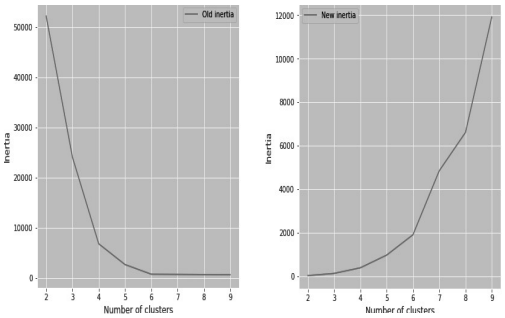


*Figure 11: The classic elbow method and the new method using the two inertia formulas*

k = 4 is the optimal k for K-means but in the other hand k = 6 is the optimal based on the new formula.

### 3.3.6 Example 6

Journal of Theoretical and Applied Information Technology
15th January 2024. Vol.102. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

We generate a dataset with seven columns and seven groups, with some groups exhibiting close proximity to each other.

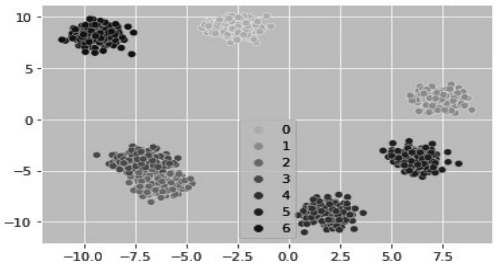


*Figure 12: Generated dataset with 7 columns and 7 groups*

By employing both methods, we will utilize the same approach to define the optimal number of clusters.

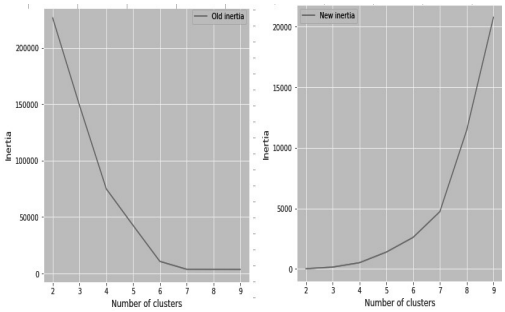


*Figure 13: The classic elbow method and the new method using the two inertia formulas*

k = 6 is the optimal k for K-means but in the other hand k = 7 is the optimal based on the new formula.

### 3.3.7 Example 7

We generate a new dataset with 6 columns and 7 groups. We note there are some groups that are close to each other.

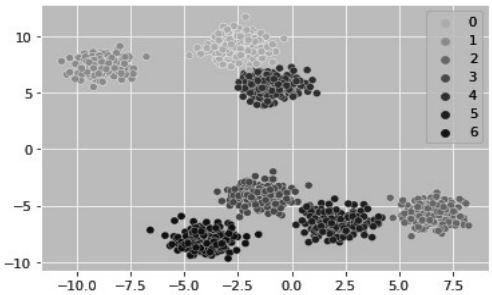


*Figure 14: Generated dataset with 6 columns and 7 groups*

By employing both methods, we will utilize the same approach to define the optimal number of clusters. k = 6 s the optimal k for K-means but in the other hand k = 7
is the optimal based on the new formula.

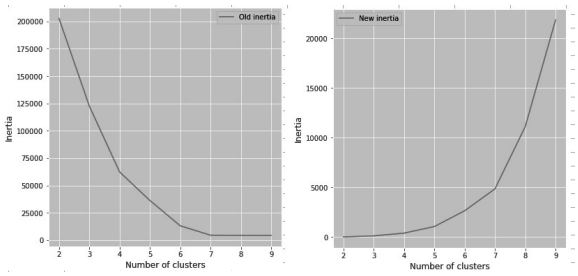


*Figure 15: The classic elbow method and the new method using the two inertia formulas*

### 3.3.8 Example 8

We generate a new dataset with 6 columns and 5 groups.

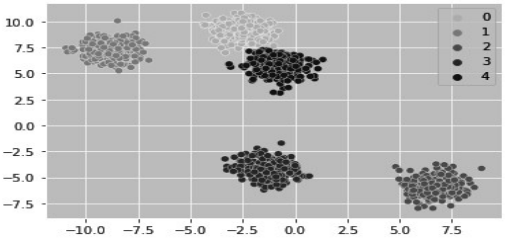


*Figure 16: Generated dataset with 6 columns and 5 groups*

We note there are some groups that are close to each other.
Using both methods, we will utilize the same approach to define the optimal number of clusters

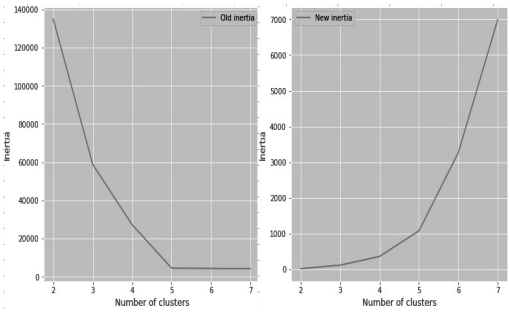


*Figure 17: The classic elbow method and the new method using the two inertia formulas*

Both of them (old and new formula) consider that 5 is the optimal k for k-means

### 3.3.9 Example 9

We generate a new dataset with 6 columns and 4 groups. We note there are some groups that are cluster to each other.

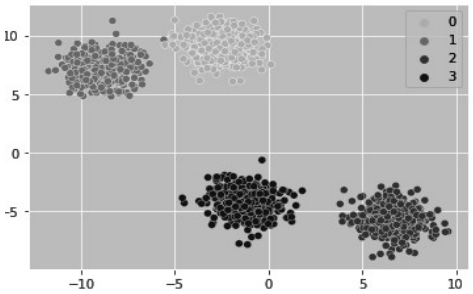


*Figure 18: Generated dataset with 6 columns and 4 groups*

Using both methods, we will utilize the same approach to define the optimal number of clusters.

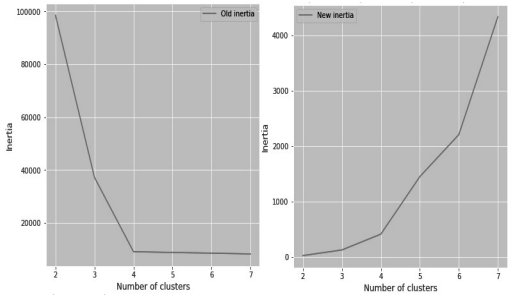


*Figure 19: The classic elbow method and the new method using the two inertia formulas*

Both of them (old and new formula) consider that 5 is the optimal k for k-means

### 3.3.10 Example 10

We note there are some groups that are cluset to each other

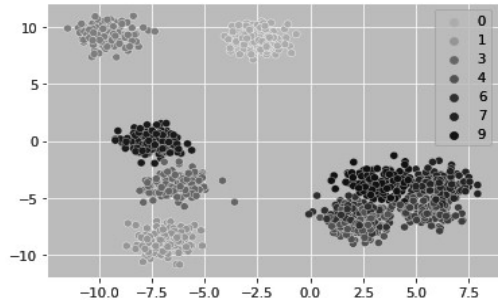


*Figure 20: Generate dataset with 3 columns and 10 groups*

Using both methods, we will utilize the same approach to define the optimal number of clusters optimal k for old formula is 8 but for the new is 10.

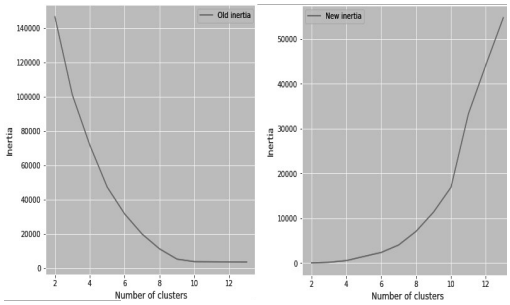


*Figure 21: The classic elbow method and the new method using the two inertia formulas*

## 4. CONCLUSION

Clustering analysis method is one of the main analytical methods in data mining; this paper discusses the standard Elbow method and analyzes the shortcomings of this algorithm. We propose an improved Elbow algorithm for determining the optimal number of clusters in the K- means using a weighted barycenter approach. The method was developed to address the limitations of existing technique. The proposed method uses to determine the optimal number of clusters a new criterion. Experimental results on various datasets demonstrated that the proposed method outperformed the exiting method.

Future research directions for this topic include exploring variations and adaptations of the proposed algorithm to assess its robustness in specific contexts and its performance on large-scale datasets and diverse applications. Further investigation into the theoretical foundations and advanced mathematical aspects of the algorithm is warranted. Additionally, the impact of different similarity metrics on the algorithm's results should be studied, and optimization methods for adjusting parameters in specific scenarios should be examined.

## REFERENCES:

[1] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall, Inc.

[2] Tan, P. N., Steinbach, M., Kumar, V. [2005]. Introduction to Data Mining. Addison-Wesley

[3] Hartigan, J. A., Wong, M. A. (1979). Algorithm AS 136: A K-means Clustering Algorithm. *Journal of the Royal Statistical Society*: Series C (Applied Statistics), 28(1), 100-108.

[4] MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), 281-297.

[5] Thirumalai, V., Lakshmi, M. S. (2018). Clustering Based Techniques for Finding Optimal Number of Clusters. In Advances in Machine Learning and Data Science (pp. 401-412). Springer.

[6] Waghmare, V. S., Joshi, S. S. (2016). An Improved Method for Determination of Number of Clusters in K-means Clustering. *International Journal of Computer Science and Mobile Computing* 5(11), 248-256.

[7] Balakrishnan, P., Dhanalakshmi, R. (2018). Analysis of Elbow Method and Silhouette Method for Determining the Optimal Number of Clusters. *International Journal of Engineering Technology*, 7(4.19), 325-330.

[8] Huang, S., Avila-Garcia, M. S., Goulermas, J. Y. (2020). Improving the Elbow Method to Better Estimate the Number of Clusters. Expert Systems with Applications, 153, 113447.

[9] ESCOFIER, B. ET J. PAGERS [1990], Analyses factorials simples et multiples: objectives, méthodes et interrostation, 2é ed., Paris, Dunod.

[10] HOTELLING, H. [1933], Analysis of a Complex of Statistical Variables into Principal Components, *Journal of Educational Psychology*, 24,417-441.

[11] PEARSON, K. [1901], On Lines and Planes of Closest Fit to Systems of Points in Space, Philosophical Magazine, 2, 559-572.

[12] JEDDIN Sara and BENTALEB Youssef [2022], Hierarchical Classification Method Based on Weighted Barycenter to Resolve the Problem of Group Separation, The *Proceedings of the International Conference on Smart City Applications*, 853–858.

[13] Inaga KP, Yang MS [2020], Unsupervised K-means clustering algorithm, IEEE access, 80716-80727.

[14] Marutho D, Handaka SH, Wijaya E [2018], The determination of cluster number at K-means using elbow method and purity evaluation on headline news, *International seminar on application for technology of information and communication*,533-538.

[15] Shi C, Wei B, Wei S, Wang W, Liu H, Liu J. A [2021], quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm, EURASIP *Journal on Wireless Communications and Networking*,1-6.

[16] M A Syakur et al [2018], quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm, EIOP Conf. Ser.: Mater. Sci. Eng,1-6. : 2018 . 336012017

[17] Tan, P. N., Steinbach, M., Kumar, V. [2005]. Introduction to Data Mining. Addison-Wesley A. K.