International Journal of Advanced Statistics and Probability, 13 (1) (2026) 32-41

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP
https://doi.org/10.14419/69dvcw11
Research paper

An Improved Initialization Method for k-Means Clustering of Noisy Datasets Based on Rough Set Neighbourhood Model

Abeng J. Abeng, Mbanefo S. Madukaife *

Department of Statistics, University of Nigeria, Nsukka, Nigeria
**Corresponding author E-mail: mbanefo.madukaife@unn.edu.ng*

Received: January 19, 2026, Accepted: March 20, 2026, Published: March 29, 2026

Abstract

This study improves one of the initialization methods for the k-means clustering algorithm based on a rough set neighbourhood model to enhance performance in noisy datasets. The method involves data normalization, obtaining a neighbourhood threshold based on the 0.25th trimmed mean of pairwise Minkowski distances, calculating cohesion and coupling degrees of the neighbourhoods and between them respectively, and obtaining the initial cluster centres as the k points having maximum cohesion degrees with minimum coupling degrees among themselves. The approach was evaluated on six datasets using Silhouette, Davies–Bouldin, Calinski–Harabasz, and Dunn–Hubert indices in comparison with an existing method. Results showed that the improved method outperformed the existing method on noisy datasets, achieving higher Silhouette and Dunn–Hubert scores, and lower Davies–Bouldin values, with a slight reduction in Calinski–Harabasz index in one of the datasets. On the non-noisy datasets, the two methods were at par in all four performance indices. With the improved performance, showing that the improved method enhanced the stability and robustness of k-means clustering in the presence of noisy data, it can be recommended for clustering noisy datasets such as gene expression, image, and signal datasets.

Keywords: *Initialization Methods; K-Means Clustering; Neighbourhood Model; Noisy Dataset; Rough Set Theory.*

1. Introduction

Clustering is a fundamental technique in multivariate data analysis, widely applied in domains such as image processing, market segmentation, bioinformatics, and social network analysis ([1]). Suppose a set of n individuals (or objects) is obtained from k different populations, $k = 2, 3, \ldots$, such that each individual has associated measurements of a p-component random vector $\mathbf{x} \in R^p$, where each individual is represented by its vector of measurements $\mathbf{x}_i$, $i = 1, 2, ..., n$. The basic data for most applications of clustering is the usual n x p multivariate data matrix, $\mathbf{X}$, containing the variable values describing each object to be clustered. That is,

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \text{M} \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \text{L} & x_{1p} \\ x_{21} & x_{22} & \text{L} & x_{2p} \\ \text{M} & \text{M} & \text{M} & \text{M} \\ x_{n1} & x_{n2} & \text{L} & x_{np} \end{pmatrix} \tag{1}$$

The entry x_{ij} in $\mathbf{X}$ gives the value of the i^{th} object under j^{th} variable, where $i = 1, 2, \ldots, n$; and $j = 1, 2, \ldots, p$. A typical approach is to learn a set of cluster centres $\mu_1, \mu_2, \ldots, \mu_k$; $\mu_i \in R^p$ such that the sum of squared distances between the points $\mathbf{x}_i$ and their nearest cluster centres is minimized. It is formally represented as:

$$J = \sum_{i=1}^{k} \sum_{\mathbf{x}_{ij} \in C_i} \left\| \mathbf{x}_{ij} - \mu_i \right\|^2, \tag{2}$$

where C_i is the set of points in i^{th} cluster, $\mathbf{x}_{ij}$ is the j^{th} data point in cluster i, and μ_i is the centroid of the i^{th} cluster ([2]). This optimization aims to find the best-fitting cluster centres by minimizing intra-cluster variances. The primary objective of clustering is to partition a given dataset into k distinct clusters or groups, where each cluster contains data points that are more similar to one another than to those in other clusters ([3]). The similarity or dissimilarity between points is usually measured using distance metrics such as Euclidean distance for continuous data or Jaccard similarity for categorical data ([4]).

There are a good number of different methods of clustering in the literature such as the partitional, hierarchical and model-based approaches. Among the partitional approaches and the entire clustering methods in general, the *k*-means algorithm, introduced by MacQueen [5], is arguably one of the most popular due to its simplicity and efficiency. This method partitions data into *k* clusters by minimizing within-cluster variance and the distortion function, which is given by

$$\sum_{j=1}^{n} \min_{i=1}^{k} \left\| \mathbf{x}_{ij} - \mu_i \right\|^2, \tag{3}$$

where $\mathbf{x}_{ij}$ are data points, μ_i are centroids, n is the number of points, and k is the number of clusters. It has been noted to be highly sensitive to the choice of initial centroids, and as a result, poor initialization can lead to convergence to local minima, unstable clustering results, and degraded performance, particularly in datasets characterized by noise or irregular distributions ([6], [7], [1]). This challenge is particularly critical in large and complex datasets where inappropriate centroid initialization can lead to misleading interpretations. To address this challenge, several initialization methods have been proposed. The base method is the random initialization method which randomly selects *k* distinct points from the dataset to be clustered as the initial centres. Unfortunately, the random initialization method received dozens of criticisms. For instance, Yedla et al. [8] noted that it leads to convergence at local minima and inconsistent results across runs. This leads to proposing of several other methods of initialization of cluster centres. For instance, Arthur and Vassilvitskii [9] proposed the *k*–means ++ algorithm, which no doubt, improved the clustering quality of the *k*–means method. Also, Cao et al. [10] proposed a deterministic method to enhance clustering quality of the *k*–means clustering using the neighbourhood model based on the rough set theory according to Pawlak [11]. The proposed model was seen to be better than the *k*–means ++ in terms of both internal and external performance indices. Again, Yedla et al. [8] proposed another deterministic method to enhance accuracy and efficiency by calculating weighted distances for each data point. The weighted distance is defined by:

$$W_{d_i} = \sum_{l=1}^{m} \mathbf{x}_l . E_{d_i}, \tag{4}$$

where E_{d_i} is the Euclidean distance from the origin and $\mathbf{x}_l$ are attributes. These distances are sorted, and *k* centroids are chosen systematically at positions $(n(2j-1)+k)/(2k)$ ensuring a uniform n/k spacing. Other methods include Xu et al. [12], Ahmed and Ashour [13], Zahra et al. [14], Yang et al. [15], Mishra et al. [16], Chowdhury et al. [17], Yang et al. [18], Sujatha et al. [19], Gul and Abdul Rehman [20], Zu et al. [21], Zubair et al. [22], Gao et al. [23] and Zhang et al. [24], to mention but a few. Despite these advancements, finding an optimal initialization technique that guarantees high-quality clustering results across diverse types of data remains an ongoing challenge in multivariate analysis ([25]). On the other hand, a number of works have equally been devoted to comparing the initialization methods for *k*–means clustering, see for instance Meila and Heckerman [26] and Celebi et al. [27].

One of the initialization methods that has consistently shown good performance is the Cao et al. [10] method. Using rough set theory, introduced by Pawlak [11], this method obtains neighbourhoods by measuring the lower and upper approximation sets of each data point which is used to obtain the cohesion degree of points under the same neighbourhood and the coupling degrees of points between neighbourhoods. It therefore chooses initial cluster centres as the points within each neighbourhood with maximum cohesion degrees and at the same time having least coupling degrees across neighbourhoods. As good as this method, it was discovered that it is highly sensitive to noise. This paper therefore proposes an improvement to Cao et al. [10] which handles noise. The rest of the paper is presented as follows: Section 2 presents the basic concepts as preliminaries while Section 3 presents the improved initialization method for *k*–means clustering of noisy datasets. In Section 4, the improved algorithm is compared with the former and the paper is concluded in Section 5.

2. Preliminaries

In a general *k*–means clustering, three basic processes are involved, namely: data preprocessing, determination of initial cluster centres and clustering proper. The basic concepts are discussed in what follows in the light of the above.

2.1. Data preprocessing

Prior to clustering, all datasets are normalized to eliminate scale effects among variables. Given a dataset $\mathbf{X} = \{x_{ij}\}$, where x_{ij} denotes the value of the *j*th variable for the *i*th observation, the dataset can be normalized using different methods. One of the commonest method of data normalization is the min-max normalization which is given by:

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}, \tag{5}$$

where $\min(x_j)$ and $\max(x_j)$ represent the minimum and maximum values of the *j*th variable, respectively. This transformation ensures that all variables are rescaled to the interval $[0,1]$, thereby preventing attributes with large magnitudes from dominating distance computations.

Also, an important aspect of every clustering is computation of inter-point distances which is used for quantifying dissimilarities between multivariate data points. Such distance measures include the Minkowski and the Mahalanobis distances between observation vectors $\mathbf{x}^T = (x_1, x_2, \ldots, x_p)$, $\mathbf{y}^T = (y_1, y_2, \ldots, y_p)$ and $\mathbf{z}^T = (z_1, z_2, \ldots, z_p)$ in R^p. The Minkowski distance between any two multivariate data points $\mathbf{x}$ and $\mathbf{y}$, which is a generalized L_n–norm, is defined by:

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{p} |x_i - y_i|^s \right)^{1/s}, s \geq 1 \tag{6}$$

Clearly, d_p is a distance metric, meaning it satisfies the above properties for any objects $\mathbf{x}$ and $\mathbf{y}$ in the universe. The distance metric is called Manhattan, Euclidean and Chebyshev distances when s is 1, 2, and ∞, respectively.

2.2. Rough set neighbourhood

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, where $\mathbf{x}_j \in R^p$ form a data matrix X. For any $\mathbf{x}_i \in \mathbf{X}$, the neighborhood δ_p^ε of $\mathbf{x}_i$ with respect to a threshold, ε is defined as:

$$\delta_p^\varepsilon(\mathbf{x}_i) = \{\mathbf{x} \in \mathbf{X} | d_p(\mathbf{x}, \mathbf{x}_i) \le \varepsilon\} \tag{7}$$

It is called a rhombus region, a ball region and rectangle or square region around $\mathbf{x}_i$ for Manhattan, Euclidean and Chebyshev distances, respectively. The family of neighborhoods of an object $\mathbf{x}$, $\delta(\mathbf{x})$, covers the universe X rather than partitioning it. Thus, having the property that multiple neighborhoods may overlap, ensuring that every object in X belongs to at least one neighborhood.

2.3. Cohesion and coupling degrees

To quantify the representativeness of each observation within its neighbourhood, cohesion and coupling degrees are computed. The cohesion degree, often expressed as the approximation quality, quantifies how well a condition attributes approximate decision classes. For any $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$, the cohesion degree of $\delta_p^\varepsilon(\mathbf{x}_i)$ is defined as:

$$\text{cohesion}(\delta_p^\varepsilon(\mathbf{x}_i)) = \left| \frac{\underline{P}(\delta_p^\varepsilon(\mathbf{x}_i))}{\bar{P}(\delta_p^\varepsilon(\mathbf{x}_i))} \right|, \tag{8}$$

where $0 < \text{cohesion}(\delta_p^\varepsilon(\mathbf{x}_i)) \le 1$ and $\underline{P}(\delta_p^\varepsilon(\mathbf{x}_i))$ and $\bar{P}(\delta_p^\varepsilon(\mathbf{x}_i))$ are the lower and upper approximation of the neighbourhood, $\delta_p^\varepsilon(\mathbf{x}_i)$. The greater $\text{cohesion}(\delta_p^\varepsilon(\mathbf{x}_i))$ is, the less the boundary region of neighborhood of object $\mathbf{x}_i$ is, which means that $\mathbf{x}_i$ is a better cluster centre of its neighborhood. Therefore, $\mathbf{x}_i$ is likely taken as an initial cluster centre in X. On the other hand, the coupling degree measures the extent to which an observation shared neighbourhood elements with other observations. For any $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$, the coupling degree between $\delta_p^\varepsilon(\mathbf{x}_i)$ and $\delta_p^\varepsilon(\mathbf{x}_j)$ is defined as:

$$\text{coupling}(\delta_p^\varepsilon(\mathbf{x}_i), \delta_p^\varepsilon(\mathbf{x}_j)) = \left| \frac{(\delta_p^\varepsilon(\mathbf{x}_i) \bigcap \delta_p^\varepsilon(\mathbf{x}_j)}{(\delta_p^\varepsilon(\mathbf{x}_i) \bigcup \delta_p^\varepsilon(\mathbf{x}_j)} \right|, \tag{9}$$

where $0 < \text{coupling}(\delta_p^\varepsilon(\mathbf{x}_i), \delta_p^\varepsilon(\mathbf{x}_i)) \le 1$. The greater $\text{coupling}(\delta_p^\varepsilon(\mathbf{x}_i), \delta_p^\varepsilon(\mathbf{x}_j))$ is, the more possibly $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster. It has been noted that they are considered to belong to the same cluster if $\text{coupling}(\delta_p^\varepsilon(\mathbf{x}_i), \delta_p^\varepsilon(\mathbf{x}_i)) > \varepsilon$. On the contrary, $\mathbf{x}_i$ and $\mathbf{x}_j$ are likely taken as initial cluster centres if their individual cohesion degrees are high.

3. Improved Initialization Method for *k*–Means Clustering

Let *k* denote the desired number of clusters, initial centroids are selected by ranking observations based on descending cohesion degree and ascending coupling degree. The top *k* observations satisfying the neighbourhood threshold condition are chosen as the initial centroids for the k-means clustering algorithm. However, it is very well known that the neighbourhood of a data point, x_i, given by $\delta_p^\varepsilon(\mathbf{x}_i)$ is dependent on the threshold, ε. As a result, the cohesion and coupling degrees also depend on ε. Precisely, the higher the value of ε, the larger the neighbourhood, thereby increasing the number of candidate points for initial centres. Now, for a set of *n* observation vectors, there are $n(n–1)$ pairwise distances. Cao et al. [10] estimated the threshold based on these pairwise distances as:

$$\hat{\varepsilon} = \bar{x} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_p(\mathbf{x}_i, \mathbf{x}_j) \tag{10}$$

However, it is known that a noisy dataset will produce some of these pairwise distances as outliers. Including these outliers in the estimation of the threshold introduces unnecessary distortion of the threshold, leading to initial cluster centres that will produce unstable clusters. As a result, this paper circumvents this error by estimating the threshold as 0.25^{th} trimmed mean of the distances, so as to largely remove the pairwise distances impacted on with noise. Hence, ε is estimated by:

$$\hat{\varepsilon} = TM_{0.25}(d_p(\mathbf{x}_i, \mathbf{x}_j)); \forall i, j, \tag{11}$$

where $TM_{0.25}(\cdot)$ denotes the trimmed mean obtained by removing the lowest 25% and highest 25% of the distance values and quantile-based trimming ensures that neighbourhood formation is less sensitive to outliers.

Based on the new improved threshold estimation, the algorithm for the initialization problem is presented in detail as:

Step 1: Input: The *nxp* dataset, X, normalized to *nxp* rescaled dataset, Y; k number of clusters and a set of initial centres, C = {}.

Step 2: Calculate the distances $d_p(\mathbf{x}_i, \mathbf{x}_j)$ and obtain the 0.25^{th} trimmed mean of the $n(n – 1)$ distances as $\hat{\varepsilon}$.

Step 3: For any $\mathbf{x}_i$ and $\mathbf{x}_j$, $i \neq j$, use $\hat{\varepsilon}$ to compute the neighbourhood of each $\mathbf{x}_i$ and compute the corresponding cohesion degree of $\mathbf{x}_i$ and the coupling degree between $\mathbf{x}_i$ and $\mathbf{x}_j$. Find the point with the highest cohesion degree as the first centre point c_1 such that $C = \{c_1\}$.

Step 4: Find the point with second highest cohesion degree as *b*. Is the coupling degree between c_1 and $b_1 < \hat{\varepsilon}$? Choose b_1 as the second centre point; else, find the point with the next highest cohesion degree and check until b_j such that the coupling degree between c_1 and $b_j < \hat{\varepsilon}$. Then $C = \{c_1, c_2\}$.

Step 5: Repeat step 4 until $|C| = k$.

Step 6: Output: $C = \{c_1, c_2, \ldots, c_k\}$.

The output of the initialization algorithm is therefore passed into the clustering as the initial centres to produce *k* independent clusters of the dataset.

4. Comparative Clustering Validation

To evaluate clustering performance of the improved initialization method, it is compared with clustering that is based on Cao et al. [10]. Order to conduct the comparison, experimental datasets are clustered using the two methods of initialization and some performance indices are employed.

4.1. Experimental datasets

A 2–component dataset consisting 20 observations is obtained from Cao et al. [10] as Example data for the purpose of comparing the clustering performances of the two competing initialization methods. Others are five real-world datasets obtained from the University of California, Irvine machine learning repository at https://archive.ics.uci.edu/datasets and the GSE1650 dataset at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1650. These include the Iris dataset with 150 observation vectors and four variables, the Cancer dataset comprising 569 observation vectors and 30 variables, the Wine dataset with 178 observation vectors and 13 variables, the Wheat seed dataset with 210 observation vectors and 7 variables and the gene expression dataset, GSE1650, with 30 observation vectors and 22,833 genes as variables. The datasets are selected to represent varying data sizes, dimensionalities, and noise levels and are described in Table 1.

Table 1: Description of the Six Datasets

Dataset	Number of observations	Number of variables	Number of clusters
Example data	20	2	3
Iris data	150	4	3
Cancer data	569	30	2
Wine data	178	13	3
Wheat seed data	210	7	3
GSE1650	30	22833	2

In order to determine the noisy nature of the six datasets, the boxplots of their pairwise distances, based on Manhattan distance as “Manh”, Euclidean distance as “Eucli” and Chebyshev distance as “Maxi” are plotted and presented in Figure 1. From the boxplots, it is obvious that the Example and Iris data are free from noise while Cancer dataset, the Wine dataset, the Wheat seed dataset and GSE1650 gene expression dataset are very noisy.

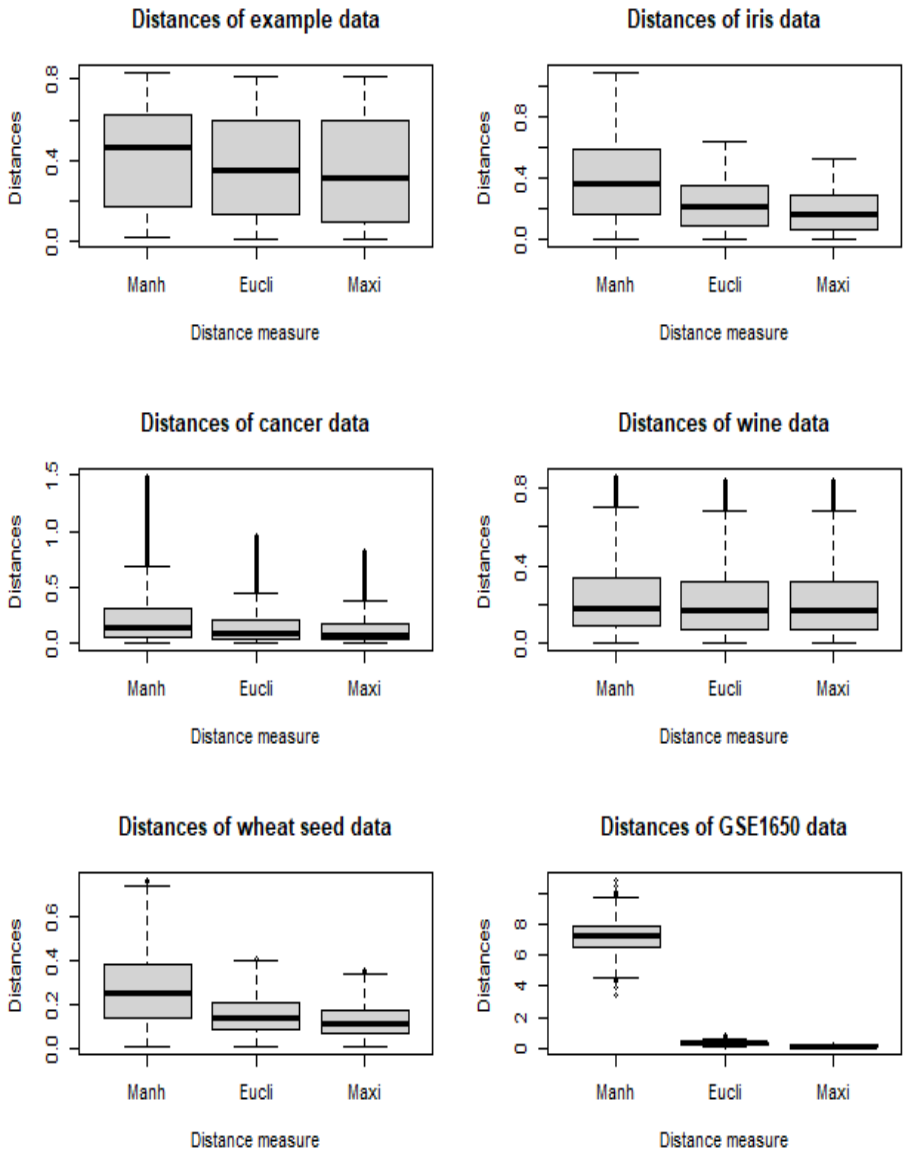

Fig. 1: Boxplots of the Pairwise Distances of the Six Real-Life Datasets.

4.2. Internal validation indices

In clustering, there are two classes of validation indices, namely: internal and external validation indices. While the external validation indices require data point labels, such is not needed for the internal validation indices. As a result, the latter is employed in this work and the specific indices employed include the Silhouette index (SI), the Davies–Bouldin index (DBI), the Calinski–Harabasz index (CHI) and the Dunn–Hubert index (DHI). The Silhouette index, also known as the Silhouette coefficient measures how well-separated and compact the clusters are simultaneously, at the level of individual points and the overall clustering. In other words, it measures how well each observation fits within its assigned cluster compared to other clusters by capturing both cluster cohesion and cluster separation. The index has range of values from -1 to 1 and higher values indicate better defined clusters. Also, the Davies–Bouldin index (DBI) measures how similar each cluster is to its closest neighboring cluster by combining within-cluster scatter and between-cluster separation, with range of values $(0,\infty)$ where smaller values indicate better clustering. Again, the Calinski–Harabasz index (CHI) also known as the variance ratio criterion measures how well clusters are separated relative to how compact they are. In other words, it measures the ratio of between-cluster variance to within-cluster variance, adjusted for the number of clusters, with higher values indicating better clustering quality. Lastly, the Dunn–Hubert index, usually called the Dunn index measures how well clusters are separated relative to their internal compactness by evaluating clustering quality as the ratio of the minimum inter-cluster distance to the maximum intra-cluster diameter. It evaluates the worst-case separation between clusters relative to the worst-case cluster compactness, with higher values indicating better clustering. The functional forms of these internal validation indices are presented in Table 2.

Table 2: Considered Internal Validation Indices

Index	Definition	Direction of better clustering
Silhouette Index	For any data point, x in *i*th cluster, $SI = \frac{1}{k}\sum_{i=1}^{k}\frac{b(\mathbf{x}_i)-a(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i),b(\mathbf{x}_i)\}}$ where $a(\mathbf{x}_i)$ is the average distance from $\mathbf{x}_i$ to all other points in cluster i and $b(\mathbf{x}_i)$ is the minimum average distance from $\mathbf{x}_i$ to points in other clusters.	Higher value
Davies–Bouldin index	For any two clusters C_i and C_j, $DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{j,j\neq i}\left\{\left(\frac{1}{n_i}\sum_{\mathbf{x}\in C_i} d_p(\mathbf{x},c_i)+\frac{1}{n_j}\sum_{\mathbf{x}\in C_j} d_p(\mathbf{x},c_j)\right)\Big/\left(d_p(c_i,c_j)\right)\right\}$ where c_i and c_j are centre points of clusters C_i and C_j respectively.	Lower value
Calinski–Harabasz index	$CHI = \frac{tr(B_k)/(k-1)}{tr(W_k)/(n-k)}$ where B_k is the between-cluster scatter matrix and W_k is the within-cluster scatter matrix.	Higher value
Dunn–Hubert index	Given clusters C_i, $i = 1, 2, \ldots, k$, $DHI = \min_{i\neq j} d_p(C_i,C_j)\Big/\max_i \Delta(C_i)$, where $d_p(C_i,C_j)$ is the distance between *i*th and *j*th clusters, and $\Delta(C_i)$ is the maximum pairwise distance within cluster C_i.	Higher value

Note: k is the number of clusters and n_i is the number of observations in the ith cluster.

4.3. Clustering performances using the two initialization methods

To analyze the efficiency of the improved initialization method relative to the Cao et al.'s method, the four datasets are clustered using each of the two initialization methods under the three Minkowski distances and their performances based on the four internal validation indices are obtained and presented in Table 3.

Table 3: Performance Scores of the Improved Initialization and the Cao et al. [10] Methods in the *k*-means Clustering of Six Different Datasets

Data	Index	Cao et al. method				Improved method			
		Manh	Eucli	Maxi	Ave	Manh	Eucli	Maxi	Ave
Example data	Silhouette	0.7367	0.7367	0.7367	0.7367	0.7367	0.7367	0.7367	0.7367
	Davies-Bouldin	0.2728	0.2728	0.2728	0.2728	0.2728	0.2728	0.2728	0.2728
	Calinski-Harabasz	139.63	139.63	139.63	139.63	139.63	139.63	139.63	139.63
	Dunn-Hubert	1.3829	1.3829	1.3829	1.3829	1.3829	1.3829	1.3829	1.3829
Iris data	Silhouette	0.5526	0.5526	0.5526	0.5526	0.5526	0.5526	0.5526	0.5526
	Davies-Bouldin	0.6623	0.6623	0.6623	0.6623	0.6623	0.6623	0.6623	0.6623
	Calinski-Harabasz	560.40	560.40	560.40	560.40	560.40	560.40	560.40	560.40
	Dunn-Hubert	0.0988	0.0988	0.0988	0.0988	0.0988	0.0988	0.0988	0.0988
Cancer data	Silhouette	0.6973	0.6973	0.6973	0.6973	0.6974	0.6974	0.6974	0.6974
	Davies-Bouldin	0.5044	0.5044	0.5044	0.5044	0.5035	0.5035	0.5035	0.5035
	Calinski-Harabasz	1300.2	1300.2	1300.2	1300.2	1300.1	1300.1	1300.1	1300.1
	Dunn-Hubert	0.0173	0.0173	0.0173	0.0173	0.0173	0.0173	0.0173	0.0173
Wine data	Silhouette	0.5731	0.5731	0.5731	0.5731	0.5823	0.5823	0.5823	0.5823
	Davies-Bouldin	0.5475	0.5475	0.5475	0.5475	0.5475	0.5438	0.5423	0.5445
	Calinski-Harabasz	490.92	490.92	490.92	490.92	490.92	490.92	490.92	490.92
	Dunn-Hubert	0.0238	0.0238	0.0238	0.0238	0.0421	0.0421	0.0421	0.0421
Wheat seed data	Silhouette	0.4020	0.4020	0.4020	0.4020	0.4224	0.4224	0.4224	0.4224
	Davies-Bouldin	0.9279	0.9279	0.9279	0.9279	0.9002	0.9002	0.9002	0.9002
	Calinski-Harabasz	249.78	249.78	249.78	249.78	249.98	249.98	249.98	249.98
	Dunn-Hubert	0.1188	0.1188	0.1188	0.1188	0.1284	0.1284	0.1284	0.1284
GSE1650 data	Silhouette	0.0307	0.0307	0.0307	0.0307	0.1174	0.1174	0.1174	0.1174
	Davies-Bouldin	3.7083	3.7083	3.7083	3.7083	2.2415	2.2415	2.2415	2.2415
	Calinski-Harabasz	1.8506	1.8506	1.8506	1.8506	2.8725	2.8725	2.8725	2.8725
	Dunn-Hubert	0.6069	0.6069	0.6069	0.6069	0.6441	0.6441	0.6441	0.6441

From Table 3, the performance results of the two methods for the non-noisy datasets (Example and Iris datasets) are the same in all the validation indices used. This shows that clustering a dataset using any of the two methods will give the same output provided the dataset is non-noisy. On the other hand, the results show that for noisy datasets (Cancer, Wine, Wheat seed and GSE1650 datasets), the improved method generally produced better clustering results almost in all the four internal validation indices used. In the Cancer data, the improved method produced higher Silhouette coefficient of 0.6974 compared to 0.6973 obtained by the Cao et al., method indicating an improved cluster cohesion and separation. Similarly, DBI decreased from 0.5044 in the Cao et al. method to 0.5035 in the improved method indicating a better cluster result with smaller within-cluster scatter and higher between cluster separation. Unfortunately, the CHI dropped from 1300.2 to 1300.1 from the Cao et al. to the improved method, indicating that the improved method failed to perform better in terms of how well the clusters are separated relative to how compact they are. Finally, for the DHI, the two methods scored equal index value of 0.0173. Concerning the Wine data, the Cao et al. method produced the SI of 0.5731 while the improved method obtained an improved index value of 0.5823. Also, the DBI dropped from 0.5475 to 0.5445 from the Cao et al. to the improved method indication better clustering using the improved initialization method. Again, the CHI remained the same for the two methods while the DHI improved from 0.0238 to 0.0421 in favour of the improved method, reflecting enhanced minimum inter-cluster separation relative to intra-cluster dispersion ([28]). For the Wheat seed data, the Cao et al., method produced the SI of 0.4020 as against the better result of 0.4224 from the improved method. Also,

the DBI dropped from 0.9279 to 0.9002 from the Cao et al., method to the improved method indicating an improved clustering quality in favour of the improved method. In a similar manner, the CHI improved from 249.78 to 249.98, from the Cao et al., to the improved method. Again, the DHI showed an improvement from 0.1188 to 0.1284 in favour of the improved method. A similar pattern of results was recorded with the GSE1650 dataset, where the SI for the Cao et al., method was 0.0307 against 0.1174 obtained with the improved method, indicating a better clustering with the improved method. Also, the Cao et al., method produced a DBI of 3.7083 against a lower value of 2.2415 produced with the improved method. Again, the Cao et al., method produced CHI and DHI of 1.8506 and 0.6069 respectively while their counterparts with the improved method were 2.8725 and 0.6441 respectively which clearly shows that the improved method produced an enhanced clustering quality than the former. The comprehensive results demonstrate that incorporating a trimmed mean threshold into rough set neighborhood construction of the initialization improves k-means clustering when dataset is noisy. By reducing the influence of extreme distance values, the improved method enhances neighborhood stability and centroid representativeness. These findings are consistent with previous studies, which stated that k-means initialization methods are sensitive to noise. The present study therefore extends this line of research by introducing a quantile-based trimming strategy that directly addresses this limitation.

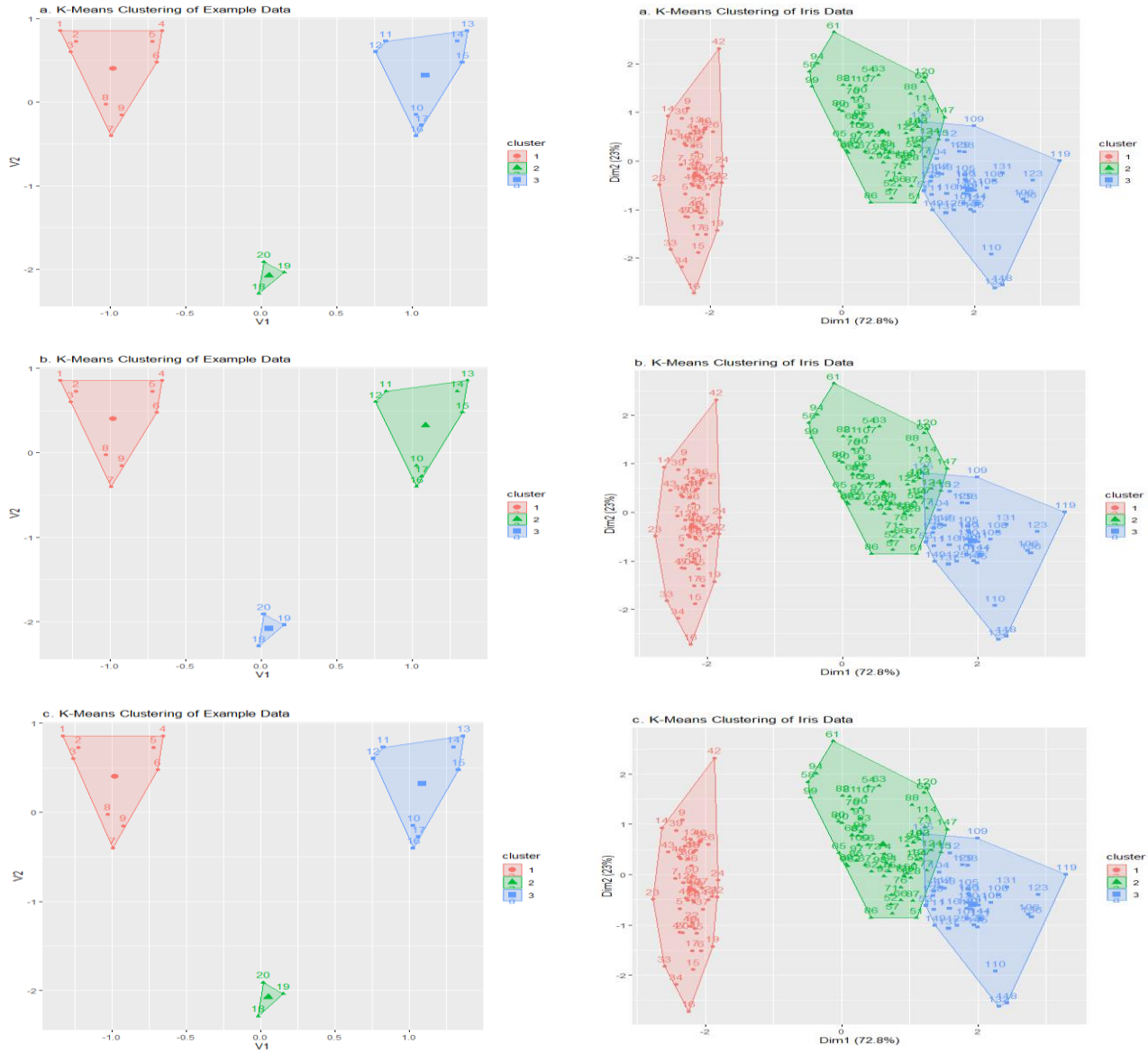

Fig. 2: *k*-Means Clustering Plots of the Example and Iris Datasets ($k = 3$) for Different Distances.

It is also important to state that both methods are insensitive to the type of Minkowski distance used, both in noisy and non-noisy datasets. This is because all the measures produced almost the same results in all three distances (Manhattan, Euclidean, and Chebychev) except in a few instances where non-significant differences were observed. As a result, any of the three distance measures can be used in k-means clustering. To show this further, the six datasets are clustered using the improved method of initialization with the three distance measures and the results are presented in Figures 2, 3 and 4.

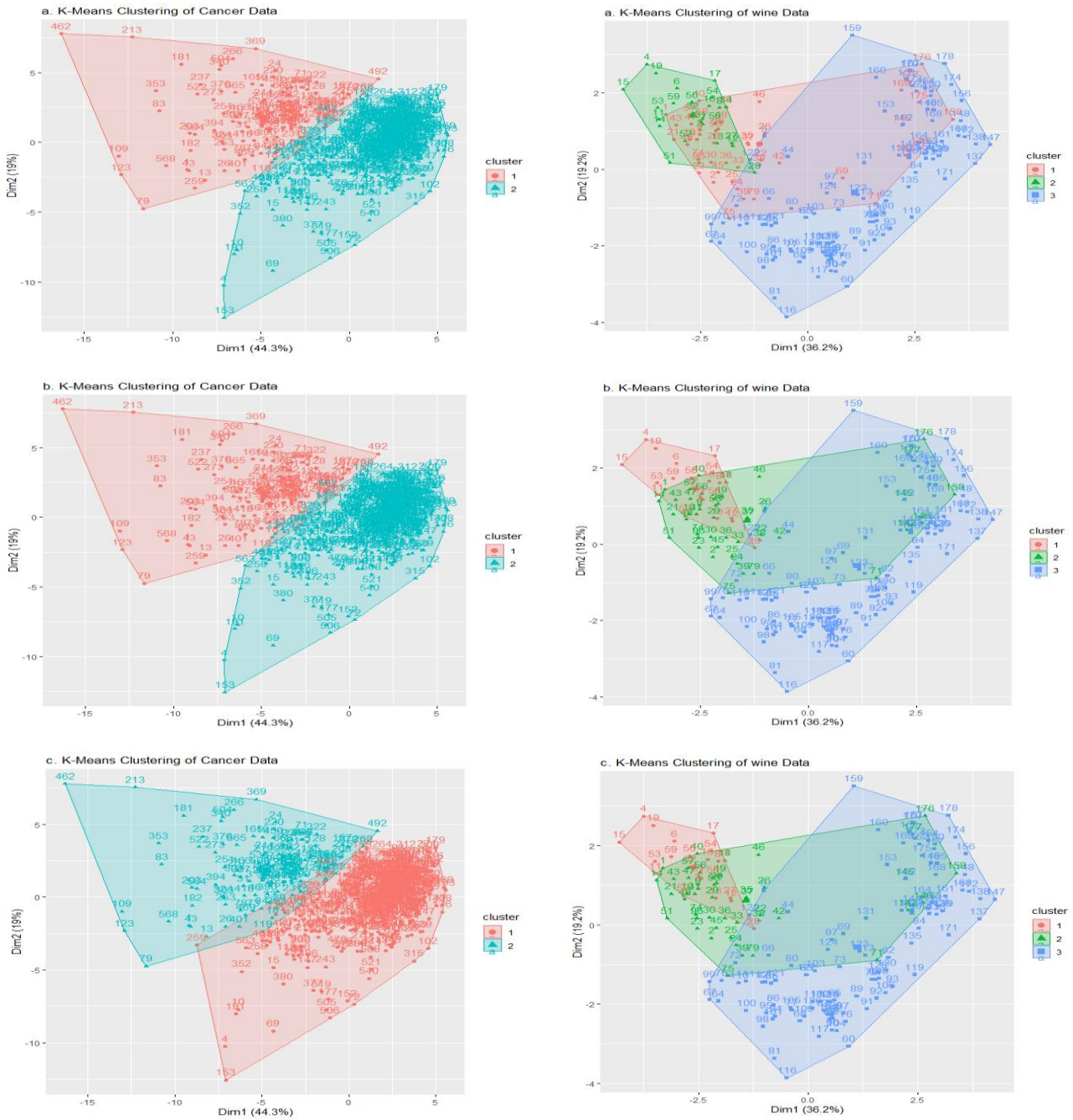

Fig. 3: *k*-Means Clustering Plots of the Cancer and Wine Datasets ($k = 2$ and 3 respectively) for Different Distances.

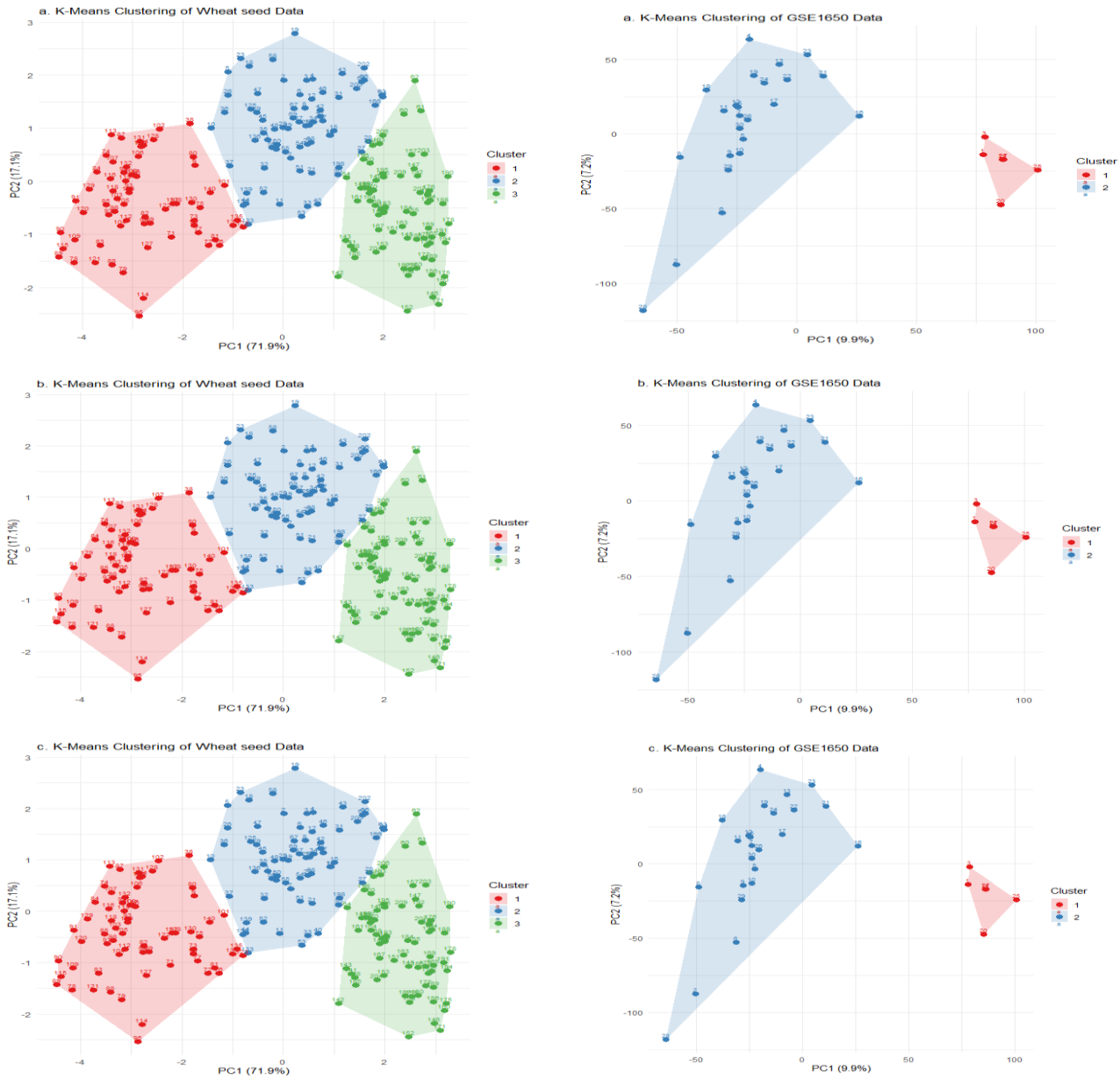

Fig. 4: *k*-Means Clustering Plots of the Wheat Seed and GSE1650 Datasets (k = 3 and 2 respectively) for Different Distances.

5. Conclusion

This study investigated the problem of centroid initialization in the *k*-means clustering algorithm, with particular emphasis on improving clustering performance in noisy datasets. Although *k*-means is widely used due to its simplicity and computational efficiency, its sensitivity to initial centroid selection often leads to unstable clustering results and convergence to suboptimal solutions. To address this limitation, an improved rough set neighborhood–based initialization method is introduced. The method refined neighborhood construction by introducing a quantile-based trimming approach for determining the neighborhood threshold. By using the 0.25th trimmed mean of inter-point Minkowski distances, the influence of extreme values and noise on neighborhood formation was effectively reduced. This adjustment enabled the identification of more stable and representative neighbourhoods, which in turn facilitated the selection of suitable initial centroids based on cohesion and coupling degrees. Experimental evaluations conducted on six datasets demonstrated that the improved initialization method outperformed the rough set neighborhood–based initialization approach of Cao et al. [10] in noisy environments. Improvements were observed in key internal validation indices, including consistently higher Silhouette and Dunn–Hubert values and lower Davies–Bouldin scores. Although a marginal reduction was noted in one of the datasets for Calinski–Harabasz index, this trade-off was minimal and did not detract from the overall robustness and stability achieved by the improved method. On relatively noise-free datasets, both methods produced the same clustering results, indicating that the improved approach does not compromise performance under no noise (or low noise) conditions. As a result, the improved method can be recommended for use to cluster noisy datasets.

Acknowledgements

The authors wish to acknowledge the editors and the anonymous reviewer for their suggestions which significantly improved the quality of the paper.

References

[1] Jain AK (2010), Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* 31, 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.
[2] Mardia KV, Kent JT & Taylor CC (2024), *Multivariate analysis* (2nd ed.). John Wiley & sons, New York.
[3] Tan P-N, Steinbach M & Kumar V (2006), *Introduction to data mining*. Pearson Addison Wesley.
[4] Han J, Kamber M & Pei J (2012), *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers.
[5] MacQueen J (1967), Multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281-297).
[6] Bradley PS, Fayyad UM & Reina C (1998), Scaling clustering algorithms to large databases, knowledge discovery and data mining.
[7] Likas A, Vlassis N & Verbeek JJ (2003), The global k-means clustering algorithm. *Pattern Recognition 36*, 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2.
[8] Yedla M, Pathakota SR & Srinivasa TM (2010), Enhancing k-means clustering algorithm with improved initial center. *International Journal of Computer Science and Information Technologies* 1, 121–125.
[9] Arthur D & Vassilvitskii S (2006), *k*-means++: The advantages of careful seeding. Stanford.
[10] Cao F, Liang J & Jiang G (2009), An initialization method for the K-means algorithm using neighborhood model. *Computers & Mathematics with Applications 58*, 474-483. https://doi.org/10.1016/j.camwa.2009.04.017.
[11] Pawlak Z (1982), Rough sets. *International Journal of Computer & Information Sciences*, *11*, 341–356. https://doi.org/10.1007/BF01001956.
[12] Xu X, Li J & Zhou Z (2009), A weighted k-means clustering algorithm based on distance optimization. *Expert Systems with Applications 36*, 6983–6987.
[13] Ahmed AH & Ashour W (2011), An initialization method for the k-means algorithm using RNN and coupling degree. International Journal of Computer Applications 25, 1–6. https://doi.org/10.5120/2999-4030.
[14] Zahra S, Ghazanfar MA, Khalid A, Azam MA, Naeem U & Prugel-Bennett A (2015), Novel centroid selection approaches for KMeans-clustering based recommender systems. Information Sciences 320, 156-189. https://doi.org/10.1016/j.ins.2015.03.062.
[15] Yang J, Ma Y, Zhang X, Li S & Zhang Y (2017), An initialization method based on hybrid distance for k-means algorithm. *Neural Computation 29*, 3094–3117. https://doi.org/10.1162/neco_a_01014.
[16] Mishra BK, Rath AK, Nanda SK & Baidyanath RR (2019), Efficient intelligent framework for selection of initial cluster centers. *International Journal of Intelligent Systems and Applications 11*, 44–55. https://doi.org/10.5815/ijisa.2019.08.05
[17] Chowdhury K, Chaudhuri D & Pal AK (2021), An entropy-based initialization method of k-means clustering on the optimal number of clusters. *Neural Computing and Applications 33*, 6965–6982. https://doi.org/10.1007/s00521-020-05471-9.
[18] Yang J, Wang Y-K, Yao X & Lin C-T (2021), Adaptive initialization method for the K-means algorithm. *Frontiers in Artificial Intelligence 4*, 740817. https://doi.org/10.3389/frai.2021.740817.
[19] Sujatha N, Latha Narayanan V, Prema A, Rathiha SK & Raja V (2022), Initial centroid selection for K-means clustering algorithm using the statistical method. *International Journal of Science and Research Archive 7*, 474–478. https://doi.org/10.30574/ijsra.2022.7.2.0309.
[20] Gul M & Rehman M (2023), Big data: An optimized approach for cluster initialization. *Journal of Big Data 10*, Article 120. https://doi.org/10.1186/s40537-023-00798-1.
[21] Zu Y, Wu J, Zhao G, Wang M & Zhou X (2024), II-LA-KM: Improved initialization of a learning-augmented clustering algorithm for effective rock discontinuity grouping. *Mathematics, 12*(20), Article 3195. https://doi.org/10.3390/math12203195.
[22] Zubair M, Iqbal MA, Shil A, Chowdhury MJM, Moni MA & Sarker IH (2024), An improved k-means clustering algorithm towards efficient data-driven modeling. *Annals of Data Science, 11*(5), 1524–1544. https://doi.org/10.1007/s40745-022-00428-2.
[23] Gao C, Yong X, Gao Y.-L & Li T (2024), An improved black hole algorithm designed for k-means clustering method. *Complex & Intelligent Systems, 10*, 5083–5106. https://doi.org/10.1007/s40747-024-01420-4.
[24] Zhang S, Chen S, Yu X & Mei S (2025), Research on collaborative filtering algorithm based on improved k-means algorithm for user attribute rating and co-rating. *Scientific Reports, 15*, Article 19600. https://doi.org/10.1038/s41598-025-96705-0.
[25] Ugwu MC & Madukaife MS (2022), Two-stage cluster sampling with unequal probability sampling in the first stage and ranked set sampling in the second stage. *Statistics in Transition new series 23* 199–214. https://doi.org/10.2478/stattrans-2022-0038.
[26] Meilă M & Heckerman D (1998), An experimental comparison of several clustering and initialization methods. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI'98) (pp. 386–395). Morgan Kaufmann.
[27] Celebi ME, Kingravi HA & Vela PA (2013), A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications 40*, 200–210. https://doi.org/10.1016/j.eswa.2012.07.021.
[28] Handoyo S, Marji M, Effendi MR & Kusnadi K (2014), The use of silhouette and fuzzy c-means clustering for customer segmentation. *International Journal of Engineering & Technology 3*, 354–358.