EVALUATION OF INITIALIZATION METHODS FOR THE PERFORMANCE OF THE K-MEANS ALGORITHM

Sinuhé Ginés-Palestino, Eduardo Roldán-Reyes*, Marcela Quiroz-Castellanos y Guillermo Cortés-Robles

Tecnológico Nacional de México (México)

DOI: https://doi.org/10.52152/D11325

1.- INTRODUCTION

Even though there are clustering algorithms with a high response capacity, it is common to observe using K-means as the preferred method by researchers [1]. Supported by the fact that the algorithm is easy to implement also because its effectiveness has been tested and validated for a longer time, in contrast to new proposals for data clustering, which are more effective [2]-[3]. Another aspect of relevance is that it sets a benchmark for the use of better algorithms due to its capacity for local data processing, such as the Genetic Algorithm [4], Particle Swarm Optimization [5], and density clustering [6], among others.

The heuristic of the K-means algorithm is segmented into four core phases: initialization, assignment of elements to centroids, centroid updating, and decision. The number of K clusters is established in the initial phase, indicating the first segmentation of random elements under study. The second phase randomly adds elements to each cluster to be contained in each existing K group. In the next phase, the centroid of each group is calculated by averaging their distances. In the last phase, if there is the convergence of data close to the centroids, it is evaluated, intending to obtain a new grouping, taking new decisions as the iterations evolve until it stops obtaining improvements in the groupings, that is to say, that these iterations are cyclized, giving similar results to the previous ones obtained [7].

All of the above results in a grouping of data sensitive to local changes within each cluster; however, from a global perspective, the nature of the algorithm prevents the centroids from adjusting between clusters. The reason is that centroids cannot move between clusters if their distance is large or if other stable clusters prevent movements. Poor initialization can cause iterations to get stuck at a lower local minimum [8]-[9]. Experiments with three different indices have shown that K-means rarely achieves the correct number of clusters, while in random exchange, it succeeds in most cases [10].

The initialization is a success factor for good performance and operation, which supports the approach of this research, that the key is in the “initialization” stage, contemplating the variants of initial grouping and order of instances [11]. According to the literature review, some initialization methods are suggested as agents of change in the performance of the algorithm. Also, some contributions from the application of K-means in different industrial sectors of relevance are discussed, where the algorithm is used in a hybrid way, in combination with initialization techniques, or implemented naturally, without any alteration in its heuristic. Therefore, dataset performance measures such as group overlap, cluster number, dimensionality, and cluster size imbalance should be considered critical to algorithm success [12].

This work is mainly focused on discussing the contributions documented in the literature on the aspects in which K-means can be improved according to the different initialization techniques, repetition, and other criteria that impact its performance. In a complementary way, some cases of applications of the algorithm in the industrial sector are discussed in order to analyze the effect of its use with different input variables. This article is divided into five segments: introduction, performance, initialization methods, applications, and conclusion.

2.- K-MEANS PERFORMANCE FACTORS

To begin this section, the methodology for the selection of representative articles for this research is described, which is shown in Figure 1. First, the following keywords are established: “K-means algorithm,” “initialization methods,” “performance,” and “applications.”; subsequently, Science Direct, Web of Science, and Scopus are selected as the databases to be used for the search of scientific contributions contained in the literature. In addition, categories are defined to filter the required information, such as: “research articles,” “publication time interval,” “application areas,” and “specialist evaluation.” In the next step, the search by categories and keywords mentioned above was carried out, finding 73 relevant manuscripts. However, 18 did not fully comply with the information needed to carry out this research. According to the previous result, the search parameters were adjusted to narrow and specify the topic of interest, achieving a total of 62 research articles that were selected for the analysis and discussion of their content in this document.

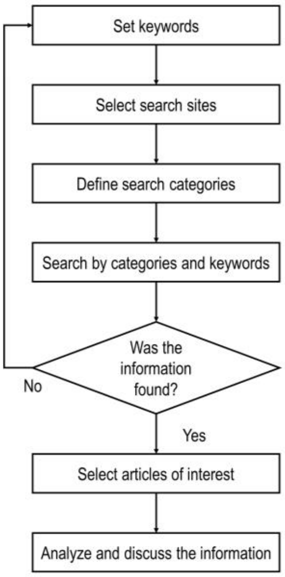
Set keywords
Select search sites
Define search categories
Search by categories and keywords
Was the information found?
No
Yes
Select articles of interest
Analyze and discuss the information

Fig. 1. Methodology used for article selection

The above methodology for selection and literature review is intended to demonstrate the factors that revolve around the good performance of the K-means algorithm. For a better understanding of the topic, it is important to highlight the difference between a clustering method and a clustering algorithm; in the former, the data clusters are evaluated from an objective function, and in the latter, the optimization of the cluster generation process [13]. As mentioned before, the nature of the initial clustering of the data can mark a starting point in the algorithm's operation. It should be mentioned that in the initialization stage, where the number of "K" groups is established, it is possible to identify balanced or unbalanced features. K-means will perform poorly when the group balance is not homologous [14]. Specifically, the relatively large size of the cluster can cause its fragmentation to be erroneous, and in small clusters, its adherence to other groups can also generate errors when joining [15].

The errors raised in the previous paragraph rule out the possibility that a bad evaluation of the sum of squares of the error could cause the central motif. Accordingly, if the data are contained in non-uniform groups, they will be divided into uniform subgroups in the direction of the highest variance. Groups of different sizes result in sectioning into large subgroups and adherence in small subgroups. It is worth mentioning that advantages can be obtained in the algorithm because of the natural properties per each cluster, avoiding the use of the sum of squares of the error and opting for an approach based on the Mahalanobis distance objective function or Gaussian mixture model [16]-[17].

Some recent contributions elaborate on the most popular performance measures of the K-means algorithm, such as [18]. The performance measures that stand out are computational efficiency, clustering quality, and stability. For computational efficiency, it is measured mainly through the execution time of the algorithm. It evaluates how long the algorithm takes to converge to a stable solution. This aspect is critical, especially when working with large data sets. The quality of clustering is measured using several metrics, the most important of which are the Total Distances, Silhouette Index, and Davies-Bouldin Index. In general, the above three measures assess inter-cluster and intra-cluster distance. Experimental analyses indicate a comparison of several variants of k-means using six standard datasets widely used in the machine learning and data mining community. The results highlight that standard k-means acted as a basis for comparing other variants; k-means with Cuckoo-enhanced initialization showed significant improvements in clustering quality and stability. K-means adapted for categorical data achieved competitive results on mixed datasets.

In the proposal made by [19], the efficiency of the algorithm is evaluated mainly through the execution time, which is measured in milliseconds. This time varies according to the number of clusters and the size of the data set, which in this study includes sets of 100,000 to 500,000 points in a two-dimensional space. The evaluation of the run time allows us to determine how fast the algorithm can process large volumes of data. The quality of the clustering is analyzed by several distance metrics used to calculate the similarity between points and centroids. In this study, three main metrics are implemented: Euclidean Distance, Manhattan Distance, and Minkowski Distance. The first one shows better performance for configurations with 6 and 16 clusters; the second metric generally provides the best results in terms of execution time for specific configurations (4, 8, 12, and 14 clusters). Finally, Minkowski turns out to be more effective only for configurations with 10 clusters. In general, as the size of the dataset increased (up to 500,000 points), it was noticed that Euclidean and Minkowski distances presented similar times.

In the research carried out by [20], the weaknesses of the standard algorithm are addressed, such as the dependence on random initialization of centroids, which can lead to suboptimal results and convergence to local minima. It is also pointed out that its sensitivity to the similarity metric (usually Euclidean distance) limits its ability to identify non-spherical cluster shapes and overlapping clusters. A comprehensive review is made of modifications to K-means designed to improve its performance in key areas, such as initialization optimization, cluster number selection, and robustness to large data volumes. These variants seek to overcome the structural limitations of the standard algorithm, making it more suitable for applications in the context of big data. In future work, hybrid approaches combining K-means with advanced machine learning techniques are proposed to improve its accuracy and scalability.

The study proposed by [21] highlights the importance of the proper selection of initial centroids, as different choices may lead to different local optimal results. Three new algorithms for initial seed selection are proposed, using statistical measures such as mean, median, and partition center to compute the cluster centroids. Methods such as Elbow and Silhouette Index are used to predict the optimal number of clusters (k). The Elbow method identifies a point in the graph where the percentage variation begins to stabilize, while the Silhouette Index measures clustering quality by comparing the intra-cluster distance with the distance to the nearest cluster. Performance is evaluated using indices such as the Silhouette Index and the Dunn Index, which provide a clear indication of the quality of the clusters formed. A high value on these indices suggests good inter-cluster separation and cohesion within clusters. Although traditional methods often outperform clustering-based techniques, this study suggests that with appropriate settings, K-means can be a viable alternative for reducing inputs in coastal models and other contexts.

The most noticeable characteristics of the data clusters for evaluating their performance are detailed in the following points, considering Overlap, number, dimension, and cluster size imbalance. It is worth mentioning that the algorithmic performance approach will be maintained over that of the K-means objective function, contemplating the benchmark for clustering based on the sum of squares of the error.

2.1 GROUP OVERLAPPING

The Projection Clustering Grouping and Consistent Fuzzy Sample Transformation Method (PCGDST-IE) is proposed to learn the information of class overlapping samples better and simultaneously deal with class imbalance and class overlapping problems. The process of solving the class overlap and imbalance problem by PCGDST-IE is as follows: firstly, Weighted Projection Clustering Combination (WPCC) converts the training set with sparse and mixed distribution into a set of subsets with simple distribution, which can perform clustering and feature reduction simultaneously to facilitate further processing. Then, SHS completes each subset's overlapping and balancing operation with the simple class distribution. Finally, the Local–Global Structure Consistency Mechanism (LGSCM) constructs a subset with further reduction of class overlap and richer single sample information to improve the SHS subset. Compared with existing related algorithms, PCGDST-IE can achieve significantly better performance [22].

In this analysis, an improved k-means clustering method using medical data was applied. A hybrid method combining K- Harmonic Means and Overlapping K-means algorithms (KHM-OKM) was used, where the initial points are selected according to the results of the KHM clustering algorithm. Experimental results using ten publicly available medical datasets show that the proposed hybrid method provides better or comparable results than the original OKM algorithm [23].

The cluster overlay method can also be used in legal tasks. Multi-view Overlay Clustering of Legal Judgments (MOSTA), a novel artificial intelligence method, can identify groups of legal judgments with similar features, possibly corresponding to topics, thus reducing the human effort required to navigate, organize, and classify large amounts of legal judgments. Methodologically, MOSTA learns two different models for incorporating legal judgments. The first one aims to represent the semantics of the textual content, while the second one aims to represent citations of legal acts, also considering the granularity of the citations. Then, these representations are merged using a multi-view approach based on an automatic encoder, and the obtained representation is finally exploited using a new overlapping clustering algorithm [24].

2.2 CLUSTERS NUMBER

Automatic Clustering by Differential Evolution (ACDE) has the main purpose of automatically obtaining the number of clusters. However, it still uses manual strategies to determine the active area of k, which affects its performance. With this problem, the U-Control Chart (UCC) method can be used as an alternative to initially calculate the active area of k to obtain the variables before generating the characteristic vectors of the algorithm. A comparative analysis between ACDE-K-means versus UCC-ACDE-K-means applied to entrepreneurial opportunities in SMEs showed results that optimized the group assignment in the initial stage of the algorithm [25].

Center-Based Clustering (CBC) is another alternative for determining the number of clusters. Generally, it requires simple calculations compared to other methods. Two new methods have been analyzed: the Last Leap (LL) and the Last Major Leap (LML). Both methods determine only the minimum distance between centers, which generates an advantage for subsequent decision-making since clusters can be constituted from different perspectives and have lower computational complexity than existing methods. The ability of these two methods applies for more than 50 k; they are sensitive to small and large changes in the data on heterogeneous clusters, as well as differences in their distributions and significant cluster overlap [26].

An adaptive algorithm in K-means was used and applied to segment images of tomato leaves. It was compared with three different algorithms, DBSCAN, Mean Shift, and ExG-ExR, yielding better results in the segmentation. To support the above, color area analysis was performed through indices such as Davies-Bouldin (DB), Calinski-Harabasz (CH), and Silhouette Index (SI), greatly improving the accuracy in the calculation of group determination [27]. An algorithm based on repetitive sampling and analyzing three approaches: variance, repetitive, and structural sampling is suggested. Unlike the algorithm proposed in the previous paragraph, the use of the CH, SI, Hartigan, Jump (HJ), and Gap Static (GS), among others, is a good reference when the groups are well separated [28].

2.3 DIMENSIONALITY

The dimensionality consists of the space that holds the data groups contemplated in the algorithm. As previously stated, the more separated the clusters are, the fewer errors and the more effective the K-means performance tends to be in assigning the number of clusters. One could assume an adequate performance in a comparative study between two sets, the first with 16 properly separated clusters; however, almost 4 clusters lack a centroid or are assigned to an incorrect one, as well as 0% effectiveness, despite the well-defined separation of the clusters. This behavior is attributed to the lack of Overlap in the clusters and not necessarily to dimensionality. The second set shows a dependence between the percentage of effectiveness and dimensionality. An analysis of 4 centroids represented an imbalance by the assignment of 3 centroids in the first group and the remaining one for the second group [12].

2.4 CLUSTER SIZE IMBALANCE

The cluster size imbalance is a disadvantage, especially when a random initialization is considered since a balanced selection of the central points of each set is complicated. Considering the initialization, an inference method is used to reduce randomness. Under the combination of a Cluster Randomized Design (CRD), a new model is proposed to balance the cluster size; it is worth mentioning that the basic principle for its operation is simple Random Sampling (SRS). Therefore, the CRD incorporates the principle of Ranked Set Sampling (RSS), which can be balanced; thus, under the combination of a CRD with a BRSS, the result is a structured design called BRSS-CRD. It was shown that the proposed design uses SRS-CRD, achieving balanced groups in the algorithm [29]. Also, a model sensitive to inequalities between numbers in each group at any CRD stage is proposed, posing as URSS [30].

In some cases, the combination of an algorithm is used to solve the cluster size imbalance. In this case, the Synthetic Minority Oversampling Technique (SMOTE) is proposed; in turn, it is combined so that a hybrid is created, called LR-SMOTE, because of the consideration of the radius limit. Improving the quality of the samples by removing noise prior to subsampling is the main intention of this algorithm to balance the cluster size [31].

In retrospect, Table 1 condenses the description of the performance measures presented in the previous paragraphs with the intention of strengthening the concepts and their relationship with the references addressed in the literature, as it will be useful for analyzing their impact in subsequent sections.

Performance Factors	Description	References
Group overlapping	- Clusters share similar or overlapping points. - Makes correct identification and grouping difficult. - Represents the erroneous assignment of points in a cluster that does not correspond to them.	[22]-[23]-[24]
Clusters Number	- Critical measure that determines the accuracy of the algorithm. - Underestimation or overestimation of the clusters gives confusing results. - Depends directly on the initial centroid assignment.	[25]-[26]-[27]-[28]
Dimensionality	- Number of features or variables describing each point in the data set. - It has no direct impact on the performance of the algorithm. - The disadvantage and indirect effect of exceeding the dimensionality of a dataset can complicate the interpretation of results and increase the computational cost.	[12]
Cluster size imbalance	- It represents the inequality between the size and distribution of clusters. - Excessive imbalance may cause partitions in the clusters or unification of small clusters with others. - It may generate higher computational costs and not provide an optimal solution.	[29]-[30]-[31]

Table 1. Description of Performance Factors

3. INITIALIZATION METHODS FOR K-MEANS

The K-means algorithm turns out to be an excellent option for carrying out clustering analysis since, at the local level, the results it yields are sensitive to small changes. However, it has marked weaknesses, especially in positioning the centroids where the cluster is placed on another cluster.

Initialization turns out to be a critical factor for the good performance of the algorithm, but it is not the only thing that must be considered to obtain good results. It seems that it is a correlational behavior having a better initialization; the algorithm will also be better; however, other opportunities should be considered, such as the initial location of the centroids and the generation of subsequent clusters. The nature of the algorithm is summarized in its easy application, an essence that should not be lost when applying a complex initialization criterion, although in certain cases, it is subjective to evaluate its complexity. Based on an analysis, repair k-means leads to 5 functions and 162 lines of code, which is summarized in an acceptable degree of difficulty compared to other initialization criteria [32]. Based on the time complexity, an evaluation can also be made, classifying linear, logarithmic, and quadratic algorithms to select the one with the shortest response time [1].

In this proposal [33], a new initialization method is established based on a neighborhood model that uses two key concepts: the degree of cohesion that measures the intracluster similarity, evaluating how compact the objects within a neighborhood are, and the degree of coupling that measures the intercluster similarity, evaluating how the neighborhoods are related to each other. The study compares the performance of the new method with other traditional approaches, such as random initialization and methods like MaxMin. Experimental results show that the proposed new method significantly improves the quality of clustering by selecting more representative centers and reducing variability in the final results. The proposed method not only provides better initial centers but also reduces the time required to reach adequate convergence compared to random initializations. In addition, an improvement in the cohesion of the clusters formed was observed, indicating that the generated groups are more homogeneous and representative.

Research by [34] addresses the limitations of random initialization. Alternative methods have been proposed; k-means++ improves initial selection by choosing centers that are farther apart. It is based on a probabilistic distribution that favors the choice of distant points, which helps to improve convergence and clustering quality. Density-based initialization uses techniques that identify densely populated areas in the data to select initial centers. This approach seeks to ensure that the centers represent the underlying structures of the data well. Heuristic methods use algorithms to optimize the initial selection, such as the use of genetic algorithms or optimization techniques. The results of the study show that improvement in clustering quality, K-means++, and density-based techniques were found to produce more homogeneous and well-defined clusters. The alternative methods tend to require fewer iterations to reach convergence, which translates into less computational time. The improved techniques provide more consistent results that are less dependent on randomness, which is crucial for practical applications where reproducibility is required.

The paper [35] discusses a new center initialization method for the K-means algorithm, called CIT (Cluster Initialization Technique), focused on datasets containing two clusters. The CIT method is based on the identification and selection of nearest neighbor pairs within the dataset. First, a pair of nearest neighbors is defined for each point in the dataset, which allows grouping points that are potentially in similar clusters. Subsequently, we seek to identify two pairs of nearest neighbors that are as dissimilar as possible and are likely to belong to different clusters. At the end, points are selected that meet certain conditions defined by four theoretical assumptions, ensuring that the initial centers are not in the Overlap between clusters. The results were compared with two existing methods: kd-tree and CCIA (Cluster Center Initialization Algorithm). Although CIT was somewhat more computationally expensive compared to kd-tree and CCIA, especially on small and low dimensional sets, its performance is still acceptable.

In this section, some initialization criteria that have been the subject of analysis and discussion by experts are presented to evaluate all the factors associated with a good performance of K-means, seeking a balance between the advantages and disadvantages that each one can generate.

3.1 RANDOM CENTROIDS

One of the most widely used methods for initializing K-means is based on the random selection of centroids [36]. The nature of the heuristic of the algorithm, in its initial stage, is supported by the random selection of data without any biased criteria. Therefore, each data cluster is guaranteed to have at least one proposed centroid, even if it does not meet the desirable characteristics to establish that first central point. As an interchange measure of the centroids, the Suffling Method is used, starting from the last element of the array and moving towards the first one, in order not to repeat the selection of centroids in the same way as in previous arrays; thus, the elementary structure of the initialization based on random centroids is built. In a more basic yet functional way, the method can cover only the initial choice of random points [37]. One can choose to use the first initialization mechanism; in both cases, the result will be the same, unlike if the Shuffling Method is not applied, there is a risk of repeating the same selection [38].

3.2 RANDOM PARTITIONS

This section deals with a method associated with random centroids since its operating mechanism is based on the random partitioning of the clusters. The data are grouped without any ordering criteria, and then the centroid of each cluster is calculated; it is worth mentioning that the number of K is obtained arbitrarily. The advantage of this method is that the centroid obtained disregards the values at the extremes of each grouping, reducing the bias derived from outliers. The disadvantage is that the centroids will converge at the center of all the data since they were calculated with the average. Based on the above, the technique works well when the groups overlap; otherwise, it works poorly. Therefore, it can be considered a deterministic behavior. This method overcomes the weaknesses generated compared to random centroids [11].

There is a technique that, in a way, combines random partitions with the centroid criterion. The proposal consists of calculating the average of the points and then adding random vectors. In this way, a kind of subgroup is created, considered a parameter. It is associated with the centroid technique since the subgroup is considered such. This way, even data located at the extremes or generating empty spaces can be covered [39].

3.3 FARTHEST POINT

Another initialization method for the K-means algorithm is that of the farthest point. First, a point is selected without any criteria; then, new points are selected starting from the first. The condition to be respected is that the new centroid must be the farthest point from the previous one for a new selection. It is categorized as a heuristic of maximum and minimum values [40]-[41].

There are different variants associated with the farthest point. One consists of measuring the distances between the data concerning the first selected centroid; when there is a certain set of distances, the smallest one is evaluated, and the selection of the new centroid is updated based on this criterion. In the end, the process is related to a K-means iteration; however, it can be accelerated, applying in a certain way a sectioning of data subgroups [42].

One way to choose the first centroid is to consider the two farthest points of the whole data set; thus, two centroids will be considered the starting point of the clustering. In addition, it suggests recalculating the centroids according to the new calculation of subgroups; the next election will start from the farthest centroid as it was done at the beginning [41]. Another practical way is to use the maximum density initialization to reduce values that go out of the normality margin, which will be detailed in a basic way in later sections [43].

Another measure of initialization based on a far point is to use the cumulative distance of each of the selected centroids [44]. The above approach works well but can be sensitive to failures when dealing with two relatively close values by contemplating cumulative distances of previous centroids [45]. A recommended selection variant is to consider the first centroid randomly, assign the randomness that the method deserves, and then apply the initial criterion of the farthest point [46].

3.4 CLASSIFICATION HEURISTICS

A more comprehensive strategy that can be employed for the initialization of K-means is by selecting the data based on a specific criterion. The ranking of this ordering can depend on heuristics such as k initial points, k initial points with elimination of points near centroids, and uniform partitioning. The central point criterion [47], density, and centrality [48] can be used to perform the sorting. Each will be explained in more detail in the following paragraph.

The central point consists of ordering the data based on the distance pattern to the center of all the data. The centroids are established as the distances to the center are obtained from smallest to largest. It can be combined with some randomness by choosing a point arbitrarily close to the center of the data. It simplifies speed and avoids additional parameters [47].

Density has a characteristic functionality, considering an accumulation of points within a particular area. Thus, the first centroid will be defined where the most saturated data area is found, hence the context of density. The following centroids will be defined according to those with the lowest density, i.e., the lowest accumulation of data, considering that they should not be closer to the previously defined centroids. Applying the average pairwise distance is useful, eliminating unnecessary parameters [40].

3.5 PROJECTION HEURISTICS

This technique is based on the classification heuristic; since the points selected are considered centroids of groups of equal size, the intention is to generate a projection or convexity of the clusters obtained to generate a balance in the selection structure. The method of selecting the points based on the center of the data is applied to generate the others based on the distances in the direction of the extremes [49]. By obtaining non-negative attributes, the objective of projecting the data using a diagonal line is achieved by calculating the average of the attributes, thus making it faster to search for data that are close to the clusters. The principal axis calculation maximizes the variance, which is useful in partitional clustering methods [50].

3.6 DENSITY HEURISTICS

In the classification heuristic, mention was made of the potential represented by the data density criterion for generating defined groups and estimating the initialization centroid. However, it represents a weakness when there is no means to size the density without causing slowness and complexity in the algorithm. Three strategies can be considered to help solve the density calculation: cubes, ε-radius circle, and K-Nearest Neighbors (KNN).

The first strategy consists of generating a grid over the data and establishing a dimension for each grid section. Subsequently, the accumulation of data captured in each area is concentrated in that section as part of the sectioning dimension [51]. It can be approached differently, provided the dimensions are considered specially, for example, under a heuristic approach [52]. An additional approach to divide the mesh is under the kd-trees criterion; it is suitable when each area is considered with the same number of points [53].

The second strategy is usually applied more traditionally by assigning a neighborhood, using a sectioning based on an ε-radius circle; subsequently, the density is contemplated, with the data falling in each sectioning [54].

The third strategy contemplates KNN, i.e., the distance between points that will make up the density of a cluster, which is calculated by identifying the nearest neighbor from one point to another, i.e., the average of the nearest distances to a point within a data group. Also, the density can be calculated according to each pair of points [55].

There are some combinations, such as the one proposed using preliminary groups applying the nature of k-means and eliminating the smallest ones, after which the farthest point is identified to select a new preliminary set of groups. This combination of density with the farthest point directly supports the problems when there are high-density groups; however, it becomes complex [56].

3.7 DIVISION ALGORITHM

This algorithm starts with a single large group and then iteratively performs segmentation until a certain number of clusters is obtained. It is attractive to initialize K-means under this criterion; however, it becomes complicated to determine which group to segment and how to perform this segmentation. Thus, it is practical to indicate that this algorithm is independent of K-means because of its structure. A proposal based on this algorithm is the three-level K-means algorithm, which performs clustering in two phases. Fewer clusters than those predefined by k are created, the clusters with less variation are subdivided, and then k-means is applied traditionally. The complexity of this approach makes it practically an independent algorithm and not an initialization aid [57].

Table 2, condenses the description of the Initialization Methods of the previous paragraphs, with the intention to give in a very general way the Heuristics that constitute them and their relevance with the references addressed in the literature since it will be useful for the following analyses.

Initialization Methods	Heuristic description	References
Rand-C	1. Random selection of K points from the data set as the initial centroids. 2. Apply the following steps of the K-means algorithm and repeat from Rand-C initialization if necessary.	[36]-[37]-[38]
Rand-P	1. The data set is randomly partitioned into K groups. 2. The centroids are calculated from this partition of K groups. 3. Apply the following steps of the K-means algorithm.	[11]-[39]
Farthest Point	1. Selection of a random point as the first centroid. 2. The next centroid is chosen, considering the maximum distance from the previous one. 3. Apply K-means when all centroid assignments are available.	[40]-[41]-[42]-[43]-[44]-[45]- [46]
Classification Heuristic	1. Selection of the first K elements. 2. Such selection is contemplated while discarding points closer to the chosen centroids. 3. Each (N/k) represents a uniform partition.	[40]-[47]-[48]
Projection Heuristic	1. The data are projected into a smaller dimensional space using PCA (Principal Component Analysis). 2. According to this dimensional arrangement, centroids with Ran-C or Farthest Point are selected.	[49]-[50]

	3. The following steps of the K-means algorithm are applied.	
Density Heuristic	1. Identify high density concentration of the dataset using the DBSCAN algorithm. 2. Selection of centroids considering the most dense areas. 3. Apply the following steps of the K-means algorithm.	[51]-[56]
Division Algorithm	1. Selection of the largest cluster. 2. Application of larger cluster segmentation. 3. Two random points are considered for segmentation. 4. K-means is applied within the segmented cluster.	[57]

Table 2. Description of Initialization Methods

4. K-MEANS APPLICATION IN THE FIELD OF INDUSTRY

This segment of the paper selectively alludes to the applications of K-means in the industrial sector and the processes that require a clustering analysis under the principle of this algorithm. The intention of verifying such information in the literature is only to complement the review of initialization methods and performance measures in order to identify the opportunities presented by the results of the application of the algorithm in real operation scenarios. Therefore, the cases related to the implementation of K-means in a direct way in processes of the productive sector will be analyzed, in parallel, the contributions that include one or several algorithms, heuristics, methods, or additional techniques to the natural implementation of K-means, to improve its performance.

A clustering algorithm is called the possibilistic fuzzy k-modes algorithm (PFKM). It also implements three metaheuristics to increase the clustering performance: a genetic algorithm (GA), a particle swarm optimization (PSO), and the sine-cosine algorithm (SCA). Three clustering algorithms are proposed: the GA-PFKM, PSO-PFKM, and SCA-PFKM algorithms. The performance of the algorithms is compared with that of the classical FKM algorithm using two indices: sum of squares error (SSE) and accuracy. Experimental results show that the PSO-PFKM and SCA-PFKM algorithms perform better for most data sets [50]. An application in the healthcare sector that combines K-means to refine data with Genetic Algorithms and Automatic Support Vectors as information classification is useful to generate predictive diagnoses of diabetes, contemplating factors such as blood pressure, glucose concentration, age, and body mass index, among others. The combination that integrates the refinement and clustering of information grants K-means functionality results above 98%; in contrast to other hybrid algorithms, the evaluation metrics were calculated based on percentage indexes such as sensitivity, specificity, as well as negative and positive predictions [58].

The application of K-means is also present in the energy industry, under the need to control the randomness and intermittency of the operation of photovoltaic energy through a prediction model of its power. The factor that directly impacts the problem is the climatic changes. For this case, an initialization method is proposed: K-means++, in combination with the similar day approach and a short-term memory network. The proposal yields results above other hybrid algorithms, with improvements of up to 70% in the reduction of the stochastic aspects of the problem. The performance evaluation of the algorithm focuses on the Mean Absolute Error, Mean Square Error, and Mean Absolute Deviation [59]. A hybrid algorithm integrates k-means unsupervised learning techniques with continuous swarm intelligence metaheuristics. In this study, the optimization of a concrete and steel composite bridge with a box girder with cost and CO2 emissions as objective functions was performed. The results show that the hybrid proposal outperforms the designed algorithms [60].

The impact of applying K-means is reflected even in SME's under a strategic planning model composed of four main guidelines: acquisition, pre-processing, processing, and evaluation. The application is based on the fulfillment of strategic objectives and the detection of opportunities through a SWOT Matrix with the support of experts. The implementation was in a frozen food distribution company. Specifically, in the processing stage, the clustering is reflected, according to the opinion of the experts, that condenses the aspects of opportunity with certain similarities to make decisions in the future in different problems. The algorithm is applied naturally; however, its performance is evaluated by assigning different clusters. It can be contemplated that the performance of several groups is evaluated by the percentage of precision of 87% above other tests with different numbers of groups [61].

A similar case to the previous one is presented in the oil industry, where K-means and hierarchical clustering are used to classify areas of opportunity in selecting candidate mature wells for productive intervention. The Maxmin initialization method and Euclidean distances are highlighted [62]. A general binarization algorithm is called the K-means Transition Algorithm (KMTA). KMTA uses the K-means clustering technique as a learning strategy to perform the binarization process. This mechanism was applied to the Cuckoo Search and Black Hole metaheuristics to solve the set covering problem (SCP). To demonstrate the efficiency of the proposal, Set Covering benchmark instances from the literature show that KMT competes with state-of-the-art algorithms [63]. Some contributions suggest further performance measures, for example, the impact of data normalization as a weighting factor; similarly, the effect that this has on Euclidean distance is indicated, assuming some cluster overlap analysis in a case study related to the refinery industry [64].

With the intention of improving the measurement and monitoring of workers' job satisfaction and their performance levels, K-means is applied in conjunction with an artificial neural network (ANN). The entire improvement process is based on control, using the PDCA cycle and strategic plans [65]. With the use of genetic algorithms and K-means, an innovative portfolio mechanism is proposed that directly benefits the stock market by calculating the trading return and performance in terms of risk [66]. Concerning multimodal optimization in manufacturing processes, a new approach based on K-means is suggested for clustering initial points, then optimization is carried out based on radii surrounding each cluster and iteratively obtain multiple optimal results; this new approach refers to density and division heuristics [67]. Finally, to improve the supply chain, economic and environmental aspects between customer and supplier. MOSS software is proposed to balance the total cost, the percentage of defective materials, delivery delays and the consideration of improvements in environmental aspects. The model works based on the NSGA-III algorithm; subsequently, K-means is applied to select representative and important solutions according to the Pareto principle and the Silhouette Index [68]. In the applications of this paragraph, natural initialization methods of the K-means algorithm are considered; what increases its performance is the combination with other algorithms and heuristics. The next section discusses the results of the relationship between initialization methods and performance measures, considering industrial applications.

4.1 ANALYSIS OF INITIALIZATION METHODS IN INDUSTRIAL APPLICATIONS

The literature review analyzed in previous sections presents the results of the initialization methods that best suit the K-means algorithm's performance factors. Previously, some observations have been made regarding the best performance of the algorithm, especially those that depend directly on a better initialization method or repeat the algorithm a certain number of times, considering randomness, generating subgroups, and other strategies to seek better performance. This section discusses applications of K-means in the industrial field based on the initialization methods shown previously.

Before performing the analysis of the application of the K-means algorithm in the field of industry, it is important to visualize the contents of Table 3, which summarizes the information from the literature reviewed in sections 2 and 3 of this article. The above refers to the association that exists between Initialization Methods and Performance Measures. It is marked with a "✓" if the Initialization Method favors the Performance Measure in which it intersects and with an "x" if the opposite is the case. For example, it is perceived that Overlap is tractable with the implementation of the Farthest Point, Classification Heuristic, Projection Heuristic, Density Heuristic, and Division Algorithm; on the contrary, Rand-C and Rand-P do not generate a positive impact on such Performance Measure. It is noticeable that there is no Initialization Method that helps to improve the Performance Measures visibly, it is advisable to implement them in parallel or to use an independent algorithm. This situation is raised in the analysis discussed in the following paragraphs, where the need to use hybrid Initialization Methods is perceived as a product of K-means applications in the industrial field.

Initialization Methods	Performance Factors			
	Overlap	No. Clusters	Dimension	Unbalance
Rand-C	x	x	x	x
Rand-P	x	x	x	x
Farthest Point	✓	x	x	✓
Classification Heuristic	✓	x	x	x
Projection Heuristic	✓	x	x	x
Density Heuristic	✓	✓	x	x
Division Algorithm	✓	x	x	x

Table 3. Analysis scheme between initialization methods and performance factors of the K-means algorithm

First, it is important to highlight some applications shown in Table 4, which measure the performance from a functionality and stability perspective of the K-means algorithm applied naturally to the oil industry, i.e., without combining it with another algorithm, using only the Maxmin and Rand-C initialization methods [62]. For the case of the performance measure of sensitivity and specificity, the presence of independent algorithms can be considered, i.e., that work in a hybrid way and yield a single result, as in the case of the combination of K-means with Support Vector Machine (SVM) and Genetic Algorithm (GA) in the healthcare industry [69]-[58].

The repetition of the number of groups and Random Centroids (Rand-C) to identify the best K-value is reflected in the natural application of K-means in the transportation services industry [61]. The performance measure evaluated through the error deviation is perceived when applying K-means with Transition Algorithm Black Hole (TA-BH) as an independent algorithm in the transportation industry [63]. Similarly, alluding to error measures, the sum of squares of error and accuracy is considered in the healthcare industry by implementing an independent algorithm, K-means, Particle Swarm Optimization (PSO), Sine-Cosine Algorithm (SCA), and GA [57].

Likewise, following the guideline of the previous method, the performance of K-means in combination with Grey Relational Analysis (GRA), Cosine, and Long Short-Term Memory (LSTM) in the energy industry is measured as an independent algorithm, applying Furthest Point Initialization as an initialization method, considering the mean absolute, quadratic, normalized and mean absolute deviation error [59]. The initiative to use a performance measure based on maximum and minimum results and standard deviation is reflected in the construction industry's independent application of K-means with Cuckoo Search (CS) [60]. Finally, a new performance measure, the level of normalization and proportional influence that contributes to the Overlap of groups in K-means, is integrated naturally into the refinery industry.

The study proposed by [70] focuses on the implementation and evaluation of the K-means algorithm for data reduction in the prediction of coastal bed evolution using an annual data set of wave features at a coastline. The K-means method was implemented with several configurations, including data normalization and initial centroid selection using the K-means++ algorithm. The results showed that the use of the K-means algorithm allows the selection of representative wave conditions that significantly reduce the computational effort without sacrificing accuracy in predicting changes in the coastal bed. In general, it was concluded that the K-means initialization method used is a valuable tool, facilitating more accurate and efficient predictions of the morphological evolution of the seabed.

The paper developed by [71] presents the implementation of the K-means algorithm using an innovative approach called Seed-Detective, which combines the ModEx method to generate high-quality initial seeds. This approach seeks to overcome the limitations of random Rand-C selection in K-means, which often results in low-quality clusters. The implementation was performed on several datasets obtained from the machine learning repository. The use of high-quality initial seeds not only reduces computational time but also improves clustering accuracy, providing a more efficient method.

The research of [72] presents an improved implementation of the K-means algorithm for the analysis of users' electrical behavior, addressing common problems such as sensitivity to initial clustering centers and the need to specify the number of clusters. A method is proposed that uses a nearest neighbor density matrix, constructing a K-D tree to improve neighbor search and select more representative initial centers. The implementation focuses on the classification of power users in order to optimize dynamic load management in smart power systems.

The study generated by [73] implements a recommendation system for e-commerce based on an improved K-means algorithm, which combines this with a Genetic Algorithm (GA) to optimize the selection of initial centers and improve the stability of the clustering results. This method focuses on product information management, addressing the limitations of conventional K-means, which suffer from handling large volumes of data. This approach was implemented in the context of e-commerce, where efficient data management and personalized recommendations are crucial to improve user experience and increase sales.

The proposal documented by [74] presents an initialization method for the W-k-means algorithm, which is based on the selection of attribute weights using two approaches: the coefficient of variation and entropy. These methods seek to improve the quality of clustering by assigning weights that reflect the variability of the data, allowing for more accurate segmentation compared to random initialization. The area of implementation of the algorithm was in image processing, specifically in color image segmentation, demonstrating its practical application in this field.

The work developed by [75] examines an encryption processing method for accounting data based on the K-means Clustering Algorithm (KMCA), implemented in the context of accounting information security. For the initialization of the K-means algorithm, initial cluster centers are selected using Euclidean distance, and the accounting information is classified into specific categories such as revenue, cost, and profit. The research highlights the importance of KMCA in the classification and protection of accounting data, thus addressing the growing challenge of information security in a digital environment.

The paper prepared by [76] presents a strategy for clustering electric vehicles based on an improved k-means algorithm implemented to mitigate wind power fluctuations. An algorithm called IDOA-KM is used, which optimizes the k-means clustering centers using the Improved Dingo Optimization Algorithm (IDOA). This method addresses dependence and sensitivity to initial centers, improving initial population diversity and balancing exploration and exploitation. The research was carried out in the context of the laboratory-level power system, using predicted data from a wind farm over 24 hours.

The paper constructed by [77] describes an initialization method for the K-means algorithm that uses Voronoi diagrams to select initial cluster centers. The proposed technique selects centers from points on the edges of Voronoi circles of larger radius, associating this method with the Furthest Point, which allows a better initial distribution and improves the quality of clustering. The implementation was performed using datasets from the UCI repository, covering various dimensions and characteristics, demonstrating the versatility of the approach in different contexts.

The study designed by [78] presents an initialization method for the customized K-means algorithm applied to the optimization of the location of charging stations for electric vehicles. A penalty term is introduced in the K-means algorithm that avoids the relocation of stations near points with low charging demand, thus improving the distribution and efficiency of the stations. This approach was implemented in the context of the energy system.

The paper published by [79] proposes an optimized model of public opinion monitoring in social networks based on the K-means algorithm enhanced with Particle Swarm Optimization (PSO). This approach optimizes the selection of initial centroids, overcoming local optima problems and improving the clustering accuracy. When implemented in social networks, the model stands out for efficiency in spreading emotions, making it an effective tool for analyzing public opinion trends.

In the client power consumption analysis designed by [80], an optimized version of the K-means algorithm was used. For centroid initialization, the Improved Harris Hawk Optimization (IHHO) algorithm was implemented, which increased the initial diversity and convergence. The optimal number of clusters was determined by the Davies-Bouldin index (DBI). Applied to customers in an electric grid, the IHHO-K-means classified users into four groups according to their electrical behavior, highlighting high-value and high-potential customers.

It is important to mention that each author implements and uses K-means in a particular way, depending on the problem to be solved and the performance measures to be considered. Table 4 shows the information contained in the discussion of the previous contributions, with the purpose of showing the applications of the K-means Algorithm in the industrial field. In addition, the initialization method used is integrated to conclude which of these methods are most used. According to the above, the recurrence of the independent algorithms can be seen when the characteristics of the data set of the problem are of a high degree of complexity.

Method	Industry	Initialization Method					References
		Rand-C	Rand-P	Furthest Point	Classification Heuristic	Independent Algorithm	
K-means/SVM	Health care					x	[69]
K-means/GA/SVM	Health care					x	[58]
K-means	Transportation Service	x			x		[61]
K-means/TA-BH	Transportation					x	[63]
K-means	Petroleum	x		x			[62]
K-means/PSO-SCA	Health care					x	[57]
K-means/CSI-SC-CS	Construction					x	[60]
K-means/K-NN	Refinery					x	[64]
K-means/GRA/Coseno/LSTM	Energy			x		x	[59]
K-means/K-means++	Oceanography			x		x	[70]
K-means/Seed-Detective	Computer Science					x	[71]
K-means/KD tree	Electronics					x	[72]
K-means/GA	Marketing					x	[73]
W-K-means	Image processing					x	[74]
K-means	Accounting	x	x				[75]
K-means/IDOA	Energy					x	[76]
K-means/Voronoi	Computer Science			x		x	[77]
K-means/Customized	Energy					x	[78]
K-means/PSO	Informatics					x	[79]
K-means/IHHO	Energy					x	[80]

Table 4. Comparative analysis of the K-means in industrial applications and initialization methods

According to Table 4, it can be interpreted as a tendency to use independent algorithms for the initialization of K-means derived from a different domain of data set, where the main difference is the complexity and the amount of information to be analyzed. Also, the use of a conventional K-means Algorithm can be identified in [61]-[62]-[75], since it simply sought a basic clustering without the intention of improving the initialization. In some cases, the incorporation of auxiliary algorithms, such as evolutionary, supervised learning, and collective intelligence, among others, is used, making K-means a hybrid whose main characteristic is the optimization of the parameters in the initialization cover.

5. CONCLUSIONS

Reviewing scientific contributions on K-means performance using the initialization method opens opportunities to suggest new methods to obtain the best results from this algorithm. Cluster overlapping is a natural problem in all data clusters and, simultaneously, difficult to solve. There are two scenarios in K-means; the overlapping can be taken care of directly by the algorithm without needing an initialization method. The second is that if there is low Overlap, a good clustering can be subjected to an initialization method; however, it must be quite effective to overcome the weakness generated in K-means, derived from the considerable separation of the clusters and the possible high number of groups. In aspects related to cluster imbalance, the problem grows since most of the initialization methods are not efficient. Dimensionality is not a problem associated with the initialization method's performance; it only impacts the superposition of groups since it is inherent to them in a certain way. The possibility of analyzing K-means from a global perspective, which complies with all performance measures based on hybrid algorithms associated with Swarm Intelligence Optimization, is not ruled out. Supported by the analysis of K-means applications, where different performances are visualized.

In retrospective, this research explores the most recurrent performance measures presented in the K-means Algorithm, as well as the Initialization Methods that affect or benefit them. It was found that exhaustive search methods are generally deficient in all four performance measures. Therefore, it is recommended to discard Rand-C and Rand-P to initialize K-means, which can be useful in the case of repeating the algorithm more than 100 times; however, it increases the computational cost. The Farthest Point, manages to improve the Overlap and imbalance of the groups quite well, as these are the most critical measures for determining the success of K-means. The other Heuristic options, widen the range of possibilities of success by correctly covering the Overlap and the determination of the number of groups, as is the case of the Density Heuristic. The analysis of the application of the K-means Algorithm in the field of industry, made it clear that in the last 10 years, the use of classical Initialization Methods is no longer sufficient to overcome the problems of large data size and variability since improving all performance measures becomes an optimization problem with increasingly complex solutions. It is necessary to resort to Metaheuristics and the design of new methods that turn K-means into a hybrid algorithm, dependent on factors that directly support its heuristics.

REFERENCES

[1] Celebi, M. E., Kingravi, H. A. & Vela, P. A., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications,* 40(1),pp.200-210.https://doi.org/10.1016/j.eswa.2012.07.021.

[2] Raykov, Y. P., Boukouvalas, A., Baig, F. & Little, M. A., 2016. What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm. *Plos one.* 11(9), https://doi.org/10.1371/journal.pone.0162259.

[3] Pitafi, S., Anwar, T. & Sharif, Z., 2023. A Taxonomy of Machine Learning Clustering Algorithms, Challenges, and Future Realms. *Applied Sciences,* 13(6), p. 3529. https://doi.org/10.3390/app13063529.

[4] YiFei, L. et al., 2023. Structure damage identification in dams using sparse polynomial chaos expansion combined with hybrid K-means clustering optimizer and genetic algorithm. *Engineering Structures.* https://doi.org/10.1016/j.engstruct.2023.11589.

[5] Zhang, H. & Peng, Q., 2022. PSO and K-means-based semantic segmentation toward agricultural products. *Future Generation Computer Systems,* pp. 82-87. 1 https://doi.org/0.1016/j.future.2021.06.059.

[6] Bai, L., Cheng, X., Liang, J. & Shen, H. G. Y., 2017. Fast density clustering strategies based on the k-means algorithm. *Pattern Recognition,* pp. 375-386. https://doi.org/10.1016/j.patcog.2017.06.023.

[7] García, J. et al., 2018. *Ciencia de Datos técnicas analíticas y aprendizaje estadístico en un enfoque práctico.* Publicaciones Altaria, S.L. ed. Bogotá: Alfaomega.

[8] Sridevi, S., Parthasarathy, S. & Rajaram, S., 2018. An effective prediction system for time series data using pattern matching algorithms. *International Journal of Industrial Engineering: Theory, Applications and Practice,* 25(2), pp. 123-136, ISSN 1943-670X.

[9] Zou, P., Rajora, M. & Liang, S., 2019. Multimodal optimization of job-shop scheduling problems using a clustering-genetic algorithm based approach. *International Journal of Industrial Engineering: Theory, Applications and Practice,* 26(5). https://doi.org/10.23055/ijietap.2019.26.5.4043.

[10] Zhao, Q. & Fränti, P., 2014. WB-index: A sum-of-squares based index for cluster validity. *Data & Knowledge Engineering,* pp. 77-89. http://dx.doi.org/10.1016/j.datak.2014.07.008.

[11] Peña, J., Lozano, J. & Larrañaga, P., 1999. An empirical comparison of four initialization methods for the K-Means algorithm. *Pattern Recognition Letters,* 20(10), pp. 1027-1040. https://doi.org/10.1016/S0167-8655(99)00069-0.

[12] Fränti, P. & Sieranoja, S., 2018. K-means properties on six clustering benchmark datasets. *Applied Intelligence,* 48(12), p. 4743–4759. https://doi.org/10.1007/s10489-018-1238-7.

[13] Jain, A. K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters,* 31(8), pp. 651-666. https://doi.org/10.1016/j.patrec.2009.09.011.

[14] Morissette, L. & Chartier, S., 2013. The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology,* 9(1), pp. 15-24. http://dx.doi.org/10.20982/tqmp.09.1.p015.

[15] Liang, J., Bai, L., Dang, C. & Cao, F., 2012. The K-means-type algorithms versus imbalanced data distributions. *IEEE Trans. Fuzzy Syst. ,* Volume 20, pp. 728-745. https://doi.org/10.1109/TFUZZ.2011.2182354.

[16] Melnykov, I. & Melnykov, V., 2014. On K-means algorithm with the use of Mahalanobis distances. *Statistics & Probability Letters,* Volume 84, pp. 88-95. https://doi.org/10.1016/j.spl.2013.09.026.

[17] Melnykov, V., Michael, S. & Melnykov, I., 2015. Recent developments in model-based clustering with applications. In: *Partitional Clustering Algorithms.* s.l.:Celebi, M. Emre, pp. 1-39. https://doi.org/10.1007/978-3-319-09259-1_1.

[18] Ahmed M, Seraj R, Islam S. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*. 2020; 9(8): p. 1295.

[19] Ghazal T, Hussain M, Said R, Nadeem A, Hasan MK, Ahmad M, et al. Performances of K-Means Clustering Algorithm with Different Distance Metrics. *Intelligent Automation and Soft Computing*. 2021; 30(2): p. 735-742.

[20] Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*. 2023; 622: p. 178-210.

[21] Abdulnassar AA, Nair LR. Performance analysis of Kmeans with modified initial centroid selection algorithms and developed Kmeans9+ model. *Measurement: Sensors*. 2023; 25: p. 100666.

[22] Li, F. et al., 2023. An overlapping oriented imbalanced ensemble learning algorithm with weighted projection clustering grouping and consistent fuzzy sample transformation. *Information Sciences,* Volume 637, p. 118955. https://doi.org/10.1016/j.ins.2023.118955.

[23] Khanmohammadi, S., Adibeig, N. & Shanehbandy, S., 2017. An improved overlapping k-means clustering method for medical applications. *Expert Systems with Applications,* Volume 67, pp. 12-18. https://doi.org/10.1016/j.eswa.2016.09.025.
[24] De Martino, G., Pio, G. & Ceci, M., 2023. Multi-view overlapping clustering for the identification of the subject matter of legal judgments. *Information Sciences,* Volume 638, p. 118956. https://doi.org/10.1016/j.ins.2023.118956.

[25] Viloria, A. & Pineda Lezama, O. B., 2019. Improvements for Determining the Number of Clusters in k-Means for Innovation Databases in SMEs. *Procedia Computer Science,* Volume 151, pp. 1201-1206. https://doi.org/10.1016/j.procs.2019.04.172.

[26] Gupta, A., Datta, S. & Das, S., 2018. Fast automatic estimation of the number of clusters from the minimum inter-center distance for k-means clustering. *Pattern Recognition Letters,* Volume 116, pp. 72-79. https://doi.org/10.1016/j.patrec.2018.09.003.

[27] Tian, K. et al., 2019. Segmentation of tomato leaf images based on adaptive clustering number of K-means algorithm. *Computers and Electronics in Agriculture,* Volume 165, p. 104962. https://doi.org/10.1016/j.compag.2019.104962.

[28] Saha, J. & Mukherjee, J., 2021. CNAK: Cluster number assisted K-means. *Pattern Recognition,* Volume 110, p. 107625. https://doi.org/10.1016/j.patcog.2020.107625.

[29] Ahn, S., Wang, X. & Lim, J., 2017. On unbalanced group sizes in cluster randomized designs using balanced ranked set sampling. *Statistics & Probability Letters,* Volume 123, pp. 210-217. https://doi.org/10.1016/j.spl.2016.12.007.

[30] Wang, X., Ahn, S. & Lim, J., 2017. Unbalanced ranked set sampling in cluster randomized studies. *Journal of Statistical Planning and Inference,* Volume 187, pp. 1-16. https://doi.org/10.1016/j.jspi.2017.02.005.

[31] Liang, X. et al., 2020. LR-SMOTE— An improved unbalanced data set oversampling based on K-means and SVM. *Knowledge-Based Systems,* Volume 196, p. 105845. https://doi.org/10.1016/j.knosys.2020.105845.

[32] Kinnunen, T., Sidoroff, I., Tuononen, M. & Fränti, P., 2011. Comparison of clustering methods: A case study of text-independent speaker modeling. *Pattern Recognition Letters,* 32(13), pp. 1604-1617. https://doi.org/10.1016/j.patrec.2011.06.023.

[33] Cao F, Liang J, Jiang G. An initialization method for the K-Means algorithm using neighborhood model. *Computers & Mathematics with Applications*. 2009; 58(3): p. 474-483.

[34] Zhang Y, Lan T, Li C, Cai W, Lin Z, Lin J. Holomorphic embedding method based Three-Phase power flow algorithm considering the sensitivity of the initial value. *International Journal of Electrical Power & Energy Systems*. 2024; 162: p. 110271.

[35] Li CS. Cluster Center Initialization Method for K-means Algorithm Over Data Sets with Two Clusters. *Procedia Engineering*. 2011; 24: p. 324-328.
[36] Forgy, E. W., 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics,* Volume 21, pp. 768-769.

[37] MacQueen, J. a. o., 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.* Berkeley(California): s.n., pp. 281-297.

[38] Norušis, M. J., 2011. IBM SPSS statistics 19 guide to data analysis. *Prentice Hall.*

[39] Thiesson, B., Meek, C., Chickering, D. M. & Hackerman, D., 1998. Leraning Mixtures of Bayesian Networks. *Technical Report MSR-TR-97-30.*

[40] Steinley, D. & Brusco, M. J., 2007. Initializing K-means Batch Clustering: A Critical Evaluation of Several Techniques. *Journal of Classification,* 24(1), pp. 99-121. https://doi.org/10.1007/s00357-007-0003-0.

[41] Chiang, M. M.-T., 2010. Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. *Journal of Classification,* 27(1), pp. 3-40. https://doi.org/10.1007/s00357-010-9049-5.

[42] Hämäläinen, J. & Kärkkäinen, T., 2016. Initialization of big data clustering using distributionally balanced folding. In: *ESANN 2016 Proceedings.* Finlandia: s.n., pp. 587-592.

[43] Cao, F., Liang, J. & Bai, L., 2009. A new initialization method for categorical data clustering. *Expert Systems with Applications,* 36(7), pp. 10223-10228. SBN 978-287587027-8.

[44] Erisoglu, M., Calis, N. & Sakallioglu, S., 2011. A new algorithm for initial cluster centers in k-means algorithm. *Pattern Recognition Letters,* 32(14), pp. 1701-1705. https://doi.org/10.1016/j.patrec.2011.07.011.

[45] Gingles, C. & Celebi, M. E., 2014. Histogram-based method for effective initialization of the k-means clustering algorithm. In: *The Twenty-Seventh International Flairs Conference.* L.A. USA: s.n., pp. 333-338.

[46] Arthur, D. & Vassilvitskii, S., 2007. K-means++ the advantages of careful seeding. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms.* s.l.:s.n., pp. 1027-1035. https://doi.org/10.1145/1283383.1283494.

[47] 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), pp. 100-108.

[48] Cao, F., Liang, J. & Jiang, G., 2009. An initialization method for the K-Means algorithm using neighborhood model. *Computers & Mathematics with Applications,* 58(3), pp. 474-483. https://doi.org/10.1016/j.camwa.2009.04.017.
[49] Yedla, M. & Rao, P. S. S. T., 2010. Enhancing K-means clustering algorithm with improved initial center. *International Journal of computer science and information technologies,* 1(2), pp. 121-125.

[50] Su, T. & Dy, J. G., 2007. In Search of Deterministic Methods for Initializing K-means and Gaussian Mixture Clustering. *Intelligent Data Analysis,* 11(4), p. 319 – 338. https://doi.org/10.3233/IDA-2007-11402.

[51] Al-Daoud, M. B. & Roberts, S. A., 1996. New methods for the initialisation of clusters. *Pattern Recognition Letters,* 17(5), pp. 451-455. https://doi.org/10.1016/0167-8655(95)00119-0.

[52] Redmond, S. J. & Heneghan, C., 2007. A method for initialising the K-means clustering algorithm using kd-trees. *Pattern Recognition Letters,* 28(8), pp. 965-973. https://doi.org/10.1016/j.patrec.2007.01.001.

[53] Rodriguez, A. & Laio, A., 2014. Clustering by fast search and find of density peaks. *American Association for the Advancement of Science,* 344(6191), pp. 1492-1496. https://doi.org/10.1126/science.1242072.

[54] Lemke, O. & Keller, B. G., 2018. Common Nearest Neighbor Clustering—A Benchmark. *Algorithms,* 11(2), p. 19. https://doi.org/10.3390/a11020019.

[55] Von Luxburg, U., 2010. Clustering stability: an overview. *Foundations and Trends® in Machine Learning,* 2(3), pp. 235-274. https://doi.org/10.48550/arXiv.1007.1075.

[56] Yu, S.-S.et al., 2018. Two improved k-means algorithms. *Applied Soft Computing,* Volume 68, pp. 747-755. https://doi.org/10.1016/j.asoc.2017.08.032.

[57] Kuo, R., Zheng, Y. & Nguyen, T. P. Q., 2021. Metaheuristic-based possibilistic fuzzy k-modes algorithms for categorical data clustering. *Information Sciences,* Volume 557, pp. 1-15. https://doi.org/10.1016/j.ins.2020.12.051.

[58] Santhanam, T. & Padmavathi, M., 2015. Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis. *Procedia Computer Science,* Volume 47, pp. 76-83. https://doi.org/10.1016/j.procs.2015.03.185.

[59] Bai, R., Shi, Y., Yue, M. & Du, X., 2023. Hybrid model based on K-means++ algorithm, optimal similar day approach, and long short-term memory neural network for short-term photovoltaic power prediction. *Global Energy Interconnection,* 6(2), pp. 184-196. https://doi.org/10.1016/j.gloei.2023.04.006.

[60] Martínez-Muñoz, D., García, J., Martí, J. & Yepes, V., 2022. Discrete swarm intelligence optimization algorithms applied to steel–concrete composite bridges. *Engineering Structures,* Volume 266, p. 114607. https://doi.org/10.1016/j.engstruct.2022.114607.
[61] Vásquez Rojas, C., Roldán Reyes, E., Aguirre y Hernández, F. & Cortés Robles, G., 2018. Integration of a text mining approach in the strategic planning process of small and medium-sized enterprises. *Industrial Management & Data Systems,* 118(4), pp. 745-764. https://doi.org/10.1108/IMDS-01-2017-0029.

[62] Galicia, H. D. P., Reyes, E. R. & Sheremetov, L., 2021. Candidate wells selection and ranking based on data mining and multi-criteria decision analysis techniques. *Arabian Journal of Geosciences,* Volume 14, pp. 17-27.

[63] García, J., Crawford, B., Soto, R. & Astorga, G., 2019. A clustering algorithm applied to the binarization of Swarm intelligence continuous metaheuristics. *Swarm and Evolutionary Computation,* Volume 44, pp. 646-664. https://doi.org/10.1016/j.swevo.2018.08.006.

[64] Niño-Adan, I., Landa-Torres, I., Portillo, E. & Manjarres, D., 2022. Influence of statistical feature normalisation methods on K-Nearest Neighbours and K-Means in the context of industry 4.0. *Engineering Applications of Artificial Intelligence,* Volume 111, p. 104807. https://doi.org/10.1016/j.engappai.2022.104807.

[65] Aktepe, A. & Ersoz, S., 2012. A quantitative performance evaluation model based on a job satisfaction-performance matrix and application in a manufacturing company. *International Journal of Industrial Engineering: Theory, Applications and Practice,* 19(6).

[66] Ahn, W., Cheong, D., Kim, Y. & Oh, K. J., 2019. Developing an enhanced portfolio trading system using k-means and genetic algorithms. *International Journal of Industrial Engineering: Theory, Applications and Practice,* 25(5). https://doi.org/10.23055/ijietap.2018.25.5.3688.

[67] Zou, P., Rajora, M. & Liang, S. Y., 2021. Obtaining multiple process parameter combinations using a supervised clustering-optimization approach. *International Journal of Industrial Engineering: Theory, Applications and Practice,* 27(4). https://doi.org/10.23055/ijietap.2020.27.4.3929.

[68] Toktas-Palut, P., Onan, K., Zahid Gürbüz, M. & Gülden-Özdemir, B., 2022. Moss software: a new tool for multi-objective green supplier selection. *International Journal Of Industrial Engineering - Theory, Applications, And Practice.* https://doi.org/10.23055/ijietap.2022.29.2.5903.

[69] Yilmaz, N., Inan, O. & Uzer, M. S., 2014. A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases. *Journal of Medical Systems,* 38(5), p. 48. https://doi.org/10.1007/s10916-014-0048-7.

[70] Andreas P., Tsoukala V. Evaluating and enhancing the performance of the K-Means clustering algorithm for annual coastal bed evolution applications. *Oceanologia*. 2024; 66(2): p. 267-285.

[71] Md Anisur R, Md Zahidul I, Terry B. ModEx and Seed-Detective: Two novel techniques for high quality clustering by using good initial seeds in K-Means. *Journal of King Saud University - Computer and Information Sciences*. 2015; 27(2): p. 113-128.

[72] Chen Y, Tan P, Li M, Yin H, Tang R. K-means clustering method based on nearest-neighbor density matrix for customer electricity behavior analysis. *International Journal of Electrical Power & Energy Systems*. 2024; 161: p. 110165.

[73] Zhang W, Wu Z. E-commerce recommender system based on improved K-means commodity information management model. *Heliyon*. 2024; 10(9): p. 29045.

[74] Hung WL, Chang YC, Lee ES. Weight selection in W-K-means algorithm with an application in color image segmentation. *Computers & Mathematics with Applications*. 2011; 62(2): p. 668-676.

[75] Wei Q. Accounting Data Encryption Processing Based on K-Means Clustering Algorithm. *Procedia Computer Science*. 2024; 247: p. 819-825.

[76] Yang Yu MLDCYHWL. Dynamic grouping control of electric vehicles based on improved k-means algorithm for wind power fluctuations suppression. *Global Energy Interconnection*. 2023; 6(5): p. 542-553.

[77] Reddy D, Jana PK. Initialization for K-means Clustering using Voronoi Diagram. *Procedia Technology*. 2012; 4: p. 395-400.

[78] Abdullahi MR, Lu QC, Hussain A, Tripura S, Xu PC, Wang S. Location optimization of EV charging stations: A custom K-means cluster algorithm approach. *Energy Reports*. 2024; 12: p. 5367-5382.

[79] Qi M, Zhao J, Feng Y. An Optimized Public Opinion Communication System in Social Media Networks Based on K-means Cluster analysis. *Heliyon*. 2024; p. 40033.

[80] Wu R. Behavioral analysis of electricity consumption characteristics for customer groups using the k-means algorithm. *Systems and Soft Computing*. 2024; 6: p. 200143.

ACKNOWLEDGMENTS

The team that carried out this research extends its most sincere and respectful gratitude to the Congreso Veracruzano de Investigación Científica y Desarrollo Tecnológico (COVEICYDET) since its support was indispensable for its publication in the journal DYNA.

In addition, we are grateful for the invaluable support of the Tecnológico Nacional de México (TecNM) and the Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) for the development of this research.