Clustering Approach in Wireless Sensor Networks Based on K-means: limitations and Recommendations


Abstract: Clustering approach in wireless sensor network is very important, the structure of cluster and how to improve it is a first challenge that faced the developers, because of it represent as a base for design the cluster-based routing protocol. One of most popular cluster algorithms that utilizing into organize sensor nodes is K-means algorithm. This algorithm has beneficial in construct the clusters for various real-world applications of WSN. K-means algorithm suffering from many drawbacks that hampering his work. The lack of adequate studies that investigates in the limitations of this algorithm and seek to propose the solutions motivated us to do this study. In this paper the limitations of K-means and some suggestions are proposed. These suggestions can improve the performance of K-means, which will be reflected on saving the energy for sensor nodes and consequently maximize the lifetime of the wireless sensor networks.

I. INTRODUCTION

Wireless Sensor Networks WSN sis consisting on a big number of sensors, which has a finite battery power. Due to it is working in the risk and harsh environments, so that impossible or very hard for battery replacing or recharging[1][3]. Consequently, conserve the energy is very significant for this network. Routing protocols has an immense influences on the energy consumption[4][5], where the energy consumption is considering a major factor in routing protocol design. A cluster base routing protocol is considering the best type of routing protocols in concept of energy saving for sensors and prolong network lifetime. In this protocol, sensor nodes is organize in specific groups called clusters. Each cluster has member nodes called ordinary nodes (ON) and special node called cluster head (CH). The CH has higher energy and utilized to collect and transmit data from ONs to base station (BS)[6][7][8]. In this approach of the routing protocols, the messages that convey through network can be reduce.

Consequently, the network lifetime will be maximize. One of challenges that faced clustering method is how to form the nodes in specific clusters i.e. choosing clustering method faces with problems and considerations that must be considered in order to reduce the energy consumption and maximum lifetime in the network [9][10]. Furthermore, do not decided the optimal clusters number in the cluster formation is producing some issues. K-means is considering one of algorithms that has been wide used for organizing the nodes in the clusters at wireless sensor networks. In this paper we investigate in K-means algorithm in terms the limitations of this algorithm and what the suggestion solutions that will be used for overcome these drawbacks. The rest of this study is arranged as follows:

In section 2 clustering process based on K-means algorithm. In section 3 clustering protocols based on k-means method will be studied. Limitations of K-means will illustrated in section 4. And in section 5, various Approaches for calculate optimum number of cluster will proposed. Finally, conclusion illustrated in section 6.

II. CLUSTERING PROCESS BASED ON K-MEANS ALGORITHM

The clustering technique is considering a widespread method which utilize to minimize the energy consumption in WSNs. Clustering procedure are organizing the sensor nodes into specific setknown as a clusters. Single node in every cluster is selected as a major of cluster known as a cluster head (CH). The CH has many function in addition to sensing the environment such as; data gathering from all sensors and cluster. The CH has higher energy and utilized to collect and transmit data from ONs to base station (BS)[6][7][8]. In this approach of the routing protocols, the messages that convey through network can be reduce.

Revised Manuscript Received on December 22, 2018.

Ali Abdul-hussian Hassan Faculty of Information and Communications Technology, UniversitiTeknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

WahidahMd Shah College of Education for Pure Sciences, University of Kerbala, Iraq

Ali Mohamed Husien College of Administration and Economics, University of Babylon, Iraq Email: ali.ab@uokerbala.edu.iq Orcid :0000-0002-0933-9807
Then comes the most common challenge of this type of protocol, which is how to select the best node to be cluster head. One of most popular algorithms that used to gather the nodes in WSNs is K-means (KM) algorithm. This algorithm is very beneficial in construct the clusters for various real-world applications of WSN [16]. This algorithm is one of unsupervised clustering methods, which efficiently utilized to form spherical shapes clusters [17]. Stuart Lloyd in 1982, was firstly researcher suggested this algorithm[18]. It divided points of data into specific number of clusters[19]. It mostly increases the distances between the clusters along with decrease distance inside the cluster. The goal of this algorithm is seek to find best cluster center when diminishing the objective function based on a Squared-Error-Function (SEF). We can defined the objective function of KM as:

$$E(\Gamma,\Lambda) = \sum_{i=1}^{k} \sum_{j=1}^{n} ||x_j - \bar{v}_i||^2$$

Where $||x_j - \bar{v}_i||^2$ represent The Euclidean distance that used to determine the distance between npoints of data $x_j$ withits cluster center $\bar{v}_i$.

The processes of this algorithm is including the following phases[20]:
Phase 1: Locate the k centroids points in the space which is representing by the data set, where K is predefined number.
Phase 2: Allocate every point of datato the specific cluster, which has nearest centroid distance.
Phase 3: Once all point of data have been clustered, re-determine the locations of the k centroids.
Phase 4: Reiterate the Phase 2 and Phase 3 till no shown change in the location of centroids.

### III. CLUSTERING PROTOCOLS BASED ON K-MEANS METHOD

In this section, we introduced some K-means algorithms that used for the nodes gathering in cluster-based routing protocols:

#### A. KPSO/KGA[21]

In this work, the authors extended their previous work that existing in [22]. Wherein, they implement a two-phase hybrid K-Means (KM) with Particular Swarm Optimization (PSO) was called (KPSO), and Genetic Algorithm (GA) clustering algorithms was called (KGA), for clustering sensor nodes of WSNs into clusters and selection of CH for each cluster. They improved their work through increase an extrastep for selection the optimum cluster members, which called it KPSO-PSO and KGA-GA.

This study was developed to presents an improved of the clustering algorithm into [22]. The algorithm of PSO was applied for WSN in order to organize the nodes in the specific clusters. It pointed to determine the optimum clusters’ number, and optimal CHs along with the best strategy of the clusters (i.e. the number of members and location of centroid in each cluster). Herein, the algorithm of PSO has molecules which refer to the clusters number, then the CH index followed for the each cluster. As showed in the Figure 2.

![Figure 2 PSO particle.](image)

So, (k) is mean the number of clusters, as well as CH, is the indicator of the CH for each cluster “i”. For that reasons, this study has promising results were compared to other traditional studies.

#### B. EECKP means[16]

In this work, Energy-Efficient Clustering Protocol based on K-means midpoint algorithm (EECKP-means) for WSN has been presented. Wherein the midpoint algorithm is utilized for enhance the initial selection of cluster centroid process. Because of the initial centroid selection was achieved by random way in K-means algorithm, which resulted unbalanced in the clusters, the presented work resulted a balanced clusters and consequently load-balance for CHS has achieved, and maximize the network lifetime. Moreover, the CH selection procedure will be optimization through depend on residual energy as a one of parameter in selection method as well as Euclidean distance that already utilized in KM. It determines the optimal number of required clusters based mathematical format, which includes the sensing area size along with the sensor number. Assume that N is meaning the sensors number which regularly distributed in the square area MxM. The optimal clusters number (kopt) can be determined as [23]:

$$k_{opt} = \frac{\sqrt{N}}{\sqrt{2} \pi} \sqrt{\frac{\varepsilon_{fs}}{\epsilon_{mp}}} \frac{M}{d_{BS}}$$

Here $d_{BS}$ is indicator to the distance between’s and CHs, $\varepsilon_{fs}$ is indicator toffee space model as well as $\epsilon_{mp}$ is indicator to multipath model.

$$DB = \sum_{i=1}^{k} \max \left( \frac{S_C(Q_i) + S_C(Q_j)}{d_i(Q_i, Q_j)} \right)$$

$$S_C = \frac{\sum ||x_i - C_k||^2}{N_k}$$
The authors in this work proposed a Multi-Hop Routing Energy Efficient Scheme (MRRCE) which depended on parameters such as distance residual and energy to get cluster heads and optimum clusters for energy efficiency objective, and accordingly the network lifetime is prolonging. In this work, they utilized the Steiner Points (SPs) concept[25]. Where after the network is gridding, draw a sequence of vertical and horizontal lines from each location node coordinates. It causes generating a lot of crossings among all sensors. In instance, each 2-sensors which are not placed in the equal direction, they have four joint crossings points, while the two of them are their coordinate’s locations[26]. The SPs are utilized as an alternative way for depending on random selection of initial cluster centroid step in K-means. Due to, SPs are the Joints all the nodes in the cluster, CHs find in best location for initial CH while in the traditional K-means algorithm, random nodes were utilized as initial CH, that mean these location could be in any place, even this place do not has any node around. Concluding, K-means algorithm has big dependency on the initial selection of CH, by choosing the correct location as initial CH, the number of iteration will be decrease.

**D. K means-Davies Bouldin index[27]**

Authors $c_p = (1/|C_p|) \sum_{h = 1}^{n} x_h \in C_p x_h$ generate a balanced in energy consumption on the clusters. In addition to utilized Gaussian elimination for select the best node as a CH which caused the distribution in the energy consumption. To solve the problem of determine the optimum number of clusters; they utilized the Davies Bould in Index (DBI) for this objective, in order to maximize the network lifetime. The DBI address each class separately and strive for determine how similar it is to the class which is nearest to it. It assess the intracluster objects similarity and intercluster dissimilarity. As a result, it generates a well clustering. The DBI is formed as:

Where $S_c(Q)$ represent the average distance between a cluster nodes $Q_k$ and its center. Where $N_c$: the number of nodes in the cluster $k$, and $d_{ce}$: the distance between the clusters centers $Q_k$ and $Q_{l}$ where:

$$d_{ce} = \|C_k - C_l\|$$

Then the centroid of each cluster is selecting as a first CH based on K-means. Next CH will be selection through the Gaussian elimination algorithm based on residual energy by BS.

**E. EBRP[28]**

In this protocol, authors’ proposed an Energy Balanced Routing Protocol (EBRP) for wireless sensor networks. In this protocol, they distribute the sensor network into various clusters based on K-means++ algorithm:

Random selection for node in the set X as cluster center $c$. The distances is calculated between the new center of cluster and other nodes by BS. A new node $x_m$ is selected as a new center of cluster $C_p$, while the probability of selection.

$$d^2(x_m, c_p) / \sum_{j=1}^{n} d^2(x_j, c_p)$$

Where $p \in 1…k, d (x_m, c_j)$ is the Euclidean distance between $x_m$ and $c_j$, and the $p$ is the cluster number. In the second stage, reiterated this stage till all the centers of clusters $K$ are determined. In 3th stage, the distance between every node to cluster center is determined, and the node is calculated to the adjacent center. It is indicated by $x_i \in C_p, C_p$ represent the nodes set cluster $p$. While in the 4th stage, a new center for every cluster is represented by:

Where $p \in \{1, ..., k\}$. Lastly, reiterated stages3 and this step till getting the number iterations setting. The K clusters are found by the algorithm through BS. After that, BS is doing the broadcast message contain the coordinate of center and the cluster number for all clusters.

**F. GAK-means[29]**

Authors in this work presented a novel mixture of K-means and improved Genetic Algorithm (GA) in order to decrease energy consumption and prolong network lifetime. GA-K-means, seek to minimize the energy consumption through determining the optimal number of cluster heads (CHs) by utilized improved of Genetic Algorithm (GA), where the number of CHs also point to the clusters number in the network. Due to GA have some issues like it is very slowly in the converging for huge data, hung in local optimum solution that mean not global optimum. Consequently, GA is employed along with K-Means algorithm. So in this protocol, GA with a few number of iterations is employed to realize the optimal solution.

The main point is the set of points that found by utilizing GA is employed as initial points for the K-Means algorithm. After that every datum according to its symmetry is assigned to single cluster. In case, when depend on the K-means only in order to form the clusters, it cannot find the optimum solution, because of K-means has sensitivity for data center, and failure to identify for the noise data and failure to discover non spherical cluster and last lyitmay hung in local optimum solution.
### IV. LIMITATIONS OF K-MEANS

- The initial cancroids are selected by random way for the input data set. As very iteration, the K-means construct different clusters depending on the different random selected for initial cancroids. As a result the initial cancroids in random selection cause the local optima.
- No guarantee for K-means will converge into optimal or better solution. Due to it run in many iteration, and no any analysis that doing to find the better result, consequently sometime produce unbalanced cluster size.
- Cluster number is not determined accurately and automatically, accordingly it needed to be set cluster number according to user input.
- K-means has very high complexity of time, where the points of data must re-assigned to number of iterations times for every loop run.
- In some rare cases, K-means may produce an empty cluster, because of the random initial centroids selection.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Protocol name</th>
<th>Clusters number</th>
<th>Initial selection point</th>
<th>Cluster formation method/parameters</th>
<th>CH selection method/parameters</th>
<th>CH selection parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheta, Alaa F</td>
<td>2015</td>
<td>KPSO/KGA</td>
<td>PSO</td>
<td>×</td>
<td>K-means/Euclidean distance</td>
<td>PSO/GA</td>
<td>Residual energy and Euclidean distance</td>
</tr>
<tr>
<td>Ray, Anindita</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>De, Debasish</td>
<td>2016</td>
<td>EECPK-means</td>
<td>mathematical</td>
<td>midpoint algorithm</td>
<td>K-means/Euclidean distance</td>
<td>K-means</td>
<td></td>
</tr>
<tr>
<td>Rezaei, Elham</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baradaran, Amir Abbas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heydariyan, Atefeh</td>
<td>2016</td>
<td>MRRCE</td>
<td>×</td>
<td>Steiner Points</td>
<td>K-means/Euclidean distance</td>
<td>K-means</td>
<td>Euclidean distance</td>
</tr>
<tr>
<td>Elkamel, Rabiaa Cherif,</td>
<td>2017</td>
<td>K means-Davies</td>
<td>Davies</td>
<td>×</td>
<td>K-means/Euclidean distance</td>
<td>Gaussian elimination</td>
<td>residual energy</td>
</tr>
<tr>
<td>Adnane</td>
<td></td>
<td>Bouldin Index</td>
<td>Bouldin Index</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li, Lin, Li, Donghui</td>
<td>2018</td>
<td>EBRP</td>
<td>×</td>
<td>×</td>
<td>K-means++/Euclidean distance</td>
<td>FLS-GA</td>
<td></td>
</tr>
<tr>
<td>Prabhu, Thiyagarajan</td>
<td>2016</td>
<td>GA-Kmeans</td>
<td>GA</td>
<td>GA</td>
<td>K-means/Euclidean distance</td>
<td>Improved GA</td>
<td>Residual energy</td>
</tr>
</tbody>
</table>
As mention above, K-means algorithm suffering from many drawbacks that hampering his work. Among these problems, determine the clusters number which is considered the most influential problem on performance of K-means in WSNs. Because of the nature of the CH activities, it expend extra energy to perform these various functions[30]. So, balancing the energy consumption of CHs is very significant problematic to long term for functioning of WSNs. Consequently, in the next section we present with some methods that will determine the value of K.

V. VARIOUS APPROACHES FOR CALCULATE OPTIMUM NUMBER OF CLUSTER

There are many methods in literature that used to selecting cluster number[20]. Authors focused on the specific approaches, where they are characterized as a fast convergence, simplicity, and it can performs well in large scale of WSNs, they are:

1. By rule of thumb
2. Information Criterion Approach
3. Choosing k Using the Silhouette
4. Elbow method
5. Cross-validation
6. An Information Theoretic Approach

By rule of thumb

It is very simple method[31]. This method can by apply to any type of data.

\[ K \approx \sqrt{\frac{N}{2}} \]

Where \( N \) is the number of sensor nodes (data points).

Information Criterion Approach

The clusters number in the combination paradigm increase results in a raise in the dimensionality of the paradigm, producing a drab raise in its likelihood. If user was need to finding the max likelihood model with any value of clusters number, he would finally end up by getting result that each data point is the unique member in the cluster. Clearly, he wishes to avoid this problem, where he must select some criteria which do not based likelihood as a singular factor. The Information Criteria Parameter is utilized for choosing amongst models along with different parameters number. It tries to balance the increasing in the likelihood because of extra parameters through setting a threshold value for each parameter. This process of selection consist from 2-stages of clustering phase, wherein the cluster number can selected as automatic choosing by user, which depended on any of information criteria. The techniques of model selection are depend to calculate the clusters number using mixture models[32].

Traditionally, selection of model is executed in two-phases. Firstly, the candidate set of models produce through some principles of learning (execute by max. likelihood learning (ML)) for a models scope.

Secondary, the appropriate model will be selected which depend on criterion of model selection. The Minimum Description Length (MDL) criterion[33], the Consistent Akaike’s Information criterion (CAIC)[34], and Akaike’s Information criterion (AIC)[35], are considering as famous examples for model selection criteria, that actually synchronize with the Bayesian inference criterion (BIC)[36], AIC and BIC are defined as

- \[ \text{BIC} = -2 \ln(\text{likelihood}) + k \ln(N) \]
- \[ \text{AIC} = -2 \ln(\text{likelihood}) + 2k \]

Where K is representing the degrees of model for freedom calculated as the rank of Variance Covariance Matrix of the parameter e(V) and N is the number of objects that utilized in the estimation, for more specifically, it represent the number of the likelihood independent terms. Practically, N is defined as e(N).

from the others. More precisely, the silhouette width is calculate as follow:

The \( a(i) \) represent the average distance between object \( (i) \) and other cluster objects, \( b(i) \) represent the min. average distances between object \( (i) \) and all the objects in the all other clusters. Range values of Silhouette Width \( s(i) \) is between -1 to 1. If all the silhouette width values are close to 1, it means that the set I is well clustered.

If \( S(i) \approx 1 \), that mean is a good clustered, and vice versa. But when the \( S(i) = 0 \), that mean this object can be re-assign to any other side. The clusters can be categorized by the Average Silhouette Width (ASW) for singular objects. The max of ASW for various clusters number is considering the correct number of clusters.

\[ s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \]
**Elbow Method**

The earlier way for estimation the right number of clusters in a data set (N)[40]. It is a visual method. The idea behind this method is that initiating the cluster number equal 2, then increase the cluster number one by one until reach to number less than 20% from N. In each step, the cluster cost must be calculating. At certain value of K, the cluster cost change will significantly, and then curve of cluster cost goes down gradually after that. This inflection point is representing the optimum K value. Disadvantage of elbow method is: This "elbow point" cannot at all times be clearly identified. Occasionally, there is no elbow point, or more than one elbows point as shown in Fig. 2.

![Elbow curve](image)

**Cross-validation**

Another method for calculating the clusters number is Cross Validation (CV), this method presented by[41]. It is depend on stability of cluster. This approach divided the objects for two parts or more. One them is utilized for clustering the data and for validation can utilized the other(s)part. The major concept is tending to frequently configure a similar clusters of data that generating from similar source. Where, this algorithm will be stable for randomization input. But when the data size be bigger and especially with the greater correlation, this method not has accurate to determine the suitable clusters number. Wang [42] was improve a cross validation by introduced some new parameters in order to calculating the clusters number. This novel election standard in order to determine the quality of clusters via utilized the instability among the samples. Where the instability for cluster process is evaluated via CV, in order to minimal the instability. All data is segmented for three sets; two of them is training-sets and the rest set is for validation in order to be similar to original concept of stability. After that, depended on the algorithm that used for clustering, a distance is measured for the independent training sets and use the validation set for evaluated the inconsistencies. It has been confirmed to be dynamic and powerful on a several examples.

**An Information Theoretic Approach**

An anther method for estimating the cluster numbers, that it sets limited to the parametric assumptions, may be strictly theory propose utilizing concepts from the rate distortion theory, it is easy to understanding and computing, and it has high effective for problems in the various range. It depend on “distortion” method that which calculates the dispersion in each cluster.

Sugar and James make improvement in the statistic by using jump-statistic, which they use the factor W in equation below to extend depending on Gaussian distribution method[43].

Precisely, the distance between an entity and centroid in equation below is calculated as:

The symbol Γk represent the covariance matrix of cluster, and the jump-statistic is calculated as:

\[ JS(K) = WK - M/2 - WK - 1 - M/2, \quad \text{where } W0 - M/2 = 0 \]

The max.of JS (K) is corresponding to the correct clusters number. This is reinforced by a derivation of mathematical indicating that if the data can be regarded a standard sample from the Gaussian distributions mixture, where the distances among canroids are almost sufficiently great. Thereafter, the max-jump would really happen at K equivalent the Gaussian components number in the mixture.

**IV.CONCLUSION**

In this work, we examine one of the most clustering algorithms that used in the WSNs, which is known as K-means, and its controversial issues. The problem in estimation the optimum clusters number has a great influences on balanced clusters in the network. This problem caused unbalanced energy consumption in network, consequently reduce the network lifetime. Occasionally, researchers are using mathematical formula and heuristic algorithms along with K-means to solve this problem and find optimal clusters number. In mathematical formula is not appropriate for all distributed methods, especially the random distributed. Also the heuristic algorithms is not suitable because it has a slow convergence, complex calculations and complex time processing. In order to overcome these problems, we proposed a specific methods can be utilize with K-means algorithm to determine the appropriate clusters number.

**ACKNOWLEDGMENT**

This work is fully supported by Universiti Teknikal Malaysia Melaka-Zamalah Scheme. So, the authors would like to thank Zamalah Scheme for providing the facilities and financial support for this research.
REFERENCES


34. H. Bozdogan, “Model selection and Akaike’s Information Criterion (AIC): The general theory and its...


