

K-Means Clustering: Implementation, Issues, and Approaches in WSN

Ab Wahid Bhat^{1*}, Abhiruchi Passi²

^{1, 2} Department of ECE, FET MRIIRS, Faridabad, Haryana, India

*Corresponding Author Email: abwahid1994@gmail.com

Abstract

In wireless Sensor Networks (WSNs), energy optimization and efficient routing are NP-Hard problems. Organizing sensor nodes into small clusters with a local coordinator for the collection, aggregation, and transmission of data can reduce energy consumption and prolong the lifetime of WSNs. However, the uniform and balanced clustering of nodes is a challenging task in designing an energy-efficient clustering and routing protocol. One of the simplest yet efficient clustering algorithms that can be employed in WSNs is the K-means algorithm. The paper presents a detailed discussion on the implementation of the K-means algorithm, limitations, and solutions that can be adopted to enhance its performance. The primary objective is to provide a detailed insight into the various approaches employed in WSNs for addressing the limitations of the K-means algorithm. In addition to it, different metrics for evaluating the clustering efficiency of an algorithm are discussed in detail. A tabular review of the characteristics of various modified algorithms, which combine K-means and probabilistic, deterministic, and metaheuristic techniques, is also presented at the end of the work.

Keywords

Approaches, clustering, K-means algorithm, limitations, Wireless sensor networks.

INTRODUCTION

Wireless sensor networks are characterized by a massive and random deployment of sensor nodes in an area of interest. The sensor nodes gather the desired information and then transmit it to the sink or base station. The energy-efficient transmission of information in WSNs is a challenging task, as the energy consumption of sensor nodes is directly dependent upon the communication distance. Also, due to random deployment, several sensor nodes gather the same information, and communicating it directly to the base station can cause increased energy consumption, along with network congestion and redundancy. Therefore, an energy-efficient mechanism called clustering is employed that partitions the network area into small clusters with each having a cluster head and several member nodes. The member nodes collect the desired information and forward it to the cluster head. Usually, the intra-cluster communication distance is small compared to the direct communication, which results in reduced energy consumption in member nodes. The cluster head then processes the data, aggregates it to reduce redundancy, and transmits it as a single message to the base station through single-hop or multi-hop communication. Several algorithms have been proposed for the clustering of wireless sensor networks, one being the K-means.

K-MEANS CLUSTERING ALGORITHM

The conventional K-means is an unsupervised iteration-based refinement algorithm that involves the partitioning of the data into k clusters and assigning each data point to one of the k clusters based on a distance parameter

from a randomly chosen central point called a centroid in each cluster [1],[2]. The objective of K-means is to reduce the intra-cluster distance (usually the square Euclidean Distance) of different data points from the centroid point for efficient clustering. The conventional K-means algorithm is based on several steps, as illustrated below:

Step I: Initialization

Select the value of parameter k (where $k \leq n$), based on which the entire data set of n elements will be partitioned into k clusters.

Randomly initialize k number of points called centroids.

Step II: Assignment

Calculate the distance of all the other non-k data points w.r.t all the k-points or centroids. The distance can be calculated using formulas like the square Euclidean distance, Manhattan distance, or cosine distance measure. The selection of a particular distance formula is subject to the application. Usually, the square Euclidean distance is used. The expression for calculating the square Euclidean distance of a data point from the centroid point (in a 2D XY plane) is

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]} \quad (01)$$

Where (x_1, y_1) are the coordinates of one of the centroid points and (x_2, y_2) are the coordinates of the data point.

Depending upon the distance, assign each of the data points to the centroid having the least value of d (or the nearest centroid point). Every data point is assigned to a centroid point, even if the distance is the same w.r.t. two centroid points.

Step III: Re-initialization

Calculate the mean of all the data points within a particular

cluster using the expression.

$$C_i = \frac{1}{N_i} \sum_{y=1}^{N_i} x_y \quad (02)$$

Where C_i ($i \in k$) is the centroid point having N number of associated data points with it.

Depending upon the value of C_i reinitializing the centroid points, and repeat steps II and III until the C_i does not change or become stable.

The main objective of the k -means algorithm is to minimize the total within-cluster sum of squares (WCSS) or the inter-cluster variance. This ensures that all the similar data points are clustered within a single cluster.

The expression for WCSS for a particular cluster C_i , or intra-cluster variance, is given by

$$WCSS(C_i) \sum_{x_j \in C_i} \| (x_j - \mu_i) \|^2 \quad (03)$$

The within-cluster sum of squares (WCSS) for all the clusters or the total variance is given by

$$WCSS(C_i; 1 \leq i < k) = \sum_{i=1}^k \sum_{x_j \in C_i} \| (x_j - \mu_i) \|^2 \quad (04)$$

The Lower the value of WCSS, the more compact and distinct the clusters are. After every iteration, the value of WCSS decreases until an optimum value of each centroid point is obtained. This is termed as convergence in the k -means algorithm. However, it does not always guarantee convergence to a global optimum.

Issues in the K-means Algorithm

While the K-means algorithm is an excellent technique for clustering small and simple data sets, its performance degrades in the case of large and complex data sets. Besides, the k -means algorithm has the following limitations:

1. The selection of the optimum number of clusters, i.e., the value of K , can significantly affect the results.
2. The initialization of centroid points such that the intra-cluster variance is minimized can be an NP-hard problem. The wrong selection of the centroid point can slow the convergence of the algorithm, and it can also get trapped in a local optimum. In addition to this, the final shape of the clusters can also vary with the selection of different distance formulas.
3. The time complexity of each iteration of the k -means algorithm is $O(K * n * t_{dist})$.

Where K = the number of clusters (centroids), n = the number of data points or objects, and t_{dist} is the time taken to calculate the distance between a set of data points. In general terms, the time complexity of K-means is written as $O(n^2)$ [3].

4. The K-means algorithm is too sensitive to outliers. The presence of outliers can be detrimental to efficient clustering.
5. The K-means algorithm performs poorly with variable shape, size, and density of clusters.

Selection of the Optimum Number of Clusters

The K-means algorithm is a simple and efficient clustering technique, provided an optimum value of K , or the number of clusters, is predefined or already known. The entire

performance and complexity of the method can vary depending upon the selection of K . An unfitting value of K can result in over-simplification or over-complication. Therefore, several methods have been proposed [4] to find the optimum value of K .

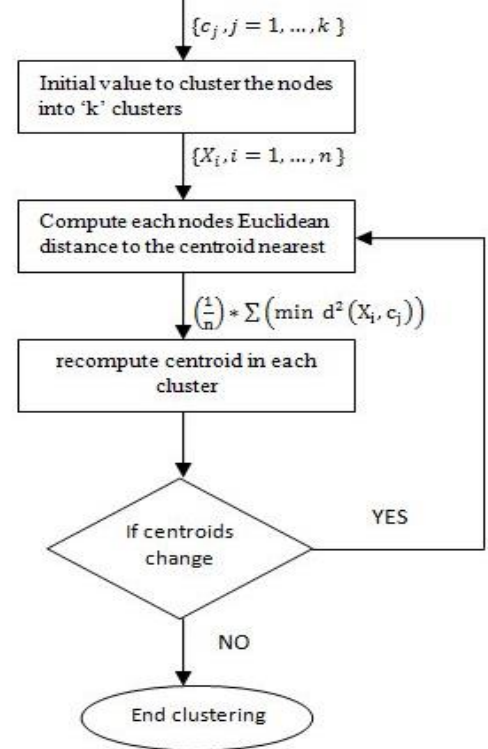


Figure 1. Flow chart of K-means Algorithm [5]

1. **Elbow method:** It is one of the prominent methods to determine the K-optimum. It is based on a graphical representation of the WCSS and the value of K . Initially, the value of K is chosen as 1, and the WCSS is calculated. After that, the value of k is incremented in steps, and the corresponding values of WCSS are noted. A graph between the value of K (X-axis) and the value of WCSS (Y-axis) is plotted, and by visual inspection, the point is identified where the reduction of WCSS starts to slow down, representing the optimum value of K for that data [Figure 2a]. However, the reliability of the elbow method decreases as the value of WSCC reduces smoothly against the increasing value of K , resulting in ambiguity in the elbow point [Figure 2b].
2. **Gap Statistics:** It is a statistical method of determining the optimum number of clusters (K value). It involves a contrast between the intra-cluster variations of the actual data set and those of a randomly generated reference distribution with no obvious clustering (bootstrapping method). The optimum value of K is obtained where the two fall farthest apart [Figure 3a] and the Gap has a maximum value [Figure 3a]. The expression for determining the gap statistic is

$$\text{Gap}(k) = \frac{1}{B} \sum_{b=1}^B \log(W_{kb}) - \log(W_k) \quad (05)$$

Where W_k is the WCSS of the original data & W_{kb} is the WCSS of the reference data set.

3. Silhouette Analysis: It is a method to determine how well the data or objects are clustered. It involves a procedure to measure the intra-cluster cohesion or the similarity of a data point w.r.t. the assigned cluster and its dissimilarity w.r.t. the other clusters [6]. It is generally expressed in terms of silhouette score, having a value that varies between -1 and +1 (Procedure discussed in a later section). The optimum K value is obtained w.r.t. the maximum silhouette score in a given data set.

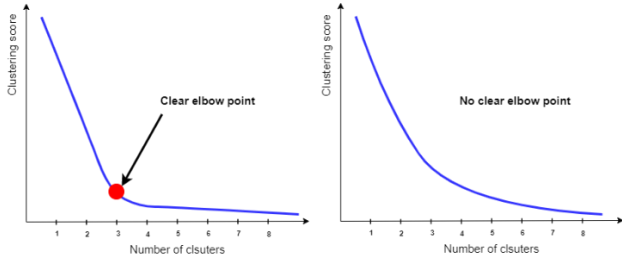


Figure 2a and 2b. Elbow method

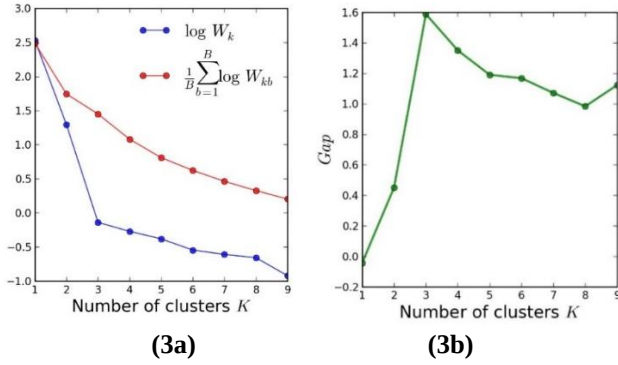


Figure 3a and 3b. Gap Statistics [Data science Lab]

Cluster Evaluation Metrics

The efficiency of any clustering algorithm can be measured in terms of various evaluation metrics. These metrics indicate how efficiently similar objects or data points are clustered together and how distinct the data points are in different clusters. The Following are some of the metrics:

- 1. Inertia:** It is a measure to determine the compactness of clusters. It is also defined as the within-cluster Sum of the Squares (WCSS) and is given by expression (03). Usually, a lower value indicates good clustering.
- 2. Silhouette Score:** It is used to measure the intra-cluster cohesion and inter-cluster separation. The expression for the Silhouette score for a data point i is

$$\text{Silhouette Score}, S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (06)$$

where $a(i)$ is the average intra-cluster distance between element i ($i \in C_i$) and all other data points in the cluster C_i .

$$a(i) = \frac{1}{C_i - 1} \sum_{j \in C_i, i \neq j} d(i, j) \quad (07)$$

$b(i)$ is the minimum of the average distances between element i and its inter-cluster neighbours.

$$b(i) = \min_{l \neq k} \frac{1}{C_k} \sum_{j \in C_k, i \neq j} d(i, j) \quad (08)$$

The Silhouette score ranges from -1 to +1, indicating a complete mismatch and perfectly matched assignment within the cluster, as shown in Table 1 [7]. Any value lying between 0.51 and 0.70 is acceptable.

Table 1. Silhouette score interpretation

Silhouette Score	Interpretation
0.71 – 1.00	Perfectly Clustered
.51 - 0.70	Well Clustered
0.26 – 0.50	Inadequately Clustered
≤ 0.25	No clustering

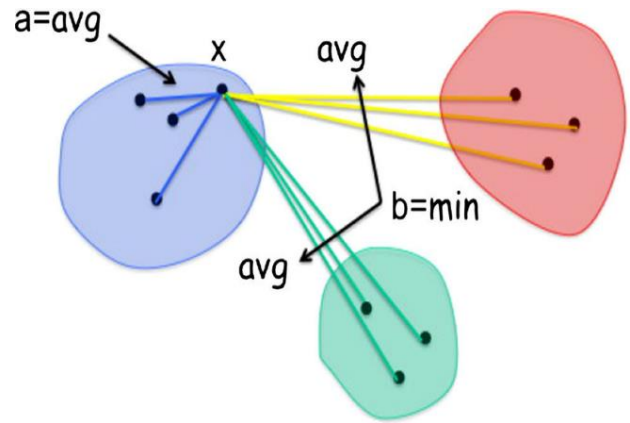


Figure 4. Silhouette score [8].

- 3. Jacard Index:** It is a measure to determine how well separated the clusters are. For two data sets A and B, it is represented as the ratio of Shared elements (intersection) to the overall elements (Union). The Jacard Index can range from 0 to 1, indicating no overlap and complete overlap of elements.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (09)$$

$$J(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (10)$$

- 4. Dice Score:** Like the Jacard index, it is a similarity detection metric. It weighs the similarities more than the dissimilarities and is calculated as the ratio of the union to the disjoint union.

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (11)$$

APPROACHES TO K-MEANS CLUSTERING IN WSNS

The node placement and energy efficiency optimization problems in WSNS are NP-hard problems [9]. While clustering can be used to achieve energy optimization goals, a robust and reliable technique is needed that can efficiently cluster the sensor nodes with minimum overhead and redundancy. The K-means algorithm is a simple and efficient clustering method that can be implemented in WSNS, provided the limitations discussed above are appropriately addressed. Although several methods have been discussed above to address the issues in K-means, a much more reliable

and robust one is needed in WSNs due to their inherent constraint of massive and random deployment. The process of clustering in WSNs can be first divided into two phases, cluster formation and cluster head selection, and then individual techniques can be applied for better performance. The K-means algorithm can be combined with other Fuzzy-based or meta-heuristic algorithms like Genetic Algorithm, Particle Swarm Optimization, Ant Lion Optimization, etc., resulting in a much more efficient and reliable hybrid algorithm. In this section, a detailed literature survey of various algorithms proposed to improve and enhance the performance of K-means in WSNs is mentioned, and the basis on which the key characteristics of those algorithms are presented in a tabulated form.

In Novel K-means [10], the K-means algorithm in its original form is used for cluster formation and selection of cluster heads. After partitioning the network, the nearest sensor node to the centroid is chosen as the cluster head. A single-hop transmission mode is used by cluster heads for forwarding the data to the base station.

In EBK-means [11], a hybrid protocol based on K-means and Leach is proposed. Initially, clustering is done by K-means with the elbow method for optimum selection of K, and then in each of the K clusters, the LEACH algorithm is used to elect the CH, and for efficient transmission of data. The cluster member selects the CH based on the signal strength and cost of communication. It employs a CSMA scheduling protocol at the CH level and a TDMA protocol for intra-cluster communication.

In K-GTBEB [12], a K-means-based clustering and General Tree-based routing protocol has been proposed. The value of K optimum is based on a deterministic function that is a function of the total number of sensor nodes and field dimensions, alongside the other network parameters. However, as the K-means, it is based on random initialization of centroid points. The cluster heads are chosen as the sensor nodes nearest to the centroid points. Routing is based on a self-organized tree topology with load distribution among different branches of the tree managed by another method called the Analytical Hierarchy Process (AHP).

In MK-means [13], a modified K-means algorithm has been proposed. The algorithm is based on K-means with random initialization, and the Cluster head is chosen based on remaining energy. After selection as cluster head, it randomly selects two other nodes as cluster heads alongside, for load sharing during the operational phase. The members are selected on the basis of Euclidean distance, and the data forwarding is governed by a threshold in order to reduce redundant transmission of data. A post-clustering phase is defined in the event that a major portion of sensor nodes does not qualify for becoming cluster heads, and is based on direct transmission of data to the base station.

In EECPPK-means [14], a deterministic approach is used for determining the k-optimum. Initially, the clusters are formed via the K-means algorithm, and then the cluster heads are chosen as the nearest ones to the midpoint of each cluster.

An energy threshold is defined for the sensor nodes to continue to act as cluster heads. The mode of data forwarding to the base station is decided by another distance threshold. If the distance to the base station is less than the threshold, then direct communication is used; otherwise, multi-hop communication will be used.

Expression for the selection of K_{opt} is

$$K_{opt} = \frac{\sqrt{N}}{\sqrt{2\pi}} \times \sqrt{\frac{\epsilon_{fs}}{\epsilon_{mp}}} \times \frac{M}{d_{BS}^2} \quad (12)$$

Where N is the total number of sensor nodes, M is the dimension of the square field.

In KPSO [15], a novel clustering algorithm based on K-means and Particle Swarm Optimization (PSO) is proposed. The K-means algorithm has been used to divide the network into a random number of clusters based on Euclidean distance with random initialization. After clustering, the PSO is employed to select the best possible cluster heads from each cluster using a fitness function that aims to maximize the lifetime of the cluster head.

$$\text{Fitness Function} = \begin{cases} \sum_{i=1}^{all} CHS \frac{E_i}{d_{BS}^2} & \text{for } d_{BS} < d_o \\ \sum_{i=1}^{all} CHS \frac{E_i}{d_{BS}^4} & \text{for } d_{BS} \geq d_o \end{cases} \quad (13)$$

Where E_i is the initial energy of CH, d_{BS} is the Euclidean distance to BS and, d_o is the threshold distance (87m).

In KH-means [16], a modified K-means algorithm based on the harmonic mean instead of the Euclidean distance is proposed. The number of clusters is selected randomly, and even though the centroids are initialized randomly, the algorithm results in a uniform distribution of cluster heads due to the assignment of different weights to the member nodes based on proximity to multiple centroids. The Cluster heads are elected based on distance from centroid points and the residual energy. The algorithm also proposes a mechanism for handling the sensor nodes lying at the intersection of one or more clusters.

In Hybrid K-means [17] and Improved LEACH [18], a modified form of K-means is proposed with cluster formation based on K-means and selection of the cluster heads based on the LEACH protocol. While [17] has implemented the LEACH protocol in its original form, in [18] the probability of a sensor node becoming a cluster head is decided on the basis of the residual energy.

In Modified EECPPK-means [19], modification is done in the earlier proposed EECPPK [14] protocol. The mid-point algorithm is used in the selection of the K optimum, initialization, and election of the cluster head. The mid-point algorithm is used to locate the centroid points along the length and breadth of the network, and the nearest sensor nodes are chosen as initial cluster heads. Such a method of locating centroids is adopted for the uniform distribution of the cluster heads across the network. After each iteration, the residual energy of the cluster heads is compared with a threshold in order to continue as cluster head; the next sensor node nearest to the centroid will be elected as the cluster head. The expression for finding the mid-point is

$$Mid - point = \frac{(R-0)}{2} \quad (14)$$

Where R is the sensing range of the nodes.

In K-means with SAW & WPM [20], the K-means algorithm is used for clustering the nodes based on data point attributes, depending upon node operation and distance from the centroid point. The election of the cluster head is based on various decision criteria like, Simple Additive Weight (SAW) and Weighted Product Method (WPM). The main aim of the proposed method is to reduce the time complexity of data aggregation.

In UCRPK-means [21], an unequal clustering algorithm is proposed based on K-means. The k-optimum is chosen by equation (15), and then the size of each cluster is adjusted based on distance from the base station by means of the expressions (16 and 17)

Expression for the selection of K_{opt} is

$$K_{opt} = \frac{\sqrt{N}}{\sqrt{2\pi}} \times \frac{\sqrt{\epsilon_{fs}} \times \sqrt{S}}{\epsilon_{mp} d_{BS}^2} \quad (15)$$

$$r = \frac{N - K_{opt} \times count}{1 + 2 + \dots + (K_{opt} - 1)} \quad (16)$$

$$Size(i)_{for \ i=1:K_{opt}} = count + r(i - 1) \quad (17)$$

Where N is the total number of deployed sensor nodes and, the count is the number of sensor nodes within a particular distance d_o to the base station.

After the initial centroid (μ_k) selection, the algorithm tries to minimize the intra-cluster variation like K-means by the following expression

$$J = \sum_{n=1}^N \sum_{k=1}^{K_{opt}} r_{nk} \| (x_n - \mu_k) \|^2 \quad (18)$$

The cluster formation is rest similar to that of K-means. Routing is based on the threshold either, as a single-hop or multi-hop inter-cluster communication.

In Modified K-means with Ant Loin Optimization [22], a hybrid clustering algorithm is proposed based on K-means and Ant Loin Optimization (ALO). The K-means is used for cluster formation with a Minkowski distance being used instead of Euclidean. The cluster head is elected by applying the ALO algorithm.

In Enhanced KEAC [23], a centralized approach is used for clustering. The K-means algorithm is used for cluster formation, while the selection of the cluster head is based on a weighted fitness function (equations 19 and 20), considering the residual energy, distance of a particular node to the centroid, and location w.r.t. base station. The nodes having a fitness value greater than a standard value get elected as cluster heads.

$$W_{ni} = c_1 * E_{ni} + c_2 * D_{cci} \quad (19)$$

$$W_{std} = c_1 * E_{Total} + c_2 * AvgD_{cci} \quad (20)$$

Where c_1 & c_2 are the arbitrary constants.

W_{ni} is the fitness weight of the node, E_{ni} is the energy, and D_{cci} is the distance between the node and the centroid. W_{std} is the standard fitness weight, E_{Total} is the energy required for each round, and $AvgD_{cc}$ is the average intra-cluster to centroid.

In OK-means [24], a clustering algorithm based on K-means with a uniform distribution of cluster heads is

proposed. It is a modified form of K-means with a deterministic selection of K-optimum, but with random initialization of the centroids. The K-optimum is obtained by the following expression.

$$K_{opt} = \sqrt{\left(\frac{2N \times \epsilon_{fs} \times R^2}{\pi \epsilon_{mp} N - 4E_{elec}} \right)} \quad (21)$$

After determining the K_{opt} , the centroid points are randomly initiated, and the nodes are assigned to each centroid based on the Euclidean distances. After convergence of the algorithm, the node closest to the centroid point with sufficient energy is chosen as the cluster head.

In Improved K-means [25], a modified K-means algorithm based on multiple fitness functions is proposed for cluster formation as well as the selection of cluster heads. The selection of K-optimum and initialization of centroids are based on fitness functions considering parameters like communication cost for each round, packet size, and a threshold for the maximum number of adjacent nodes. The algorithm has addressed the uneven distribution of cluster heads by dividing the entire network area (PxP dimensions) equally among K_{opt} cluster heads. The expression for calculating the K_{opt} is

$$K_{opt} = \frac{\sqrt{N}}{\sqrt{2\pi}} \times \sqrt{\left(\frac{\epsilon_{fs} \times P^2}{\epsilon_{mp} d_{BS}^2 - E_{elec}} \right)} \quad (22)$$

Avg number of nodes in a cluster

$$N_{opt} = \frac{N}{K_{opt}} \quad (23)$$

Where N is the total number of sensor nodes.

The cluster formation is based on the competition radius and the number of adjacent nodes. After determining the centroid, the member nodes are added based on the Euler Formula.

The selection of cluster heads in each cluster is based on a deterministic function that considers the residual energy of the node, node centrality, and node degree. All these parameters are compared with the average values in that particular cluster. The expression for selecting the cluster head is

$$T_{hi} = \lambda_1 \frac{E_i}{E_h} + \lambda_2 \frac{e_i \cdot V_{ic}}{e_h \cdot V_{ic}} + \lambda_3 \frac{d_{avg}^h}{d_{(L, NCE_h)}} \quad (24)$$

Where T_{hi} is the probability of a node to become cluster head, $e_i \cdot V_{ic}$ is the average number of adjacent nodes, and d_{avg}^h is the distance of the node from the centroid of that cluster. In the denominator of the expression, the average values of the numerator parameters over the cluster.

In Quad clustering-based K-means [26], a quad-clustering approach is proposed. The K-means algorithm is initially used to locate the centroid of the entire network for positioning the base station, and then the network field is divided into four equal quadrants. Again, the K-means is used to find the centroid points of each cluster, and then the sensor node nearest to the centroid is elected as the cluster head.

In GA-based K-means [27], a hybrid clustering and routing algorithm based on K-means and Genetic Algorithm is proposed. The cluster formation is based on random initialization of K clusters obtained by equation (25). The

selection of the cluster head is based on a multi-objective fitness function considering residual energy, average intra-cluster distance of the node, distance to the base station, and the number of times the node has been previously selected as a cluster head. The expression for the fitness function is

$$K_{opt} = \sqrt{\frac{N}{2}} \quad (25)$$

$$Cost_i = \alpha * E_i + \frac{\beta}{Dist(i, BS)} + \frac{\gamma}{ec_i} \quad (26)$$

In KCED [28], a modified K-means algorithm is proposed to prolong the network lifetime. The process of selection of K-optimum and initialization is the same as conventional K-means. The algorithm has proposed a novel method of cluster head selection based on twin score functions that consider the residual energy, distance to the centroid (d_{CC}), and the distance to the base station (d_{BS}). For the sensor nodes having residual energy greater than the average residual energy of the cluster, the score values are calculated by the following expressions.

$$Score_{BS} = \frac{E_{residual}}{E_{initial}} - \frac{d_{BS}}{Max(d_{BS})} \quad (27)$$

$$Score_{CC} = \frac{E_{residual}}{E_{initial}} - \frac{d_{CC}}{Max(d_{CC})} \quad (28)$$

The sensor nodes having the maximum value of $Score_{BS}$ and $Score_{CC}$ are elected as the cluster heads within each cluster. When maximum score values are obtained for different nodes, the node with the maximum score is selected as the cluster head. However, if multiple nodes have the same value of score functions, then the total communication distance is used for the selection of the appropriate sensor node.

$$Total Distance(i) = d_{i to BS} + \sum_{j=1}^{MN_i} d_{i,j} \quad (29)$$

The member node having the minimum value of $Total Distance(i)$ is selected as the final cluster head.

In KOCED [29], the author has addressed the random selection of the K-optimum in KCED [28]. Despite other steps and parameters being similar, the simulations have indicated an improvement in the overall performance of the algorithm in terms of clustering time as well as the lifetime of the network.

In DCK-LEACH [30], the K optimum is calculated based on a canopy pre-clustering technique. It involves the random selection of a node, calculating the Euclidean distance of all other nodes w.r.t. the selected node and comparing it with a threshold. The same process is repeated for all the nodes of the network. After selection, the K centroid points are randomly initiated, and members are added based on Euclidean distance to the centroid points. The algorithm also proposes a dual selection of cluster heads, the primary cluster

head and the vice cluster head. The primary cluster head is responsible for intra-cluster communication, while the vice cluster head forwards the data from the primary cluster head to the base station. This dual approach is employed to balance the load among the nodes. The primary and the vice cluster heads are selected by multiple fitness functions, listed below.

$$f_1(i) = \frac{E_{res}(i)}{\frac{1}{n} \sum_{j=1}^n E_{res}(j)} \quad (30)$$

$$f_2(i) = \frac{\frac{1}{n} \sum_{j=1}^n d_{to center}(j)}{d_{to center}(i)} \quad (31)$$

$$P(i) = \alpha f_1(i) + \beta f_2(i) \quad (32)$$

$$f_3(i) = \frac{\frac{1}{n} \sum_{j=1}^n d_{to BS}(j)}{d_{to BS}(i)} \quad (33)$$

$$Q(i) = \mu f_1(i) + \phi f_3(i) \quad (34)$$

The equations (30-32) are used for the selection of the primary cluster head, and the equations (30, 33, 34) are used for the selection of the vice cluster head.

In KM-MWOA [31], a modified K-means algorithm has been proposed. The algorithm is based on K-means with random initialization, and the Cluster head is chosen based on the distance to the centroid and the remaining energy. The members are selected on the basis of Euclidean distance from the cluster heads. For energy-efficient routing, a modified form of the Whale optimization algorithm has been proposed.

In Fuzzy-based K-means [9], a hybrid clustering algorithm based on K-means and fuzzy logic is proposed. The cluster formation is based on K-means, while multiple fuzzy fitness functions are used to select the cluster head.

CONCLUSION

The clustering in WSNs by K-means, though an energy-efficient technique, involves several challenging tasks. The selection of K-optimum and the proper initialization of K-centroids can be detrimental to the algorithm's performance. Various methods have been identified to address these issues involving deterministic and meta-heuristics algorithms like PSO, ACO, and GA. Although they have improved the efficiency of the K-means clustering, each has its drawbacks. The deterministic approaches are difficult to implement due to the dynamic and unreliable nature of the WSNs. Whereas, the meta-heuristic approaches have the issue of slow convergence leading to time complexity and can get waded into a local optimal solution. Although due to the NP-hard nature of these problems, a global optimal solution can really be a challenging task, a greedy approach should be adopted for establishing a trade-off between the complexity and the performance of the proposed algorithm.

Table 2. Summary of prominent Algorithms

Algorithm	Year	Cluster Formation			Cluster Head Selection		Routing
		Initialization	Selection of K	Selection of Cluster Members	Method	Parameters	
Novel K-means	2013	Random	Random	K-means	Nearest to	Euclidean Distance	Single-Hop from CH

Algorithm	Year	Cluster Formation			Cluster Head Selection		Routing
		Initialization	Selection of K	Selection of Cluster Members	Method	Parameters	
[10]					centroid	and Energy.	to BS
EBK-means[11]	2014	Random	Elbow method at BS	K-means	LEACH	RSSI and transmission cost.	LEACH based with CSMA at CH level.
K-GTBEB[12]	2015	Random	Deterministic (Mathematical Formula)	K-means	K-means	Euclidean Distance	Centralized. GTBEB routing with load balancing scheme.
MK-means[13]	2016	Random	Random	K-means based on Euclidean distance from CH.	Nearest to centroid with three CHs within a single cluster.	Euclidean Distance and Energy (70% of initial energy)	A threshold it set and data is transmitted with TDMA scheduling
EECPK-means [14]	2016	Mid-Point of each cluster	Deterministic (Mathematical Formula)	K-means based on Euclidean distance	Mid-point Algorithm	Euclidean Distance from centroid and Energy.	Based on a distance criterion; either single hop or Multi-hop.
KPSO[15]	2016	Random	Arbitrary/ Random	K-means based on Euclidean distance	Particle Swarm Optimization (PSO)	Energy and Euclidean Distance to BS	-
KH-means[16]	2017	Random	Random	Harmonic distance from CH with weight assignment	Nearest to centroid	Distance and residual energy	-
Hybrid K-means[17]	2017	Random	Random	K-means based on Euclidean distance from CH.	LEACH 5% of total population)	Euclidean Distance	-
Improved LEACH[18]	2017	Random	Random	K-means based on Euclidean distance from CH.	LEACH	Residual energy	-
Modified EECPK-means [19]	2018	Mid-point Algorithm	Mid-point Algorithm	Based on distance to the nearest CH	Mid-point Algorithm	Distance to centroid and residual energy	Based on a distance criterion; either single hop or Multi-hop.
K-means with SAW &WPM [20]	2018	Random	-	Based on Euclidean distance to the nearest CH	Multiple criteria Decision Making	SAW &WPM	-
UCRPK-means [21]	2019	Deterministic; intermediate node of each uneven set.	Deterministic (Mathematical Formula)	Based on Euclidean distance to the nearest CH	Nearest to centroid	Euclidean Distance from centroid and Energy	Based on a distance criterion; either single hop or Multi-hop.
Modified K-means with Ant loin Optimization [22]	2019	-	-	Based on Minkowski distance from centroid	Ant Loin Optimization	Fitness Function	spherical grid multi-tier routing
Enhanced KEAC [23]	2019	Random	Deterministic (Mathematical Formula)	Based on Euclidean distance to the nearest CH	Weighted Fitness Functions	Distance to central point, location and Residual energy	-
OK-means [24]	2020	Random	Deterministic (Mathematical Formula)	Based on Euclidean distance to the nearest centroid	Nearest to centroid	Euclidean Distance from centroid and Energy	Single hop intra-cluster and Multihop inter-cluster communication.

Algorithm	Year	Cluster Formation			Cluster Head Selection		Routing
		Initialization	Selection of K	Selection of Cluster Members	Method	Parameters	
Improved K-means [25]	2020	Deterministic; based on Sensing range and adjacent nodes	Deterministic; based on channel model	Based on the Euler formula	Deterministic	Residual energy, node degree and centrality.	Multihop communication
Quad clustering based K-means [26]	2020	Random	Fixed (Quad-clusters)	Based on Euclidean distance to the nearest centroid	K-means; Nearest to centroid	Euclidean Distance from centroid	-
GA-based K-means [27]	2020	Random	Deterministic; Rule of Thumb	K-means based on Euclidean distance from centroid	Deterministic	Residual Energy, Distance to BS and Node CH selection Frequency.	Genetic algorithm for optimal path establishment.
KCED [28]	2021	Random	Random	K-means based on Euclidean distance from CH.	Multi-objective Score based	Residual energy, distance to centroid and Distance to BS	LEACH
KOCED [29]	2021	Random	Deterministic (Mathematical Formula)	K-means based on Euclidean distance from centroid	Multi-objective Score based	Residual energy, distance to centroid and Distance to BS	LEACH
DCK-LEACH [30]	2022	Random	Canopy based Pre-clustering	K-means based on Euclidean distance from centroid	Dual cluster head;	Residual energy, distance to centroid and distance to BS.	LEACH
KM-MWOA [31]	2023	Random	Random	K-means based on Euclidean distance from centroid	K-means; Nearest to centroid	Euclidean Distance from centroid and Energy	Whale optimization Algorithm
Fuzzy Based K-means [32]	2023	Random	Random	K-means based on Euclidean distance from Centroid	Multi-Attribute Fuzzy-based Decision Making.	Residual energy, Nearest neighbour and Distance to BS.	-

REFERENCES

- [1] Abdullah, M., Al-Thobaity, A., Bawazir, A., & Al-Harbe, N. (2013). Energy Efficient Ensemble K-means and SVM for Wireless Sensor Network. *INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY*, 11(9), 3034–3042. <https://doi.org/10.24297/ijct.v11i9.3409>.
- [2] Devi, G., Swain, S., & Sankar Bal, R. (n.d.). *The K-Means Clustering used in Wireless Sensor Network*. <https://www.researchgate.net/publication/375331726>.
- [3] Ali Abdul-hussian Hassan, Wahidah Md Shah, Ali Mohamed Husien, Mohammed Saad Talib, Ali Abdul-Jabbar Mohammed, & Mohd Fairuz Iskandar. (2019). *International Journal of Recent Technology and Engineering (IJRTE) Volume-7, Issue-6S5, April 2019*. <https://www.ijrte.org/portfolio-item/f10200476s519/>.
- [4] Han, J., Kamber, M., & Pei, J. (2012). Cluster Analysis. In *Data Mining* (pp. 443–495). Elsevier. <https://doi.org/10.1016/B978-0-12-381479-1.00010-1>.
- [5] Zhong, S., Wang, G., Leng, X., Wang, X., Xue, L., & Gu, Y. (2012). A Low Energy Consumption Clustering Routing Protocol Based on K-Means. *Journal of Software Engineering and Applications*, 05(12), 1013–1015. <https://doi.org/10.4236/jsea.2012.512117>.
- [6] Zhao, Y., & Zhou, X. (2021). K-means Clustering Algorithm and Its Improvement Research. *Journal of Physics: Conference Series*, 1873(1). <https://doi.org/10.1088/1742-6596/1873/1/012074>.
- [7] Annachhatre, C., Austin, T. H., & Stamp, M. (2015). Hidden Markov models for malware classification. *Journal of Computer Virology and Hacking Techniques*, 11(2), 59–73. <https://www.scirp.org/reference/referencespapers?referenceid=1737599>.
- [8] Belyadi, H., & Haghighat, A. (2021). Unsupervised machine learning: clustering algorithms. *Machine Learning Guide for Oil and Gas Using Python*, 125–168. <https://doi.org/10.1016/B978-0-12-821929-4.00002>.
- [9] Hatemoglu, A. (2024). *NP-Hard Problems in Wireless Sensor Networks (WSNs) and the Internet of Things (IoT)*. https://www.researchgate.net/publication/379534168_Metaheuristic_Algorithms_for_Solving_NP-hard_problems_of_WSNs_and_IoTs.
- [10] Park, G. Y., Kim, H., Jeong, H. W., & Youn, H. Y. (2013). A novel cluster head selection method based on k-means algorithm for energy efficient wireless sensor network. *Proceedings - 27th International Conference on Advanced*

- Information Networking and Applications Workshops, WAINA 2013*, 910–915. <https://doi.org/10.1109/WAINA.2013.123>.
- [11] Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. In *International Journal of Computer Applications* (Vol. 105, Issue 9). <https://www.ijcaonline.org/archives/volume105/number9/18405-9674/>.
- [12] Kaur, P., & Bhardwaj, V. (n.d.). *K-Means based General Self-Organized Tree-Based Energy Balancing Routing Protocol for Wireless Sensor Networks*. <https://www.ijcsit.com/docs/Volume%206/vol6issue04/ijcsit2015060445.pdf>.
- [13] Periyasamy, S., Khara, S., & Thangavelu, S. (2016). Balanced Cluster Head Selection Based on Modified k -Means in a Distributed Wireless Sensor Network. *International Journal of Distributed Sensor Networks*, 2016. <https://doi.org/10.1155/2016/5040475>.
- [14] Ray, A., & De, D. (2016). Energy efficient clustering protocol based on K-means (EECPK-means)-midpoint algorithm for enhanced network lifetime in wireless sensor network. *IET Wireless Sensor Systems*, 6(6), 181–191. <https://doi.org/10.1049/iet-wss.2015.008714>.
- [15] Solaiman, B. F., & Sheta, A. (2016). Energy optimization in wireless sensor networks using a hybrid K-means PSO clustering algorithm. *Turkish Journal of Electrical Engineering and Computer Sciences*, 24(4), 2679–2695. <https://doi.org/10.3906/elk-1403-293>.
- [16] *IEEE Student Conference on Research and Development (Scored) 2017: Inspiring technology for humanity: proceedings: 13-14 December, Putrajaya, Malaysia*. (2017). IEEE. <https://www.proceedings.com/38559.html>.
- [17] Mahboub, A., Arioua, M., & En-Naimi, E. M. (2017). Energy-efficient hybrid K-means algorithm for clustered wireless sensor networks. *International Journal of Electrical and Computer Engineering*, 7(4), 2054–2060. <https://doi.org/10.11591/ijece.v7i4.pp2054-2060>.
- [18] L. Kaur and S. Kad, "Modified EECPK-means mid-point algorithm for enhancing network life expectancy in WSN," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2018, pp. 1262-1265. <https://doi.org/10.1109/ICECA.2018.8474875>.
- [19] Saheb, P., & Kanika Sharma, D. (n.d.). Improved LEACH Protocol Based on K-Means Clustering Algorithm for Wireless Sensor Network. *International Journal of Electronics & Communication Technology*. <https://www.iject.org/vol8/issue4/5-pratiksha-saheb.pdf>.
- [20] *Proceedings of the Second International Conference on Electronics, Communication and Aerospace Technology (ICECA 2018): 29-31, May 2018*. (2018). IEEE. www.proceedings.com/content/041/041134webtoc.pdf.
- [21] Khandelwal, A., & Jain, Y. K. (2018). An efficient k-means algorithm for the cluster head selection based on SAW and WPM. *International Journal of Advanced Computer Research*, 8(37), 191–202. <https://doi.org/10.19101/IJACR.2018.836022>.
- [22] Dong, X., Zhang, Y., & Yu, S. (2019). An Uneven Clustering Routing Protocol based on Improved K-means Algorithm for Wireless Sensor Network in Coal-mine. In *Journal of Information Hiding and Multimedia Signal Processing c* (Vol. 10, Issue 1). <https://bit.kuas.edu.tw/2019/vol1/JIH-MSP-2019-01-07.pdf>.
- [23] Dhand, G., Tyagi, S.S. SMEER: Secure Multi-tier Energy Efficient Routing Protocol for Hierarchical Wireless Sensor Networks. *Wireless Pers Commun* 105, 17–35 (2019). <https://doi.org/10.1007/s11277-018-6101-y>.
- [24] Devi, R. R., & Sethukarasi, T. (2019). Energy Efficient Enhanced K-Means Cluster-based Routing Protocol for WSN. *International Research Journal of Engineering and Technology*. <https://www.irjet.net/archives/V6/i11/IRJET-V6I11294.pdf>.
- [25] el Khediri, S., Fakhret, W., Moulahi, T., Khan, R., Thaljaoui, A., & Kachouri, A. (2020). Improved node localization using K-means clustering for Wireless Sensor Networks. In *Computer Science Review* (Vol. 37). Elsevier Ireland Ltd. <https://doi.org/10.1016/j.cosrev.2020.100284>.
- [26] Tang, X., Zhang, M., Yu, P., Liu, W., Cao, N., & Xu, Y. (2020). A nonuniform clustering routing algorithm based on an improved k-means algorithm. *Computers, Materials and Continua*, 64(3), 1725-1739A. <https://doi.org/10.32604/cmc.2020.010272>.
- [27] Kumar, B., Tiwari, U. K., & Kumar, S. (2020). Energy efficient quad clustering based on k-means algorithm for wireless sensor network. *PDGC 2020 - 2020 6th International Conference on Parallel, Distributed and Grid Computing*, 73–77. <https://doi.org/10.1109/PDGC50313.2020.9315853>.
- [28] Benmahdi, M. B., & Lehsaini, M. (2020). A GA-based Multihop Routing Scheme using K-Means Clustering approach for Wireless Sensor Networks. *2020 2nd International Conference on Embedded and Distributed Systems, EDiS 2020*, 155–160. <https://doi.org/10.1109/EDiS49545.2020.9296444>.
- [29] Lee, J. Y., & Lee, D. (2021). K-means clustering-based WSN protocol for energy efficiency improvement. *International Journal of Electrical and Computer Engineering*, 11(3), 2371–2377. <https://doi.org/10.11591/ijece.v11i3.pp2371-2377>.
- [30] Park, S. Y., Lee, J. Y., & Lee, D. (2021). Enhanced KOCED Routing Protocol with K-means Algorithm. *Computers, Materials and C* Wu, M., Li, Z., Chen, J., Min, Q., & Lu, T. (2022). A Dual Cluster-Head Energy-Efficient Routing Algorithm Based on Canopy Optimization and K-Means for WSN. *Sensors*, 22(24). <https://doi.org/10.3390/s22249731>.
- [31] Kaviarasan, S., & Srinivasan, R. (2023). Energy efficient based optimized k-means and modified whale optimization algorithm for cluster head selection in wsn. *Journal of Theoretical and Applied Information Technology*, 15(1). <https://www.jatit.org/volumes/Vol101No1/6Vol101No1.pdf>
- [32] Angadi, B. M., & Kakkasageri, M. S. (2023). K-Means and Fuzzy based Hybrid Clustering Algorithm for WSN. *International Journal of Electronics and Telecommunications*, 69(4), 793–801. <https://doi.org/10.24425/ijet.2023.147703>.