A comparative study of Gaussian mixture algorithm and K-means algorithm for efficient energy clustering in MWSN

Iman Ameer Ahmad\textsuperscript{1,2}, Muna Mohammed Jawad Al-Nayar\textsuperscript{1}, Ali M. Mahmood\textsuperscript{1}

\textsuperscript{1}Department of Control and Systems Engineering, University of Technology-Iraq, Baghdad, Iraq
\textsuperscript{2}Department of Basic Science, College of Dentistry, University of Baghdad, Baghdad, Iraq

ABSTRACT

Wireless sensor networks (WSNs) represent one of the key technologies in internet of things (IoTs) networks. Since WSNs have finite energy sources, there is ongoing research work to develop new strategies for minimizing power consumption or enhancing traditional techniques. In this paper, a novel Gaussian mixture models (GMMs) algorithm is proposed for mobile wireless sensor networks (MWSNs) for energy saving. Performance evaluation of the clustering process with the GMM algorithm shows a remarkable energy saving in the network of up to 92%. In addition, a comparison with another clustering strategy that uses the K-means algorithm has been made, and the developed method has outperformed K-means with superior performance, saving energy of up to 92% at 4,500 rounds.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Iman Ameer Ahmad
Department of Control and Systems Engineering, University of Technology
Baghdad, Iraq
Email: cse.21.11@grad.uotechnology.edu.iq

1. INTRODUCTION

Modern real-world applications where the sensor nodes are mobile heavily rely on mobile wireless sensor networks (MWSNs). MWSNs are a collection of sensor nodes distributed throughout a particular location. Sensors could process data, recognize data, and communicate wirelessly. Every sensor node is typically powered with worked-in battery-confined power. Since the sensor nodes in MWSNs can be installed in any circumstance and can adapt to quickly changing topology, they are far more flexible than static wireless sensor networks (WSNs). The task of mobile sensor nodes involves sensing countless physical phenomena such as light, temperature, humidity, pressure, and mobility. Hence, these networks can be utilized in several applications including environmental monitoring, mining, meteorology, seismic monitoring, acoustic detection, monitoring of processes in the healthcare industry, protection of infrastructure, context-aware computing, underwater navigation, smart spaces, inventory tracking, and tactical military surveillance are among the major uses of MWSNs [1]. Furthermore, there are several applications available for MWSNs that provide helpful solutions for certain real problems. A portion of the early applications of MWSN are estimated as: health sector, home management, policing, monitoring, natural designing applications, military applications, and directing conventions [2–6].

In fact, since sensor nodes may be placed in dangerous or unusable environments, charging or replacing the battery may be inconvenient or impossible [7]. Hence, energy consumption is one of the key distinctions between a WSNs and a conventional wireless network. In order to extend the network’s lifespan, it is important to establish methods for dependable data relay from sensor nodes to the sink and energy-efficient route creation at the network layer. Prolonging the lifespan of the network has a significant impact on how well

Journal homepage: http://beei.org
sensor network applications operate. One of the applied approaches is the energy dependent architecture (EDA). This technique is employed to reduce the overall power consumption of MWSNs according to the type of power distribution that occurs through the network.

Recent research studies have shown how organizing center nodes in MWSNs might result in energy saving. In one of three states which are (work, send, and receive), the sensor center point typically operates. These scenarios have a significant impact on the amount of power used. Continuous research has shown that sensor networks with center-point with express events can be used to arrange sensor power–hold [8]. One of these practices is to let sensor center points stir and rest. A few realistic computations are employed to regulate and control sensor power use needed [9], [10]. In addition, to save energy, the transmitting methods are also been identified as crucial concerns for WSN networks [11]. Due to the limited power available at sensor center points, data gathered based on objective conditions is directly relayed to the base station (BS) [12], [13]. A center named BS (also known as sink) negotiates for obtaining data along several sensor center locations. The BS examines the short proximity of the supplied information, which is used in the route. Additionally, the BS can transfer this information to multiple networks set up in different areas along with being ready to use it locally [14]–[17].

The clustering of sensor nodes is one of the most effective strategies to preserve the energy of the sensor node [18]. Throughout the clustering, the network is divided into numerous groupings known as clusters. A cluster head (CH) is a member of each group. CHs can gather local information from nodes of the cluster, aggregate it, and communicate the data to a distant BS directly or via other CHs. The abolition of duplicate data, improved network scalability, and preservation of the capacity of transmission are other advantages of clustering [19]. The selection of CHs is the crucial step in the clustering process, though, since it has significant implications for energy conservation in member sensing nodes and is crucial to the longevity of the network. Additionally, it has an efficiency in energy impact on the data routing procedure, which is the main goal [19]. Therefore, care should be used in selecting CHs. The selection of CH can be considered an NP-hard optimization problem [20]. The field of CH selection in WSNs has been investigated and researched in numerous research studies as shown in Table 1.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Clustering approach</th>
<th>Enhancements</th>
<th>Simulator</th>
</tr>
</thead>
<tbody>
<tr>
<td>[21]</td>
<td>Two phases clustering model</td>
<td>Improve the total network lifetime using the energy efficient clustering algorithm (EECA), for WSNs.</td>
<td>C programming language</td>
</tr>
<tr>
<td>[22]</td>
<td>An enhanced genetic algorithm and data fusion technique</td>
<td>An improved CH node is used in the proposed energy-efficient routing protocol to choose a method that can assess the remaining energy and directions of each participating node.</td>
<td>MATLAB</td>
</tr>
<tr>
<td>[23]</td>
<td>CH selection based on particle swarm optimization (PSO)</td>
<td>The lifetime of the network is increased when PSO is used for the best CH selection.</td>
<td>MATLAB</td>
</tr>
<tr>
<td>[24]</td>
<td>CH selection based on a firefly optimization</td>
<td>Increasing the network longevity by utilizing a firefly-based optimization strategy algorithm for CH selection.</td>
<td>MATLAB</td>
</tr>
<tr>
<td>[25]</td>
<td>K-means clustering-based routing protocol</td>
<td>The K-means clustering-based routing protocol that is suggested which takes into account the ideal fixed packet size to minimize the energy consumption of individual nodes and increase the network lifetime.</td>
<td>MATLAB</td>
</tr>
<tr>
<td>[26]</td>
<td>Dynamic clustering and distance aware routing protocol</td>
<td>The algorithm is focused on the role of super CH in saving the power of CH when nodes are too far from the BS.</td>
<td>MATLAB</td>
</tr>
</tbody>
</table>

Through reviewing the above related works and investigating the modern articles concerning efficient clustering, and their many types of WSN clustering techniques. In this paper, two key clustering algorithms are implemented for power saving in MWSNs, which are the K-means algorithm and the Gaussian mixture models (GMM) algorithm. A comparison between the results is made. A brief description of K-means and GMM algorithms is demonstrated in section 2. The complexity analysis of the proposed clustering algorithms is presented in section 3. The evaluation section is presented in section 4. Section 5 includes results and discussion. Finally, the conclusion is stated in section 6.

2. THE RELATION OF CLUSTERING TECHNIQUE AND ENERGY SAVING

Many challenges might face clustering in MWSN such as the number of devices or sensor nodes in MWSN and the diversity of these devices, so each device or sensor node has its distinguished features. The proposed model in this work was developed to address the aforementioned clustering issue and circumvent difficulties with MWSN clustering. In reality, the primary task of the suggested model is to identify an effective

Bulletin of Electr Eng & Inf, Vol. 12, No. 6, December 2023: 3727-3735
clustering algorithm to combat the dispersion and power loss within the MWSNs that was produced as a result of the battery consumption and the challenge of supplying continuous energy, especially for mobile sensors. Clustering methods are unsupervised techniques; thus, the input points will not be labeled at the same time. As a result, the problem solution will depend on the algorithm’s expertise from analyzing similar problems throughout the training process. Two primary categories of clustering algorithms are offered in the literature, hard, and soft clustering, such as [27]–[31]: i) K-means algorithm and ii) GMM algorithm. We will explain the K-means algorithm and the GMM algorithm in the next paragraphs.

2.1. Gaussian mixture models algorithm

A function called a Gaussian mixture is made up of several Gaussians, each of which is denoted by the notation \( k \in \{1, \ldots, K\} \), where \( K \) is the number of clusters in our dataset. The following parameters are the components of each Gaussian \( k \) in the mixture:
- An average \( \mu \) that characterizes its center.
- The covariance's width. In a situation of multi variables, this is equivalent to the dimensions of a defined ellipsoid.
- The probability of mixing defines the shape of the Gaussian function.

In this study, the proposed technique has been chosen to be the GMM algorithm. This clustering technique is characterized by its high efficiency to form clustering rings around the moving nodes in the wireless sensors network. Hence, Gaussian clustering is a suitable nomination for our proposed model due to its high efficiency and speed for accomplishing clustering groups among the moving wireless sensors in the communication network. The GMM algorithm will satisfy as (1) [32]–[34]:

\[
G(x_i) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x_i-\mu_i)^2}{2\sigma^2}}
\]

where \( x_i \) is the random variable, \( \sigma^2 \) is the variance, and \( \mu_i \) is the mean value. Also:

\[
\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_i)^2
\]

and:

\[
\mu_i = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_i)
\]

For many distributions, the coefficient of mixture for \( k \)-th distribution. To predict the parameters by the algorithm maximum log-likelihood (MLL), calculated as (4):

\[
P(x) = \sum_{k=1}^{K} G(x_k) \pi_k
\]

where \( \pi \) is the weight of probability distribution. The random variable can be determined in the following expression:

\[
\gamma_k(x) = P(k|x) = \frac{p(X|k)\pi_k}{\sum_{k=1}^{K} p(X|k)\pi_k}
\]

The updated equations are as (6) and (7):

\[
\mu_i = \frac{\sum_{k=1}^{K} \gamma_k x_i}{\sum_{k=1}^{K} \gamma_k}
\]

\[
\pi_k = \frac{1}{N} \sum_{i=1}^{N} \gamma_i
\]

The probability of finding a certain cluster set of nodes will be as presented in (1) based upon the random node variable \( x_i \) and on the overall WSN nodes variance and mean [31]–[35]. The flowchart of the Gaussian mixture clustering algorithm is shown in Figure 1.

2.2. K-means algorithm

It is one of the common unsupervised learning algorithms. The dataset firstly is unlabeled. Then it is split into many clusters. The \( K \) of the K-means algorithm refers to the minimum number of clusters, which
should be created, this number must be pre-defined. For instance, if K=2, there will be two clusters, if K=4, it means that 4 clusters will be generated, and so on. The unlabeled dataset is divided into k different clusters using an iterative process. Each cluster comprises just one dataset and has a unique set of properties. It enables us to classify the data into various sets and provides a practical technique to determine the cluster of unlabeled datasets quickly and accurately. The method is based on the centroid of the clusters. The main objective of K-means is to decrease the distances between points and centers of clusters. The steps of the K-means algorithm of n input data points x₁, x₂, x₃, ..., xₙ at K number of clusters as:

a. Choose K points either at random or the first K from the dataset to serve as the starting centroids. In the dataset containing the identified K points.

b. Calculate the Euclidean distance between each point (cluster centroids).

c. Use the calculated distance in the second step, and assign every point of data to the closest cluster centroid.

d. Compute the mean of the points in each cluster group to determine the new centroid.

e. Repeat steps b through d as many times as necessary, up until the centroids stay the same.

The Euclidean distance between two points in space is (8):

\[ d(p, q) = \sqrt{(q₁ - p₁)^2 + (q₂ - p₂)^2} \]  

(8)

Suppose p(p₁, p₂), q(q₁, q₂) are two points in the space. If every centroid of the cluster is denoted by cᵢ, then every point of data x is assigned to a cluster depending on (9):

\[ \text{arg min} \{ \text{dist} (cᵢ, x) \} \]  

(9)

Then find the new centroid from the clustered group of points (10):

\[ cᵢ = \frac{1}{|Sᵢ|} \sum_{xᵢ \in Sᵢ} xᵢ \]  

(10)

Sᵢ is the set of all points assigned to the i_th cluster.

The flowchart of K-means is shown in Figure 2. It is a hard clustering approach, which means that it will identify each point with one and only one cluster. This is a key feature. This approach has a drawback in that no probability or uncertainty measure indicates the degree to which a data point is related to a given cluster. So how about switching from a hard clustering to a soft one? Exactly this is what GMM, or just GMMs, aim to achieve. Let’s now talk more about this approach.

---

**Figure 1. Gaussian mixture clustering algorithm flowchart**

**Figure 2. K-means flowchart**
3. COMPLEXITY ANALYSIS OF THE PROPOSED CLUSTERING ALGORITHMS

The computational complexity or latency in decision-making is one of the key challenges in computer science, particularly with critical time applications. The computational complexity of the K-means algorithm mainly involves two steps: the assignment step and the update step. In the assignment step, each data point is assigned to the nearest cluster center based on a distance metric (often Euclidean distance). In the update step, the cluster centers are recalculated as the mean of all data points assigned to that cluster. These steps are iteratively performed until convergence, which typically happens when the cluster centers no longer change significantly. The time complexity of each iteration is $O(nkd)$, making K-means computationally efficient, especially when the number of iterations is relatively small, consider n, d, and k are the number of points, dimensions, and centers, respectively [36]. On the other hand, the GMM is a probabilistic model used for density estimation and clustering. In GMM, each cluster is represented as a Gaussian distribution, and the model parameters include the means, covariances, and mixture weights of these Gaussians. The paragraph mentions that the computational complexity of the GMM algorithm is $O(nkd^3)$ [37], consider n data points, k Gaussian components, and d dimensions of the GMM algorithm. This complexity primarily arises from the need to estimate the covariance matrices for each Gaussian component and their inverses; hence, K-means requires less computing power per iteration than a GMM per iteration. In summary, understanding the computational complexities of clustering algorithms like K-means and GMM is essential for selecting the appropriate algorithm for a given application. K-means is often preferred when computational efficiency is a priority, while GMM offers greater flexibility at the cost of increased computational requirements.

4. EVALUATION METHOD

In this section, an evaluation method is proposed to measure the performance of both algorithms (K-means and the GMM) in saving energy in MWSN. Figure 3 illustrates a flowchart of the evaluation method. For fair evaluation, the input parameters represent the graph parameters. The evaluation metrics mentioned in the context of energy saving and complexity can be assessed using specific methodologies and criteria which is calculated after the clustering process as depicted in Figure 3. Here is an explanation of how these metrics may be evaluated:

a. Energy saving: energy consumption monitoring: measure the energy consumption before and after implementing a both algorithms to determine the amount of energy saved.

b. Complexity analysis of the proposed clustering algorithms: the computational complexity or latency in decision-making is one of the key challenges in computer science, particularly with critical time applications. Complexity analysis involves comparing the proposed clustering algorithms with existing or alternative methods. This comparison can be based on metrics such as execution time.

---

Figure 3. The proposed evaluation models
5. RESULTS AND DISCUSSION

In design, the proposed evaluation technique is conducted for both algorithms (K-means and GMM) of MWSN with parameter settings shown in Table 2. The MWSN is deployed using the graph topology (undirected graph topology). Figure 4 shows the minimum energy in the distribution nodes in a rectangular layout. The figures between nodes are similar to rectangles. It can be shown as rectangular or circular as in Figure 5. Figure 6 shows how nodes’ position changes during the clustering process when searching for the closest cluster (in this image, it is cluster 5 and the nodes around it), where the x-axis and y-axis represent the dimension of the MWSN.

Table 2. Parameters for simulation

<table>
<thead>
<tr>
<th>Network parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation area</td>
<td>100x100 m²</td>
</tr>
<tr>
<td>Number of clusters (k)</td>
<td>5</td>
</tr>
<tr>
<td>Sensors number</td>
<td>50</td>
</tr>
<tr>
<td>BS location (x, y)</td>
<td>(50, 50)</td>
</tr>
<tr>
<td>The initial energy of nodes ($E_0$)</td>
<td>9 J</td>
</tr>
<tr>
<td>Data packet size</td>
<td>2,000 bits/packet</td>
</tr>
<tr>
<td>Rounds number</td>
<td>4,500</td>
</tr>
<tr>
<td>Mobility speed</td>
<td>5 m/s</td>
</tr>
</tbody>
</table>

Figure 4. MWSN minimum energy distribution nodes (rectangular layout)  
Figure 5. MWSN minimum energy distribution nodes (circular layout)  
Figure 6. WSN searching clustering energy distribution nodes cluster 5
After obtaining the results from the implementation program, it is preferable to compare these results. The comparison results have been illustrated in Figure 7. The decision boundary's form is the first obvious distinction between K-means and GMM. With a covariance matrix, elliptical borders can be created with GMM rather than circular bounds with K-means. Hence, GMM is a little more flexible than K-means.

Figure 7 shows that the consumption power will remain 100% using K-means for some time and then will be decreased to reach 40%. While in GMM the power consumption will decrease directly from 90% to reach 30% after a period. It is worth stating that in comparison to other clustering algorithms, GMM are typically less scale-sensitive. Thus, it might not need to rescale the variables before utilizing them for clustering. Figure 8 shows that the average saving power ratio between GMM and K-means is 60.52% and that means GMM is better than K-means since it can save power more than K-means by a ratio of 60.52% and that is because of its accurate clustering techniques.

Figure 7. Comparison between MWSN best energy-consuming performances

Figure 8. Average saving power

6. CONCLUSION

In this paper, the power saving of wireless sensors network based on clustering techniques is addressed. For efficient energy saving, the GMM algorithm can be considered in the MWSN as it showed a significant improvement in the power saving in MWSN. The results present that the power saving rate is up to 92% at 4,500 rounds compared to the K-means algorithm. This means that using the GMM algorithm will decrease the amount of power consumed over time. On the other hand, the computational overhead of GMMs is higher compared to K-means. Hence, there is a trade-off between the power saving and the corresponding complexity of each algorithm.
REFERENCES
A comparative study of Gaussian mixture algorithm and K-means algorithm for … (Iman Ameer Ahmad)


BIOGRAPHIES OF AUTHORS

Iman Ameer Ahmad received her bachelor’s degree in Computer Engineering from Department of Control and Systems Engineering, University of Technology-Iraq in 2006 and a master’s student is currently in the Department of Control and Systems Engineering, Computer and Control Branch at the University of Technology, Baghdad, Iraq. She works at Department of Basic Science, College of Dentistry, University of Baghdad. She can be contacted at email: cse.21.11@grad.uotechnology.edu.iq.

Muna Mohammed Jawad Al-Nayar received the B.Sc., M.Sc., Ph.D. in Computer Science from University of Technology Baghdad-Iraq in 1988, 1994, and 1999 respectively. She has been a faculty member since 1994. Her research interests include artificial intelligence, distributed systems, and network management. She can be contacted at email: Muna.M.Jawad@uotechnology.edu.iq.

Ali M. Mahmood received a B.Sc. degree in Computer Engineering from the University of Technology-Baghdad in 2005, and an M.Sc. degree in Computer Engineering from Al-Nahrain University in 2010. He received a Ph.D. degree in Wireless Telecommunications from the University of Salford in the UK in 2018. His research interests include IoT, WSNs, cellular networks, scalability, computational complexity minimization, and end-to-end latency. He can be contacted at email: Ali.M.Mahmood@uotechnology.edu.iq.