



NEMOCED: New Energy Model and Optimal Cluster Estimation Based on Density to Increase Lifetime in Wireless Sensor Networks

Amir Abbas Baradaran¹ · Fahimeh Rabieefar²

Received: 6 December 2020 / Revised: 11 March 2022 / Accepted: 9 August 2022 / Published online: 6 September 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

In recent years, many methods have been presented for clustering in wireless sensor networks (WSNs). Some areas have high density, and the random distribution of the nodes reduces the clustering quality. Moreover, the number of clusters is manually determined before clustering. In this paper, a new clustering algorithm called NEMOCED is presented based on the node distribution. In the NEMOCED, the best cluster head is selected according to the node distribution. Moreover, we propose a new energy model to estimate proper clusters. One of the main features of the energy model is selecting the proper clusters. It is performed based on the number of nodes and the network size. In each round, two cluster heads are selected by using the tree structure. Finally, we introduce five criteria for assessing the quality and accuracy of the proposed algorithm. The NEMOCED can perform the clustering based on the local density of nodes and choose more proper cluster heads in high-density areas. The simulation results demonstrate that the NEMOCED can significantly improve lifetime and energy consumption. Furthermore, the simulation results show that the NEMOCED algorithm has good adaptability and works well under different network lifetime definitions. All the results prove that the NEMOCED algorithm has the advantage of being suitable and efficient for large-scale WSN applications.

Keywords Wireless sensor networks · Clustering · Density · Energy model · Cluster estimation

1 Introduction

Wireless sensor networks (WSNs) are significant in applications such as traffic control, forest fire control, environment tracking, control of patient status [1–5]. The node distribution in WSNs is random or deterministic [6, 7]. Each node collects data and transmits them to a base station (BS). Finally, the BS processes the collected data. One of the major issues in WSNs is energy consumption and network lifetime [8–11]. The initial energy of nodes is low and the sensors cannot be recharged [12]. Therefore, the initial energy of nodes is quickly drained during receiving and sending of data. So, they are gradually removed from the network [13, 14].

So far, several clustering algorithms have been proposed to increase lifetime and improve energy consumption in the WSN [15–17]. In clustering, several nodes are grouped according to the common properties. In each cluster, a node with better conditions is selected as the cluster head (CH). The CH collects data from cluster members and sends it to the BS. The data collected by CH can be transmitted in a single-hop or multi-hop manner [18–22]. In the single-hop, CHs transfer the collected data to the BS directly. In the multi-hop, the CHs transfer the collected data to the BS by using other CHs [23]. The single-hop manner is proper for regions where nodes' distance is low. Moreover, the multi-hop manner is suitable for large-scale WSNs. Since the node distribution is random, the major problem of the clustering is the node density [20–22]. In the other words, the clusters formed in WSN are high-density or low-density. The high-density clusters have more nodes, and the energy of the CHs will quickly drain compared to low-density clusters. Most clustering algorithms do not consider the intra-cluster density that reduces the quality of the cluster. Therefore, these methods are not reliable. In density-based schemes, clusters are high-density regions that have separated from

✉ Amir Abbas Baradaran
a_baradaran@sharif.edu

¹ Faculty of Computer Science and Engineering, North Tehran Branch, Islamic Azad University, Tehran, Iran

² Faculty of Electrical and Computer Engineering, University of Kashan, Kashan, Iran

low-density areas. There are criteria such as entropy, intra-cluster, and inter-cluster distances to ensure cluster quality. The separate use of these criteria does not guarantee the quality of the clusters [24–26]. For example, in grid-based clustering [27–30], the nodes are divided into multiple regions called Grid. Then, clustering is performed based on the number of nodes in each grid. The number of nodes in each grid is the only clustering criterion. Therefore, these methods are not efficient because the number of the nodes in each cluster (or grid) may be low but their initial energy is high and vice versa. The number of clusters at the beginning of the clustering process is another. In some methods, such as k-means and c-means, the number of clusters is manually determined before the clustering process [31, 32]. For this reason, these methods are not very secure.

The paper presents a new energy model and by using this model, we determine the number of proper clusters (or proper cluster heads). The clustering is performed using the local density of each node. In addition to the main CH, the successor CH also exists in each cluster. The successor CH is elected to support the main CH. The main advantage of our algorithm is determining the number of clusters based on the number of nodes and the network size. In the NEMOCED, the quality of clustering and its accuracy is ensured by using five criteria. The other advantages are:

1. Determining the number of proper clusters using a new energy model, network size, and the number of nodes.
2. Performing the clustering process according to the local density of each node.
3. Selecting the best node in each cluster as the CH.
4. Using a new tree structure to determine the main and the successor CHs
5. Determining the high-density and low-density regions and selecting the more proper CH in the high-density regions.
6. Presenting the five criteria to ensure the cluster quality and accuracy

The rest of the paper is organized as follows: Sect. 2 briefly reviews related works. Section 3 provides our system model, network model, and radio energy model. The procedure of the NEMOCED algorithm is presented in Sect. 4. Section 5 presents the simulation results and a comparison with the existing algorithms. Finally, Sect. 6 concludes this paper with some discussion of future work.

2 Related Works

The energy consumption issue was addressed very early in the Low Energy Adaptive Clustering Hierarchy protocol (LEACH) [33]. It is a hierarchical, self-organization, and

single-hop protocol. In LEACH, the CH selection is randomly performed based on the threshold $T(n)$. It is defined as follows:

$$T(n) = \begin{cases} \frac{P}{1 - P \left(r \bmod \frac{1}{P} \right)} & \text{if } n \in G \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where P : the percentage of the CH number, r : is the current round number, G : the set of nodes that have not been elected as CH for the past $1/P$ rounds, n : the number of nodes.

After selecting CHs, they broadcast a message to other nodes. The common nodes are connected to their corresponding CHs based on the lowest power required to connect and the received signal strength indicator (RSSI). The CH role alternates among all of the nodes. So, each node will have a chance to be the CH. Each node selects a random number between 0 and 1. If this random number is less than $T(n)$, then the node is selected as the CH. The CH rotation between the nodes results in load balancing, energy consumption balance, and uniform distribution of the energy. Despite these benefits, LEACH has problems. Firstly, the CH is randomly selected and if a node is the CH in the current round, it cannot be the CH in the next round. Therefore, low-energy nodes also can be CH. This problem is worse in high-density regions. Secondly, the LEACH assumes the communication range of each CH is high and they can directly transfer data to the BS. This is not a realistic assumption because in most cases the BS is not available for all nodes. Third, LEACH uses a single-hop method that is not proper for large-scale networks.

In the next years, different methods were introduced that could improve the LEACH. One of these methods is the ERP algorithm [34]. It is performed in homogeneous WSNs and has a good lifetime and stability. After the formation of clusters, routing is performed in a multi-hop manner. The ERP algorithm produced a 42% lifetime improvement compared with the LEACH protocol. One of the main benefits of the ERP is the new fitness function. It is calculated based on the genetic algorithm. Logambigai et al. presented the EEGBR protocol. It performs multiple clustering based on the grid and fuzzy rules. In the EEGBR, routing is performed through a novel feature called grid coordinator (GC). The results simulations demonstrate that EEGBR can reduce the hops and energy consumption by using the fuzzy rules. The EEGBR has three phases: cluster formation, grid coordinator selection, and grid-based routing. The grid coordinator is determined by using fuzzy rules. The fuzzy inference system uses three parameters: the residual energy of the nodes, the motion model of the nodes, and the distance to sink. In [35], a protocol called OCM-FCM is proposed that uses the fuzzy c-means algorithm for clustering. The OCM-FCM uses a single-hop approach for intra-cluster

communication. Furthermore, it uses an inter-cluster manner to transfer data to the BS. In the OCM-FCM, the number of clusters is determined in advance that is not proper. If the node distribution is random, then the OCM-FCM is not a proper method because some areas have a higher density in random distribution. Another problem in the OCM-FCM is the CH selection. It is only performed based on the residual energy of the nodes. Meng et al. [30] have proposed a grid-based protocol called GBRR. It can solve the problem of CH overload by grid-based clustering. This method improves the quality of the node-to-node link between nodes in the WSN. In, authors have proposed a new routing algorithm called ENEFC HRML that has three phases: hierarchical routing using cluster identification (HRCI), hierarchical routing using multi-hop (HRMH), and hierarchical routing using multilevel (HRML). The simulation results show that the HRML phase has high efficiency in energy consumption. The FUZZY-TOPSIS [37] presents a new technique based on fuzzy rules in which CH selection is according to five criteria. By using the five criteria, the common node decides whether to be the CH or not. Then the common nodes are linked to the corresponding CHs based on the maximum RSSI value and the smallest distance. In the MCFL [36] has been proposed a new clustering protocol to reduce energy consumption and increase network lifetime. The MCFL has three rounds that in each round clustering is separately performed. In other words, every three rounds produce different clusters. In [37] has been proposed a hybrid protocol based on density and threshold. The hybrid protocol is based on density and a threshold called C-DTB-CHR and C-DTB-CHR-ADD. It uses the LEACH, T-LEACH, and MT-CHR deficiencies to select the CHs. The main benefit of the C-DTB-CHR is that all of the nodes do not participate in the data transmission. In other words, a density-based approach has been suggested for nodes that will cooperate in the data transmission process. Jinyu Ma et al. [38] introduced a protocol based on an ant colony called ADCAPTEEN. It selects two CHs: the MCH (Master Cluster Head) and VCH (Vice Cluster Head). The MCH and VCH have co-work to collect, aggregate, and send data. The simulation results show that the ADCAPTEEN protocol has better scalability compared with the APTEEN protocol. Therefore, it is proper for large-scale WSNs. The MOFCA protocol [39] is another protocol that uses the residual energy of the nodes, node distance to the BS, and node density. The MOFCA protocol determines the tentative and final CHs via local decision-making. Hua-mei et al. [40] proposed an energy-efficient and non-uniform clustering protocol based on improved shuffled frog leaping algorithm to increase lifetime in wireless sensor networks. Their method is adopted to divide the sensor nodes into clusters and finds the optimal cluster head. OK-means is another method that improves the position of nodes using the K-means algorithm [41]. The OK-means algorithm uses

single step and multi-hop manners for intra-cluster and inter-cluster communications [41].

3 System Model

3.1 Network Model

In the NEMOCED, there are N sensor nodes, and all of the nodes are randomly distributed in a square area measuring $n * n$. The distribution of sensor nodes is uniform and there is only one BS. In different scenarios is assumed the BS has different positions. The BS position is outside of the nodes' distribution. Furthermore, the following assumptions are considered:

1. Nodes are aware of their position, the position of other nodes, and the BS position. It is performed using the global positioning system (GPS) or positioning algorithms.
2. The nodes are stationary in the environment that their initial energy is equal.
3. Wireless communication between nodes is symmetrical.
4. There is a medium access control (MAC) layer that prevents interference when sending or receiving messages.
5. Energy, memory, and computational power of the BS are infinite.

3.2 Radio Energy Model

Radio energy consumption is measured based on the distance of the transmitter and receiver. The radio energy model is calculated by using an energy dissipation model [42]. Therefore, to transmit an L -bit message over a distance d , the energy consumption is defined as follows:

$$\begin{cases} E_{rcv}(L, d) = L * (E_{elect} + \epsilon_{fs} * d^2) & d < d_c \\ E_{rcv}(L, d) = L * (E_{elect} + \epsilon_{mp} * d^4) & d \geq d_c \end{cases} \quad (2)$$

where d_c is the crossover distance. ϵ_{fs} and ϵ_{mp} are power consumption of the free space propagation and power consumption of multipath propagation, respectively. They are based on the sensitivity of the sender and noise shape. E_{elect} is the energy/bit consumed by the transmitter/receiver electronics. Finally, the energy consumption to receive an L -bit message is:

$$E_{RX} = L * E_{elect} \quad (3)$$

4 Proposed Method

4.1 Estimation of Proper Number of Clusters

One of the main issues in clustering is to estimate the number of proper clusters before the beginning of the clustering. Therefore, we propose a relation based on the energy consumption model. We compute the number of proper clusters based on this relation. Since the number of clusters indicates the number of cluster heads, we can determine the number of proper cluster heads. Energy consumption in each cluster ($E_{cluster}$) includes the energy consumption of common nodes ($E_{common-node}$) and the cluster heads (E_{CH}). Energy consumption of common nodes includes both the energy required to receive data (E_r) and the energy needed to send data to CH (E_s) over a distance d . The cluster head has three types of energy:

1. The energy required to receive data from the common nodes (E_{rcv-CN}) in each cluster
2. Aggregation energy of the collected data (E_{agg})
3. The energy required to transmit the aggregated data from CH to a higher level over the distance d meter (E_{send}). E_{send} is based on the type of transmission (single-hop or multi-hop). In the single-hop manner, the CH data are directly transmitted to the BS. In the multi-hop manner, the CH data are transmitted to the BS via higher-level CHs.

Therefore, energy consumption in each cluster is defined as follows:

$$E_{cluster} = E_{CH} + E_{common-node} = (E_{rcv-CN} + E_{agg} + E_{snd}) + (E_r + E_s) \quad (4)$$

Assuming that there are k clusters, the total energy consumption of the network is defined as follows:

$$E_{total} = k * E_{cluster} = k * (E_{CH} + E_{common-node}) \\ = k * ((E_{rcv-CN} + E_{agg} + E_{snd}) + (E_r + E_s)) \quad (5)$$

If N_{Common} is the total number of common nodes in each cluster, then the energy required to receive data is defined as follows:

$$E_{rcv-CN} = L * E_{elect} * N_{Common} \quad (6)$$

Data aggregation energy is also defined as follows:

$$E_{agg} = L * E_a * N_{CH} \quad (7)$$

where N_{CH} and E_a are the number of clusters and energy needed to aggregate the received data, respectively. In the single-hop method, the energy required to receive data from a cluster head to the BS is defined as follows:

$$E_{snd} = L * E_{elect} + L * \epsilon_{mp} * d_{CH2BS}^4 \quad (8)$$

where d_{CH2BS} is the cluster distance to the BS. In the multi-hop method, the distance between the CH and the BS is greater than the CH distance to higher-level CH. Therefore, in Eq. (8), d_{CH2BS}^2 is used instead of d_{CH2BS}^4 .

The energy required for receiving and sending data by common nodes are defined as follows:

$$E_r = L * E_{elect} \quad (9)$$

$$E_{snd} = L * E_{elect} + L * \epsilon_{fs} * d_{CN2CH}^2 \quad (10)$$

where d_{CN2CH}^2 is the common node distance to the CH. We use d_{CN2CH}^2 because the distance between the common node and its cluster is less than the crossover distance. In other words, the relation of the common nodes and their cluster-head is single-hop. Therefore, the total energy consumption is:

$$E_{total} = k * E_{cluster} = k * [(L * E_{elect} * N_{Common}) + L * E_a * N_{CH}] \\ + (L * E_{elect} + L * \epsilon_{mp} * d_{CH2BS}^4) \\ + (L * E_{elect} + L * \epsilon_{fs} * d_{CN2CH}^2) \quad (11)$$

On the one hand, each cluster has one cluster head. On the other hand, the number of common nodes per cluster is on average equal to $N/k-1$. Therefore, the total number of cluster heads is k . Finally, the total number of nodes will be $k * (N/k-1)$ or $(N-k)$:

$$E_{total} = k * [L * E_{elect} * (\frac{N}{k} - 1) + L * E_a * k + L * E_{elect} \\ + L * \epsilon_{mp} * d_{CH2BS}^4 + L * E_{elect} + L * \epsilon_{fs} * d_{CN2CH}^2] \\ = k * L[E_{elect} * (\frac{N}{k}) + k * E_a + 2E_{elect} + \epsilon_{mp} * d_{CH2BS}^4 \\ + \epsilon_{fs} * d_{CN2CH}^2] \quad (12)$$

The proper value of the cluster (k_{opt}) is obtained by taking the derivative with respect to k in Eq. (12).

In the following, we propose a proper relation for the cluster head distance to the BS (d_{CH2BS}) and the common node distance to the cluster head (d_{CN2BS}). Since our clustering is based on the primary grids, we consider the nodes are distributed in square $s * s$. In later sections, we will describe an equation to compute the edge of the initial grids.

Lemma 1 *The average distance between the nodes in a square with diameter d and side length s is defined as follows:*

$$\frac{2 + 5\ln(\sqrt{2} + 1) + 2)\sqrt{2}}{30} * d, d = s\sqrt{2} \quad (13)$$

Proof Let two nodes with coordinates (x_1, y_1) and (x_2, y_2) are in a square to the side length s . The nodes are also independent of the other nodes. The average distance between the nodes is defined as follows:

$$d_{avg} = \frac{1}{s^4} \int_0^s \int_0^s \int_0^s \int_0^s \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} dx_2 dx_1 dy_2 dy_1 \tag{14}$$

According to Fig. 1 and by reducing integral in Eq. (14), d_{avg} in Eq. (15) is define as follows:

$$d_{avg} = \frac{4}{s^4} \int_0^s \int_0^{y_1} \int_0^s \int_0^{x_1} \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} dx_2 dx_1 dy_2 dy_1 \tag{15}$$

By defining $z_1 = x_1 - x_2$, $z_2 = x_1 + x_2$, and Jacobian determinant $j = \frac{\partial(x_1, x_2)}{\partial(z_1, z_2)} = \frac{1}{2}$:

$$x_1 = \frac{z_1 + z_2}{2} \tag{16}$$

$$x_2 = \frac{z_2 - z_1}{2} \tag{17}$$

Figure 2 shows the integral regions with new coordinates. According to Fig. 2, d_{avg} can be described as follows:

$$\begin{aligned} d_{avg} &= \frac{2}{s^4} \int_0^s \int_0^{y_1} \int_0^s \int_0^{2s-z_1} \sqrt{z_1^2 + (y_1 - y_2)^2} dz_2 dz_1 dy_2 dy_1 \\ &= \frac{2}{s^4} \int_0^s \int_0^{y_1} \int_0^s (2s - 2z_1) \sqrt{z_1^2 + (y_1 - y_2)^2} dz_1 dy_2 dy_1 \end{aligned} \tag{18}$$

We define $w_1 = y_1 - y_2$ and $w_2 = y_1 + y_2$. So:

$$\begin{aligned} d_{avg} &= \frac{4}{s^4} \int_0^s \int_0^s (s - z_1)(s - w_1) \sqrt{z_1^2 + w_1^2} dz_1 dw_1 \\ &= \frac{8}{s^4} \int_0^s \int_0^{z_1} (s - z_1)(s - w_1) \sqrt{z_1^2 + w_1^2} dz_1 dw_1 \end{aligned} \tag{19}$$

Finally, by changing the integration interval to the polar coordinates $z_1 = r \cos \theta$ and $w_1 = r \sin \theta$, the following equation is obtained:

$$d_{avg} = \frac{8}{s^4} \int_0^{\frac{\pi}{4}} \int_0^{\frac{s}{\cos \theta}} (s - r \cos \theta)(s - r \sin \theta) r^2 dr d\theta \tag{20}$$

By solving Eq. (20):

Fig. 1 Reducing the integration area from a square to a triangular area

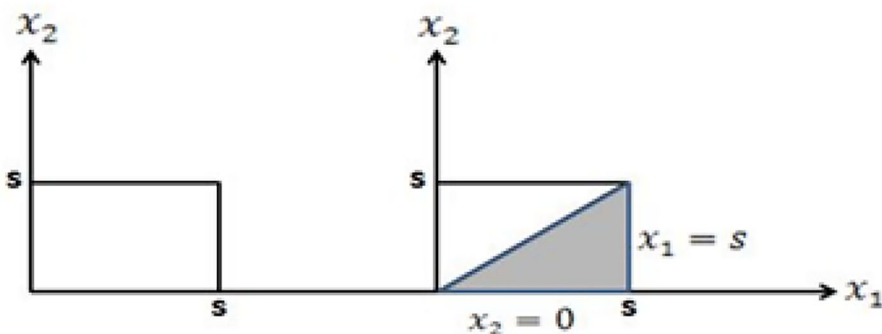
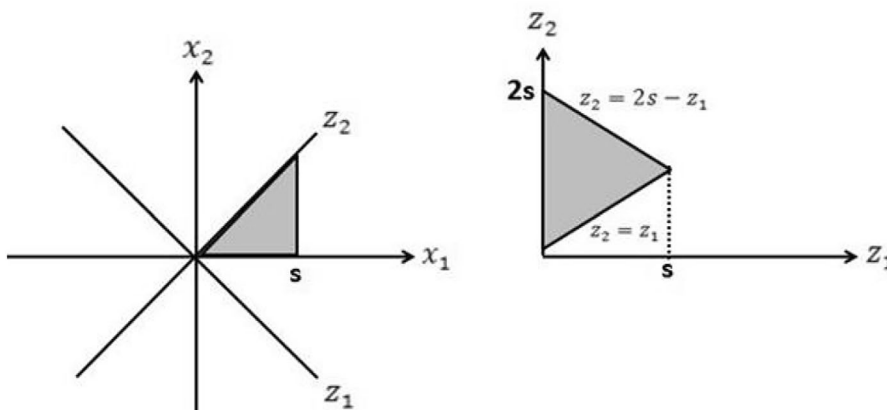


Fig. 2 Converting variables x_1 and x_2 to z_1 and z_2



$$d_{avg} = \frac{8}{s^4} * \frac{2 + 5\sqrt{2}\ln(\sqrt{2} + 1) + 2\sqrt{2}}{120\sqrt{2}} \tag{21}$$

$$= \frac{2 + 5\ln(\sqrt{2} + 1) + 2\sqrt{2}}{30} * d$$

By calculating the Eq. (21), d_{avg} will be approximately $0.36869*d$. We will use $d_{avg} = 0.36869 * d$ instead of d_{CH2BS} in the proposed method.

Lemma 2 *The distance between two random nodes in a square with side length p is:*

$$D_{avg} = \left(\frac{1}{3}\ln(1 + \sqrt{2}) + \frac{1}{15}(2 + \sqrt{2})\right) * p \tag{22}$$

Proof We first make a proof for the rectangular region with side lengths a and b and then extend it to the square area.

Let two points (x_1, y_1) and (x_2, y_2) are randomly distributed in the interval $(0, a)$ and $(0, b)$, respectively. The probability distribution function is defined as follows:

$$F_a(t) = prob(x_1 - x_2)^2 \leq tp \tag{23}$$

With respect to the probability distribution function, the corresponding density function is defined as:

$$f_a(t) = \frac{dF_a(t)}{dt} \tag{24}$$

We show the density corresponding to $G(s) = prob((x_1 - x_2)^2 + (y_1 - y_2)^2 \leq s)$ by $g(s)$, which is obtained by convolving of f_a and f_b . Finally, the distribution function for the distance is $K(v) = H(v^2)$ with the density $K(v) = 2vh(v^2)$.

The probability density for $(x_1 - x_2)^2 + (y_1 - y_2)^2 \leq s$ is equal to the convolution of g from f_a and f_b :

$$g(s) = \int f_a(s - t)f_b(t)dt \tag{25}$$

Due to different domains of f_a and f_b , there are three main states:

$$g_1(s) = \int_0^s f_a(s - t)f_b(t)dt, 0 < s \leq a^2 \tag{26}$$

$$g_2(s) = \int_{s-a^2}^s f_a(s - t)f_b(t)dt, a^2 < s \leq b^2 \tag{27}$$

$$g_3(s) = \int_{s-a^2}^{b^2} f_a(s - t)f_b(t)dt, b^2 < s \leq a^2 + b^2 \tag{28}$$

After calculating the above equations:

$$g_3(s) = \int_{s-a^2}^{b^2} f_a(s - t)f_b(t)dt, b^2 < s \leq a^2 + b^2 \tag{29}$$

$$g_1(s) = -2\frac{\sqrt{s}}{a^2b} - 2\frac{\sqrt{s}}{ab^2} + \frac{\pi}{ab} + \frac{s}{a^2b^2} - 2\frac{\sqrt{s}}{a^2b}, 0 < s \leq a^2, g_2(s)$$

$$= -\frac{1}{b^2} + \frac{2}{ab} \arcsin\left(\frac{a}{\sqrt{s}}\right) + \frac{2}{a^2b} \sqrt{s - a^2} - \frac{1}{b^2} + \frac{2}{ab} \arcsin\left(\frac{a}{\sqrt{s}}\right)$$

$$+ \frac{2}{a^2b} \sqrt{s - a^2} - \frac{1}{a^2} + \frac{2}{ab} \arcsin\left(\frac{b}{\sqrt{s}}\right) + \frac{2}{ab^2} \sqrt{s - b^2}, a^2 < s \leq b^2 \tag{30}$$

$$g_3(s) = -\frac{\pi}{ab} - \frac{s}{a^2b^2}, b^2 < s \leq a^2 + b^2 \tag{31}$$

As stated earlier, the above calculations are for a rectangular region. In a square region, a is equal to b . So, $g_2(s)$ is omitted. Since s is the square of the distance, so the density of the distance $v = \sqrt{s}$ between two random points in a rectangle with the sides a and b is equal to:

$$g_v(v) = g(v^2) \frac{ds}{dv} = 2vg(v^2) \tag{32}$$

The expectation of the distance or average distance between two random points in a rectangle is defined as follows:

$$E[rect] = \int_0^{a^2+b^2} \sqrt{s}g(s)ds = \int_0^{\sqrt{a^2+b^2}} vg_v(v)dv \tag{33}$$

So:

$$E[rect] = \frac{a^2}{6b} \ln\left(\frac{b}{a} + \sqrt{1 + \frac{b^2}{a^2}}\right) + \frac{b^2}{6a} \ln\left(\frac{a}{b} + \sqrt{1 + \frac{a^2}{b^2}}\right)$$

$$+ \frac{1}{15}\left(\frac{b^3}{a^2} + \frac{a^3}{b^2}\right) + \left(3 - \frac{a^2}{b^2} - \frac{b^2}{a^2}\right) \sqrt{a^2 + b^2} \tag{34}$$

In a square area $a = b = p$, so:

$$D_{avg} = E[rect] = p * \frac{1}{3}\ln(1 + \sqrt{2}) + \frac{1}{15}(2 + \sqrt{2}) \approx 0.5214p \tag{35}$$

In the proposed model, we will use D_{avg} instead of d_{CN2CH} . According to lemma 1 and 2, the proposed energy model is defined as follows:

$$E_{total} = k * L \left[E_{elect} * \left(\frac{N}{k}\right) + k * E_a + 2E_{elect} \right.$$

$$\left. + \epsilon_{mp} * D_{avg}^4 + \epsilon_{fs} * d_{avg}^2 \right] \tag{36}$$

As mentioned earlier, there is on average N/k node per cluster. So:

$$E_{total} = k * L \left[E_{elect} * \left(\frac{N}{k} \right) + k * E_a + 2E_{elect} + \epsilon_{mp} * D_{avg}^4 + \epsilon_{fs} * \left(\frac{N}{k} * d_{avg} \right)^2 \right] \tag{37}$$

By taking the derivative in terms of k in Formula (37), the number of proper clusters (k_{opt}) is obtained.

For example, assume 100 nodes are randomly distributed in a square area of $120 * 120 \text{ m}^2$. Let the size of the grid length is equal to $40 * 40 \text{ m}^2$. So:

$$E_{total} = k * L \left[E_{elect} * \left(\frac{100}{k} \right) + k * E_a + 2E_{elect} + \epsilon_{mp} * (0.5214 * 120)^4 + \epsilon_{fs} * \left(\frac{100}{k} * 0.36869 * 40 * \sqrt{2} \right)^2 \right]$$

With respect to the simulation parameters (see Table 1) and k , k_{opt} is equal to 13. Therefore, the initial grid proposes 13 proper clusters. In the next section, we will discuss the clustering process with respect to k_{opt} and the selection of cluster centers. The cluster centers are based on the nodes' density.

4.2 Clustering Process

4.2.1 Pre-clustering

Assume that N sensor nodes have been distributed randomly, independently, and uniformly in a square area $L * L$. At first, the network area is divided into cells (or grids) and the side length of the grid is calculated by the following equation:

$$L_g = \alpha \left(\prod_{i=1}^d \left(\frac{h_i - l_i}{n} \right) \right)^{\frac{1}{d}} \tag{38}$$

where d and n are the dimensions of the area and the number of distributed nodes in the network, respectively. Since the distribution area is two-dimensional, d is equal to 2. l_i and h_i are the beginning and the end of the area. For example, if the nodes have been distributed in the square area measuring $0 * 120$, then l_i and h_i are 0 and 120, respectively. α is the adjustment parameter of the grid side and is defined as follows:

$$\alpha = \frac{h_i}{k_{opt}} \tag{39}$$

We consider the density of each cell (or grid) equal to the number of nodes in each grid. Therefore:

$$p_g = \text{Count}(n) \tag{40}$$

Now, with respect to k_{opt} , we reduce the number of cells to k_{opt} . This state is called the merge phase. In the merge phase, the total number of cells is reduced to k_{opt} . In other words, after the merge phase, the number of final cells will be equal to k_{opt} .

In the merge phase, the cells with the highest p_g are selected. The number of choices (p_g) will be according to k_{opt} . These cells are called "candidate cells" and the centers of gravity of the nodes in these cells (X_g, Y_g) are calculated according to the following relations:

$$X_g = \frac{\sum_{i=1}^{n_g} x_i}{n} \tag{41}$$

$$Y_g = \frac{\sum_{i=1}^{n_g} y_i}{n} \tag{42}$$

where x_i and y_i are node coordinates and n_g is the number of sensor nodes in each cell. After calculating the centers of gravity, the distance between nodes in adjacent grids to the centers of gravity of candidate grids is calculated. Any node in the adjacent grids which has the smallest distance to the gravity centers of the candidate grids will be a member of the grid. This phase is called pre-clustering. For example, suppose the energy model proposes $k_{opt} = 5$. Therefore, five candidate cells are: 4, 6, 11, 1, and 10. Centers of gravity in five candidate cells are calculated and the nodes of adjacent cells are appropriated to proper candidate cells according to the nearest distance to the centers of gravity of each candidate cell. Note that the output of the pre-clustering phase is to create cells to the number of k_{opt} . In the following, we discuss the final density-based clustering process.

4.2.2 Density-Based Clustering

In this section, we describe the final density-based clustering process. Let x_i, ρ , and δ_i are two-dimensional coordinates

Table 1 Simulation parameters

Parameter name	Parameter value
Number of nodes	100–200–300–500
Network size	100*100–200*200–300*300–500*500
Initial energy	0.5 J
Relative position of the base station	Variable
Simulation end condition	Number of nodes < 5
E_{elect}	50 nJ/bit
ϵ_{fs}	10 pj/bit/m ²
ϵ_{mp}	0.0013 pj/bit/m ⁴
Data aggregation energy	5 nj/bit/signal
Crossover distance	87 m

of each sensor node, local density function, and distance function, respectively. We show the distance between two points x_i and x_j with d_{ij} . The local density function is defined as follows:

$$\rho_i = \sum_j \mathfrak{N}(x) = \sum_j \mathfrak{N}(d_{ij} - d_c) \quad (43)$$

where d_c is crossover distance between two points. In the proposed algorithm, d_c is the maximum distance between a node with other nodes. If the number of nodes is low, then ρ_i is defined by the Gaussian kernel function as follows:

$$\rho_i = \sum_j e^{\left(-\frac{d_{ij}^2}{d_c^2}\right)} \quad (44)$$

The distance function is defined as:

$$\delta_i = \begin{cases} \max(d_{ij}) & : \text{if } \rho_i > \rho_j \\ \min(d_{ij}) & \text{otherwise} \end{cases} \quad (45)$$

After calculating ρ_i and δ_i , any node with the largest value $\|\delta_i - \rho_i\|$ has a better position than local density. So, it is selected as the center of the cluster. Note that d_{ij} is based on a matrix called the adjacency matrix.

4.2.3 Cluster Head Selection

In this section, by using a new tree structure and six criteria including the residual energy of the node, node distance to the BS, density (ρ), δ , node distance to the center of gravity in each cluster, and $\|\delta_i - \rho_i\|$ value, the cluster head is selected. The advantage of the new cluster head selection is to elect the original CH and the successor CH in each cluster. Upon completion of the energy of the original CH, the successor CH is substituted. The entropy and information gain are other criteria that play a significant role in our proposed approach. These two criteria increase the accuracy of the CH selection. For each node in the cluster, a table called the decision table is formed. It has six introduced criteria. Then the decision table is normalized and is calculated the entropy of each parameter. Entropy is calculated by using the following equation:

$$\text{Entropy}(D) = - \sum_{i=1}^c p_i * \log_2 p_i \quad (46)$$

In Eq. (46), D and C are set of cluster members and the number of clusters, respectively. p_i is the probability of belonging of nodes to their clusters. Information gain is defined as follows:

$$\text{InformationGain}(A) = \text{Entropy}(D) - \text{Entropy}(D)_A \quad (47)$$

$$\text{Entropy}(D)_A = \sum_{j=1}^v \frac{|D_j|}{|D|} * \text{Entropy}(D)_y \quad (48)$$

where v is the member's number of A and D_j is a part of the initial nodes for which A is equal to v_j . In the NEMOCED algorithm, the highest amount of information gain will be considered.

There are three phases to calculate the entropy. In the first phase, the decision table is formed according to the six introduced criteria. The computations of the six criteria have been discussed in the previous sections. In the second phase is performed the normalization of the decision table. Normalization uses the simple normalization method (or arithmetic normalization). The equation of simple normalization is defined as follows:

$$p_{ij} = \frac{X_{ij}}{\sum_{i=1}^m X_{ij}}, j = 1, 2, \dots, n \quad (49)$$

In Eq. (49), X_{ij} is a parameter that must be normalized. Also, m is the number of nodes in the decision table (or the number of nodes in each cluster).

In the third phase, the entropy of each parameter is calculated by using the following equation:

$$E_j = -k \sum_{i=1}^m p_{ij} * \log_2 p_{ij}, \quad i = 1, 2, \dots, n \quad (50)$$

where k is the entropy value of each parameter and is defined as follows:

$$k = \frac{1}{\log_2 m} \quad 0 \leq k \leq 1 \quad (51)$$

After performing the three phases, the parameters of the decision table are labeled. This work is performed to the BS and its base table (BT).

$$E_{Rs} = \begin{cases} \text{LOW} & E_{Rs} \leq \min \\ \text{MEDIUM} & \min < E_{Rs} \leq \text{Avg} \\ \text{HIGH} & E_{Rs} > \text{Avg} \end{cases} \quad (52)$$

$$d_{BS} = \begin{cases} \text{NEAR} & d_{BS} \leq \min \\ \text{AVERAGE} & \min < d_{BS} \leq \text{Avg} \\ \text{FAR} & d_{BS} > \text{Avg} \end{cases} \quad (53)$$

$$\rho = \begin{cases} \text{LOW} & \rho \leq \min \\ \text{MEDIUM} & \min < \rho \leq \text{Avg} \\ \text{HIGH} & \rho > \text{Avg} \end{cases} \quad (54)$$

$$\delta = \begin{cases} \text{GOOD} & \delta \leq \min \\ \text{FAIRLYGOOD} & \min < \delta \leq \text{Avg} \\ \text{BAD} & \delta > \text{Avg} \end{cases} \quad (55)$$

$$d_g = \begin{cases} \text{NEAR} & d_g \leq \min \\ \text{AVERAGE} & \min < d_g \leq \text{Avg} \\ \text{FAR} & d_g > \text{Avg} \end{cases} \quad (56)$$

$$\|\delta_i - \rho_i\| = \begin{cases} \text{LOW} & \|\delta_i - \rho_i\| \leq \min \\ \text{MEDIUM} & \min < \|\delta_i - \rho_i\| \leq \text{Avg} \\ \text{HIGH} & \|\delta_i - \rho_i\| > \text{Avg} \end{cases} \quad (57)$$

In the above equations, *min* and *max* are the smallest amounts and the average value in the decision table, respectively. The main task of the decision table is to identify the successor CHs. The original and successor CHs are determined by the new tree structure. To better understand, we show the details of the CH selection with two examples at Appendix 2.

5 Simulation Results

In this section, we present the results of the proposed method. The proposed NEMOCED algorithm has been evaluated using MATLAB and is compared with other protocols namely LEACH, ENEFC HRML, ADCAPTEEN, MOFCA, MT-CHR, C-DTB-CHR-ADD, EEGBR, ERP, OCM-FCM, and FUZZY-TOPSIS by the same initial values and the scenario with different parameters. Each sensor has initial energy of 0.5 J and is randomly distributed in the square $n * n$. The simulation parameters are summarized in Table 1.

The efficiency of the proposed algorithm is evaluated in the following cases, and its results are compared with previous works:

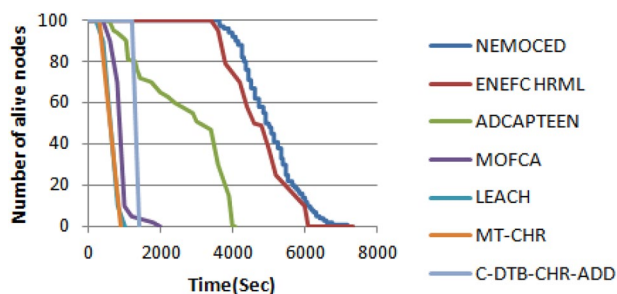


Fig. 3 Network lifetime in NEMOCED protocol and similar methods

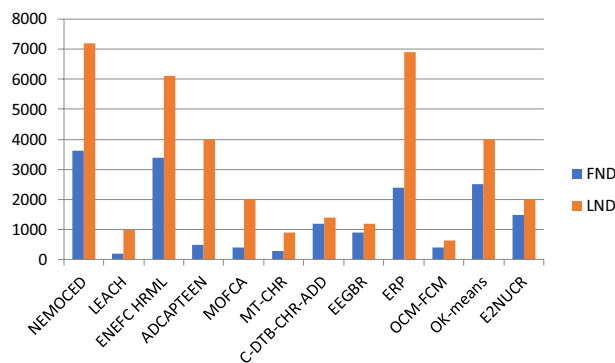


Fig. 4 FND and LND parameters in the NEMOCED protocol and similar methods

1. Network lifetime
2. The number of cluster heads (or clusters) and selection of them in high-density and low-density regions.
3. The Residual energy in each round and each cluster
4. Determining the number of clusters using the new energy model, network size, and the number of nodes
5. Measuring the quality and accuracy of clustering

5.1 Network Lifetime

One of the NEMOCED goals is reducing energy consumption and increase network lifetime. First Node Die (FND) and Last Node Die (LND) are two main parameters in network lifetime. Figure 3 illustrates the network lifetime compared to similar methods. Figure 4 indicates FND and LND in the NEMOCED and other approaches. From Figs. 3 and 4 it is clear that the NEMOCED protocol has a better lifetime than similar methods.

5.2 Number of Cluster Head

If the distribution of nodes is random, then more nodes may be located in some areas. Therefore, more cluster heads should be selected so that energy consumption and the loss

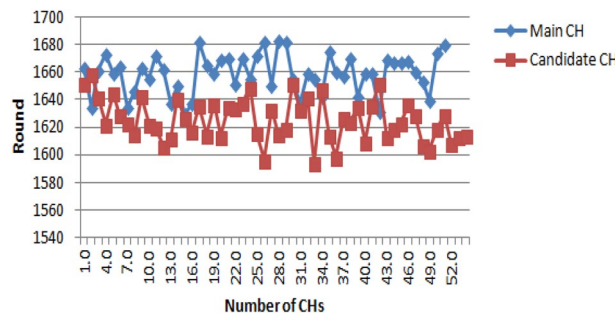


Fig. 5 FND and LND parameters in the NEMOCED protocol and similar methods

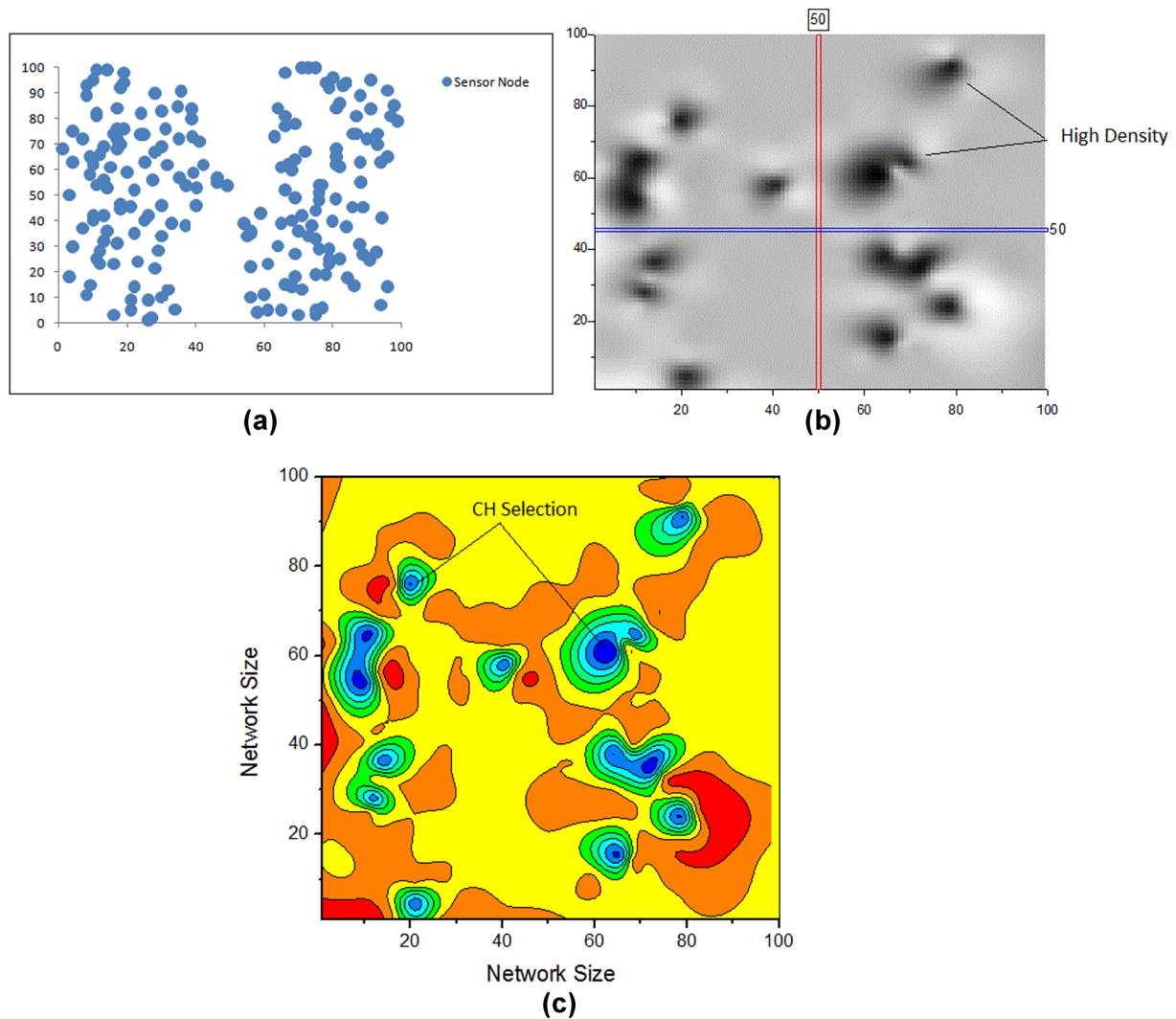


Fig. 6 a Random distribution of nodes, b-High density regions and c-Cluster heads selection in high density regions

percentage of nodes in the high-density regions can be well balanced. Figure 5 shows how to choose the original and successor CHs over 160 rounds. From Fig. 5, it is clear that the selection of the original and successor CHs has a proper balance in the environment. Figure 6 indicates the random distribution of 200 nodes, the high-density areas, and selected CHs. As Fig. 6 indicates, in high-density areas more CHs have been selected. The relation given in the preceding sections for calculating the number of clusters (relation 37) is based on the number of nodes and the network size. This is an advantage because, in most of the presented methods, the number of clusters (or cluster heads) is determined only by the number of nodes or network size. Also, in some methods such as k-means and C-means clustering, the number of clusters is predetermined. In the proposed method, if the network size is constant but the number of nodes in the network increases, then the number of proper clusters does not

Table 2 K_{opt} estimation based on network size, node's number, d_{avg} , and D_{avg}

Network size	D_{avg}	d_{avg}	Number of nodes	K_{opt}
120*120	62.568	14.7476	120	15
120*120	62.568	14.7476	100	13
120*120	62.568	14.7476	50	7
100*100	52.14	36.869	120	15
100*100	52.14	36.869	100	13
300*300	156.42	110.607	500	22
400*400	208.56	147.476	500	19
500*500	260.7	184.345	500	30

increase significantly and will increase by a small ratio. For example, in a network of $(120 \times 120) \text{ m}^2$ with 100 nodes, the

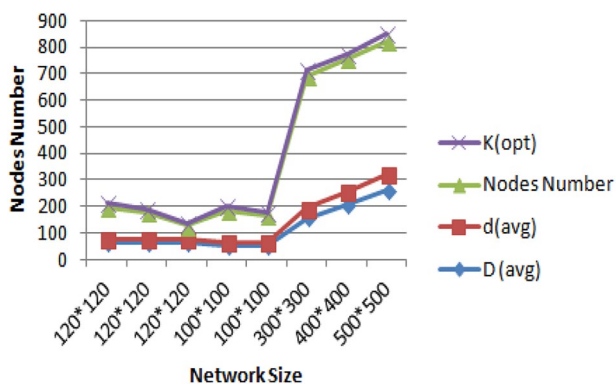


Fig. 7 The effect of the network size and nodes' number on K_{opt}

proper cluster number is 13. If there are 120 nodes in the same network, then the proper number of clusters will be equal to 15. Similarly, if the number of nodes is 50, then 7 clusters are suggested. In other words, the number of clusters in a network with 50 nodes is approximately equal to half of the proposed clusters with 120 nodes. This means that the proposed algorithm has high accuracy. Table 2 shows the various comparisons between different states. Figure 7 illustrates the effect of increasing the number of nodes and increasing the network size on the proper number of clusters, D_{avg} , and d_{avg} . From Fig. 7, it is clear that with increasing the number of nodes and the size of the network simultaneously, the number of proper clusters has been increased.

5.3 Residual Energy

If at the end of each round the amount of the residual energy is high, then the total lifetime will be better. One of the important advantages of the proposed method is that the residual energy level is more uniform than other methods. Figure 8 indicates the amount of residual energy per 1000 rounds of the proposed protocol compared with similar methods. As Fig. 8 shows, the residual energy in the

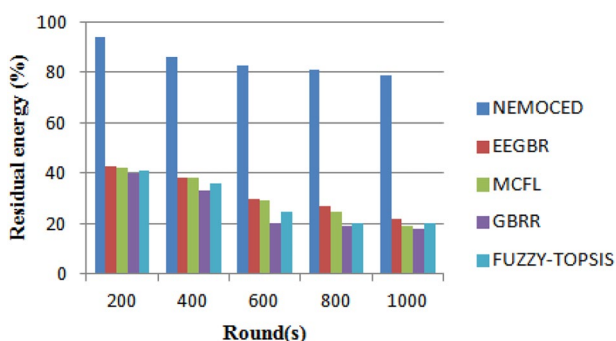


Fig. 8 The residual energy in the NEMOCED and other methods based on the round's numbers

Table 3 Confusion matrix

Cluster name	Cluster A	Cluster B	...	Cluster P	Σ
Cluster A	a_1	b_1	...	p_1	$a_1 + b_1 + \dots + p_1$
Cluster B	a_2	b_2	...	p_2	$a_2 + b_2 + \dots + p_2$
...
Cluster P	a_k	b_k	...	p_k	$a_k + b_k + \dots + p_k$
Σ	$\sum_{i=1}^n a_i$	$\sum_{i=1}^n b_i$...	$\sum_{i=1}^n p_i$	

NEMOCED protocol is better. It is due to the presence of successor CHs.

5.4 The NEMOCED Evaluation

We evaluate the proposed method performance by using the confusion matrix. In general, if the number of clusters is k , the confusion matrix will be $k * k$. The confusion matrix for n clusters has been defined in Table 3.

In the confusion matrix, the principal diagonal elements represent the number of nodes that are properly clustered. For example, a_1 and b_2 are the number of nodes the actual clusters of them are A and B respectively and are correctly located in these clusters. The elements on the secondary diagonal elements are the number of nodes that are not properly clustered. For example, b_1 is the number of nodes that its actual cluster is A, but they have been mistakenly located in cluster B.

According to the confusion matrix, the sum of the quantities in the secondary and principal diagonal elements is the number of correct and incorrect cases in the proposed clustering method, respectively. In order to proper clustering method, the matrix elements that are not in the principal diameter should have a value close to zero. We propose five criteria namely accuracy, error rate, sensitivity, specificity, and precision in the proposed method for evaluating the clustering model. All criteria are defined according to the confusion matrix. The accuracy criterion is defined as follows:

$$Accuracy = \frac{\text{sum of the principal diagonal elements of confusion matrix}}{\text{sum of the confusion matrix elements}} = \frac{a_1 + b_2 + \dots + p_k}{\sum_{i=1}^n (a_i + b_i + \dots + p_i)} \tag{58}$$

In fact, the accuracy of the proposed algorithm is equal to the percentage of nodes that are properly clustered. By subtracting this value from 1, the error rate of the model is obtained (Eq. 59).

$$ErrorRate = \frac{\text{sum of the secondary diagonal elements of confusion matrix}}{\text{sum of the confusion matrix elements}} = \frac{p_1 + \dots + a_k}{\sum_{i=1}^n (a_i + b_i + \dots + p_i)} \tag{59}$$

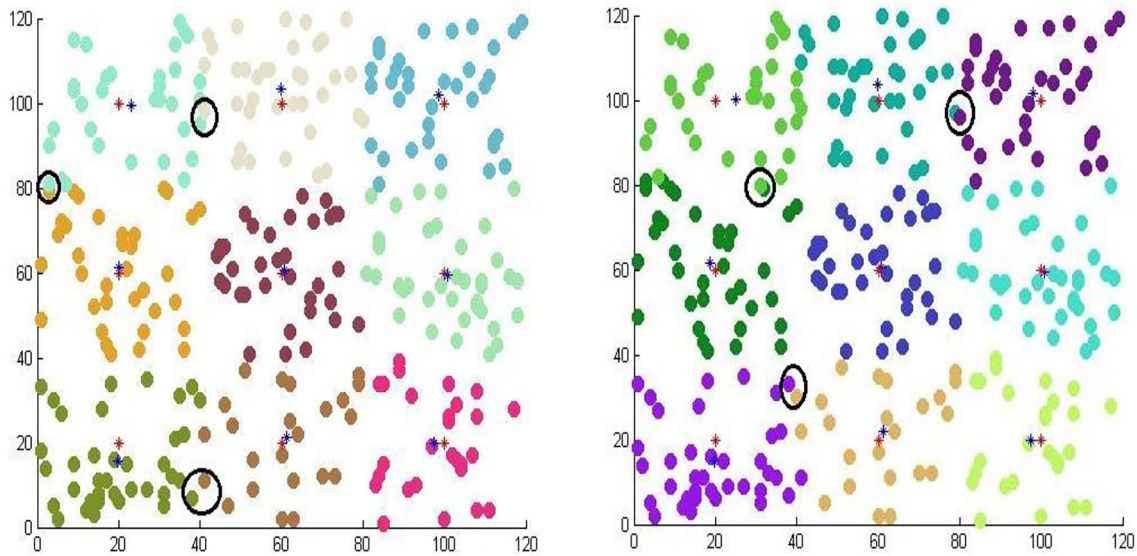


Fig. 9 Clustering in the NEMOCED with 200 nodes. The blank circles represent clustering error. The nodes inside the blank circles are nodes that have been mistakenly located in the other clusters

Table 4 Confusion matrix in Fig. 9

Cluster name	1	2	3	4	5	6	7	8	9	Σ
1	35	1	0	0	0	0	0	0	0	36
2	1	35	0	0	0	0	0	0	0	36
3	0	1	31	1	0	0	0	0	0	33
4	1	0	0	23	0	0	0	0	0	24
5	0	0	0	0	36	0	0	0	0	36
6	0	0	1	0	0	35	0	0	1	37
7	0	0	0	0	0	0	27	0	0	27
8	0	0	0	0	0	0	0	36	0	36
9	0	0	0	0	0	1	0	0	34	35
Σ	37	37	32	24	36	36	27	36	35	

Other criteria are sensitivity and specificity. A proper trade-off between these two criteria can be helpful. These two criteria are defined as follows:

$$sensitivity = \frac{a_1}{a_1 + b_1 + \dots + p_1} \tag{60}$$

$$specificity = \frac{p_k}{a_k + b_k + \dots + p_k} \tag{61}$$

The last criterion is precision. This criterion is defined as follows:

$$precision = \frac{a_1}{\sum_{i=1}^n a_i} \tag{62}$$

We have used the five criteria above because any of the above criteria alone cannot guarantee the validity of the

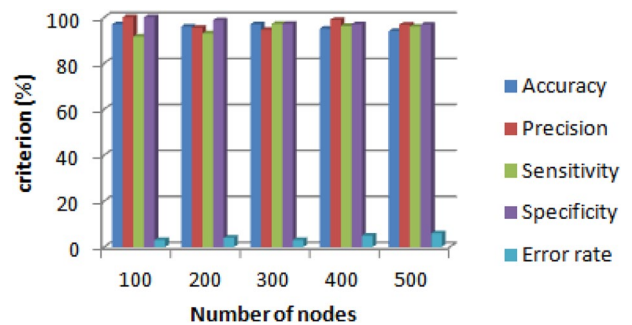


Fig. 10 Five proposed criteria in the NEMOCED protocol with different nodes

proposed algorithm. Therefore, the combination of the above criteria ensures the quality of the NEMOCED protocol. Figure 9 indicates the proposed clustering algorithm.

As shown in Fig. 9, some of the nodes are not properly clustered. Table 4 shows the confusion matrix in the proposed method for Fig. 9.

According to Fig. 9 and Table 4, the five proposed criteria are calculated as follows:

$$accuracy = \frac{292}{300} = 0.97, \quad errorrate = 1 - 0.97 = 0.03$$

$$sensitivity = \frac{35}{36} = 0.972, \quad specificity = \frac{34}{35} = 0.9714$$

$$precision = \frac{35}{37} = 0.9459$$

As can be seen, the proposed clustering method has high accuracy and low error rate. Figure 10 illustrates the five criteria with different numbers of nodes.

6 Conclusion and Future Work

This paper presents a density-based clustering algorithm and a new energy model in wireless sensor networks. The new energy model determines the number of proper clusters based on the network size and the number of nodes. The NEMOCED protocol provides a new tree structure by which two CHs are identified in each cluster. The first type is the main CH, which is responsible for collecting cluster data and sending them to the BS. The second type is the successor CH that is a good successor to the main CH at the time of its energy termination. The CH selection is performed by using residual energy, node distance to the BS, density, δ , node distance to the center of gravity, and $\|\delta_i - \rho_i\|$ criterion. One of the strengths of the NEMOCED is that the choice of the main CH is based on the density of nodes. The clustering process is performed in two steps. In the first phase, environment density is evaluated using grids. In the second phase, clustering is performed based on local density. To validate the performance of the NEMOCED, we made a comprehensive comparison of the NEMOCED with some similar algorithms. The simulation considered different sensor and base station deployments and the various data aggregation ratios. The result shows that the NEMOCED can increase the network lifetime through the proper selection cluster head. To prove the efficiency of our proposed algorithm, we also considered several validation criteria to increase clustering accuracy. An extensive simulation has been performed to test the performance of the NEMOCED under different options. All the conclusions demonstrate that our proposed protocol is suitable and efficient for large-scale WSNs.

In the future, in addition to wireless sensor networks, the proposed protocol can also be used in classification and data

mining. In future work, we will reduce base table rules (BS rules) using fuzzy logic to improve the efficiency of the proposed algorithm.

Appendix 1: An Example for Calculating the Local Density of a Cluster According to the NEMOCED Algorithm

Assume that five nodes in a cell are in the coordinates $a=(10,20)$, $b=(5,7)$, $c=(9,10)$, $d=(15,20)$, and $e=(30,35)$. The adjacency matrix d_{ij} , which indicates the Euclidean distance of nodes with each other is defined as follows:

$$d_{ij} = \begin{matrix} & \begin{matrix} a & b & c & d & e \end{matrix} \\ \begin{matrix} a \\ b \\ c \\ d \\ e \end{matrix} & \begin{pmatrix} 0 & 13.9 & 10 & 5 & 25 \\ 13.9 & 0 & 5 & 16.4 & 37.5 \\ 10 & 5 & 0 & 11.6 & 32.6 \\ 5 & 16.4 & 11.6 & 0 & 21.2 \\ 25 & 37.5 & 32.6 & 21.2 & 0 \end{pmatrix} \end{matrix}$$

According to the adjacency matrix, the local density for each node is:

$$\rho_a = \sum_{j=1}^5 e\left(-\frac{d_{ij}^2}{d_c^2}\right) = e\left(-\frac{0^2}{25^2}\right) + e\left(-\frac{13.9^2}{25^2}\right) + e\left(-\frac{10^2}{25^2}\right) + e\left(-\frac{5^2}{25^2}\right) + e\left(-\frac{25^2}{25^2}\right) = 3.9201$$

$$\rho_b = \sum_{j=1}^5 e\left(-\frac{d_{ij}^2}{d_c^2}\right) = 4.0523 \rho_c = \sum_{j=1}^5 e\left(-\frac{d_{ij}^2}{d_c^2}\right) = 4.1399$$

$$\rho_d = \sum_{j=1}^5 e\left(-\frac{d_{ij}^2}{d_c^2}\right) = 3.6113 \rho_e = \sum_{j=1}^5 e\left(-\frac{d_{ij}^2}{d_c^2}\right) = 3.2136$$

Since ρ_a is smaller than ρ_b and ρ_c , so the maximum value in the row of ρ_b and ρ_c is considered as δ_a (or $\delta_a = 37.5$). Since ρ_c is not smaller than any value, so the minimum value in the corresponding row is considered as δ_c (or $\delta_c = 5$). Other values are:

$$\delta_b = 32.6, \delta_d = 37.5, \delta_e = 37.5$$

The maximum value of $\|\delta_i - \rho_i\|$ is related to the nodes ρ_d and δ_d . Therefore, node d with coordinates $(15, 20)$ is chosen as the center of the cluster. As can be seen, the coordinates of node d are closer to the center of gravity than the other nodes. Therefore, the proposed method has high accuracy in the selection of the cluster centers. After selecting the cluster centers and identifying them, nodes in other candidate grids are allocated to the corresponding cluster. This work is performed based on the proximity of the geographical distance.

Appendix 2

Example 1 Assume that five nodes are in a cluster with coordinates $1=(5,6)$, $2=(10,15)$, $3=(9,11)$, $4=(20,1)$ and $5=(14,8)$. The nodes have the residual energy of $0.8 J$, $0.4 J$, $0.9 J$, $0.7 J$ and $0.6 J$, respectively. Also, assume that the BS is located at $(100,100)$. As before, with the calculation of d_{ij} , the parameters ρ , δ and $\|\delta_i - \rho_i\|$ are defined as follows:

$$\rho_1 = 3.5898 \quad \delta_1 = 17.2047 \quad \|\delta_1 - \rho_1\| = 16.6139$$

$$\rho_2 = 3.8196 \quad \delta_2 = 14.8661 \quad \|\delta_2 - \rho_2\| = 11.0465$$

$$\rho_3 = 3.9868 \quad \delta_3 = 4.12310 \quad \|\delta_3 - \rho_3\| = 0.13630$$

$$\rho_4 = 3.0308 \quad \delta_4 = 17.2047 \quad \|\delta_4 - \rho_4\| = 14.1739$$

$$\rho_5 = 2.8340 \quad \delta_5 = 17.2047 \quad \|\delta_5 - \rho_5\| = 14.3707$$

Given the coordinates of the nodes, the center of gravity is $(11.6, 8.2)$. Tables 5 and 6 are the decision and normalization tables, respectively.

According to relations 52–57:

$$k = \frac{1}{\log_2 5} = 0.4307$$

$$Entropy(D)_{E_{RS}} = 0.9789$$

Similarly, entropy is calculated for other parameters. The last row in normalization and decision tables represents the

Table 5 Decision table in Example 1

ID	Residual energy (E_{RS})	Distance to BS (d_{BS})	Density (ρ)	δ	Distance to gravity center (d_g)	$\ \delta_i - \rho_i\ $	Candidate CH
1	0.8	133.6451	3.5898	17.2047	6.9570	13.6149	–
2	123.7942	0.4	14.8661	3.8196	11.0465	6.9857	–
3	127.2871	0.9	4.1231	3.9868	0.1363	3.8210	–
4	127.2831	0.7	17.2047	3.0308	14.1739	11.0635	–
5	125.9365	0.6	17.2047	2.8340	14.3707	2.4083	–
$\sum i$	637.946	3.4	70.6033	17.2610	53.3423	31.2355	–
Avg	127.5892	0.68	14.1207	3.4522	10.6685	6.2471	–

Table 6 Normalization table in Example 1

ID	Residual energy (E_{RS}) -Joule	Distance to BS (d_{BS})	Density (ρ)	δ	Distance to gravity center (d_g)	$\ \delta_i - \rho_i\ $	Candidate CH
1	0.2353	0.2095	0.2080	0.2437	0.2227	0.2552	–
2	0.1176	0.1941	0.2213	0.2106	0.2236	0.2071	–
3	0.2647	0.1995	0.2310	0.0584	0.1223	0.0026	–
4	0.2059	0.1995	0.1756	0.2437	0.3542	0.2657	–
5	0.1765	0.1974	0.1642	0.2437	0.0771	0.2694	–
Entropy	0.9789	0.9997	0.9947	0.9482	0.9267	0.8671	–

Table 7 Labeling of the decision table

ID	Residual energy (E_{RS}) -Joule	Distance to BS (d_{BS})	Density (ρ)	δ	Distance to gravity center (d_g)	$\ \delta_i - \rho_i\ $	Candidate CH
1	HIGH	FAR	HIGH	BAD	FAR	HIGH	YES
2	LOW	NEAR	HIGH	BAD	FAR	HIGH	NO
3	HIGH	AVERAGE	HIGH	GOOD	MEDIUM	LOW	NO
4	HIGH	AVERAGE	MEDIUM	BAD	FAR	HIGH	NO
5	MEDIUM	AVERAGE	LOW	BAD	NEAR	HIGH	YES

average (Avg) and entropy. Also, *min* is the minimum value of each column. For example, the *min* value for the residual energy is 0.4. So, the residual energy in the decision table is labeled as follows:

$$E_{Rs} = \begin{cases} \text{LOW} & E_{Rs} \leq 0.4 \\ \text{MEDIUM} & 0.4 < E_{Rs} \leq 0.68 \\ \text{HIGH} & E_{Rs} > 0.68 \end{cases}$$

Other parameters are also labeled with the following functions. Table 7 illustrates the labeling of the decision table.

$$d_{BS} = \begin{cases} \text{NEAR} & d_{BS} \leq 123.7942 \\ \text{AVERAGE} & 123.7942 < d_{BS} \leq 127.5892 \\ \text{FAR} & d_{BS} > 127.5892 \end{cases}$$

$$\rho = \begin{cases} \text{LOW} & \rho \leq 2.8340 \\ \text{MEDIUM} & 2.8340 < \rho \leq 3.4522 \\ \text{HIGH} & \rho > 3.4522 \end{cases}$$

$$\delta = \begin{cases} \text{GOOD} & \delta \leq 4.1231 \\ \text{FAIRLYGOOD} & 4.1231 < \delta \leq 14.1207 \\ \text{BAD} & \delta > 14.1207 \end{cases}$$

$$d_g = \begin{cases} \text{NEAR} & d_g \leq 2.4083 \\ \text{AVERAGE} & 2.4083 < d_g \leq 6.2471 \\ \text{FAR} & d_g > 6.2471 \end{cases}$$

$$\|\delta_i - \rho_i\| = \begin{cases} \text{LOW} & \|\delta_i - \rho_i\| \leq 0.1363 \\ \text{MEDIUM} & 0.1363 < \|\delta_i - \rho_i\| \leq 10.6685 \\ \text{HIGH} & \|\delta_i - \rho_i\| > 10.6685 \end{cases}$$

As previously stated, the estimation of the candidate CHs is performed by BS and the base table. From the five nodes, three nodes have a "NO" label. Also, two nodes have a "YES" label. So, the total entropy is:

$$Entropy = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.9710$$

To calculate the information gain, the entropy corresponding to all the values in the decision table must be calculated. So:

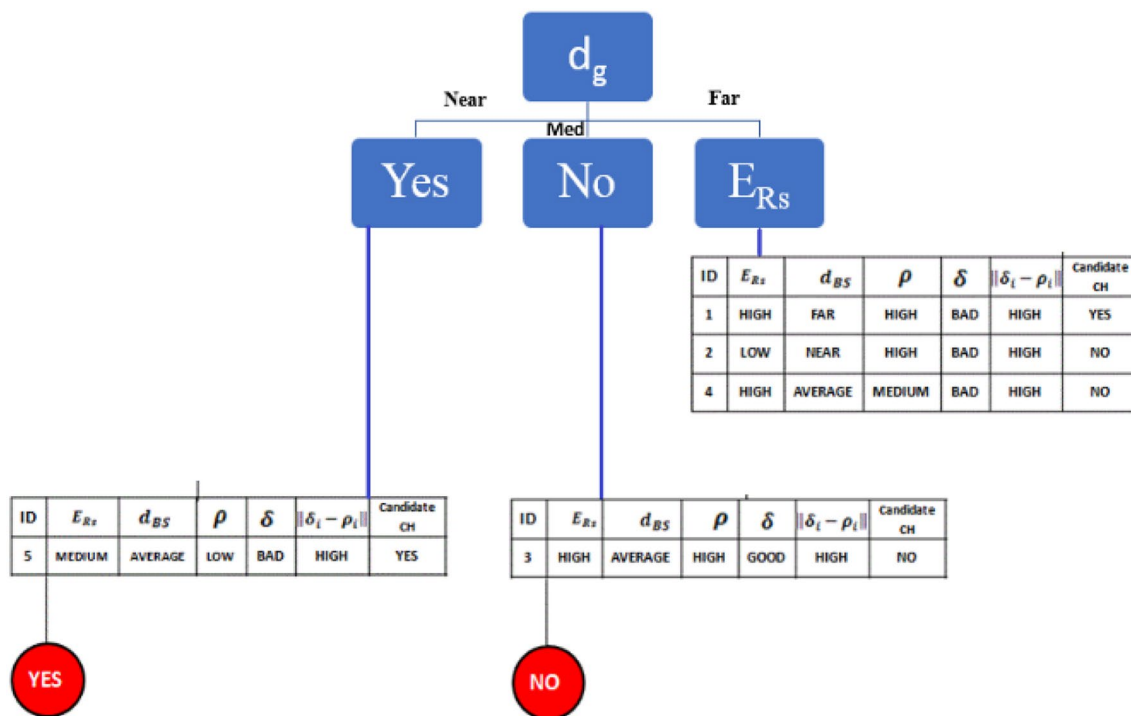


Fig. 11 Tree structure with three branches in example 1

$$Entropy(E_{RS}) = Entropy(LOW) + Entropy(MEDIUM) + Entropy(HIGH) = 0.5510$$

$$Entropy(d_{BS}) = Entropy(NEAR) + Entropy(AVERAGE) + Entropy(FAR) = 0.5510$$

$$Entropy(\rho) = Entropy(LOW) + Entropy(MEDIUM) + Entropy(HIGH) = 0.5510$$

$$Entropy(\delta) = Entropy(BAD) + Entropy(GOOD) = 0.8$$

$$Entropy(d_g) = Entropy(NEAR) + Entropy(Medium) + Entropy(FAR) = 0.5510$$

$$Entropy(\|\delta_i - \rho_i\|) = Entropy(LOW) + Entropy(HIGH) = 0.8$$

On one hand, the difference of entropy and information gain in the residual energy, distance to the BS, density, and distance to the center of gravity is the greatest and on the other hand, one case must be selected. Therefore, by referring to the normalization table, d_g is selected as the root of the tree. Since it has three labels, so the tree has three branches (see Fig. 11).

By repeating the calculations for the right table in Fig. 11, the tree structure is converted into Figs. 12 and 13.

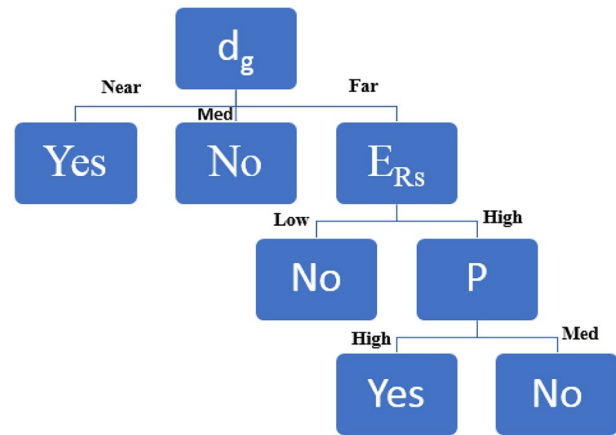


Fig. 13 Tree structure with three branches in example 1 and based on Fig. 12

Therefore, nodes 5 and 1 are selected as the original and successor CHs, respectively. As is clear, these two nodes are better than other nodes. In the process of selecting the cluster head and forming the tree, we may reach a phase that has not any improvement. This mode is when all the parameters of the labeling have the same characteristics and entropy. This mode is called “TRAP”. One of the basic conditions for stopping work is the Trap state. Example 3 illustrates this case.

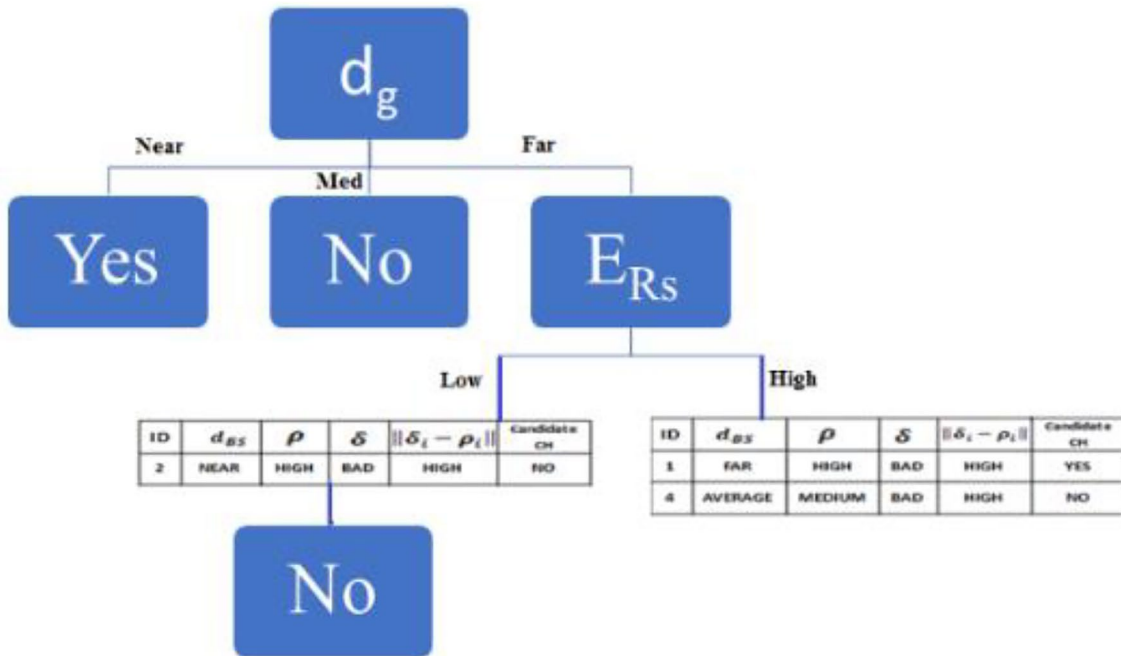


Fig. 12 Tree structure with three branches in example2 and based on Fig. 13

Table 8 Labeling of the decision table in Example 2

ID	Residual energy (E_{Rs})-Joule	Distance to BS (d_{BS})	ρ Density ()	δ	Distance to gravity center (d_g)	$\ \delta_i - \rho_i\ $	Candidate CH
1	HIGH	FAR	LOW	BAD	MEDIUM	HIGH	NO
2	HIGH	AAVERAGE	HIGH	FGOOD	FAR	MEDIUM	NO
3	HIGH	FAR	HIGH	BAD	MEDIUM	HIGH	YES
4	HIGH	FAR	HIGH	BAD	FAR	HIGH	YES
5	HIGH	NEAR	MEDIUM	BAD	MEDIUM	HIGH	YES
6	MEDIUM	FAR	HIGH	BAD	FAR	HIGH	NO
7	LOW	AVERAGE	HIGH	GOOD	NEAR	LOW	NO

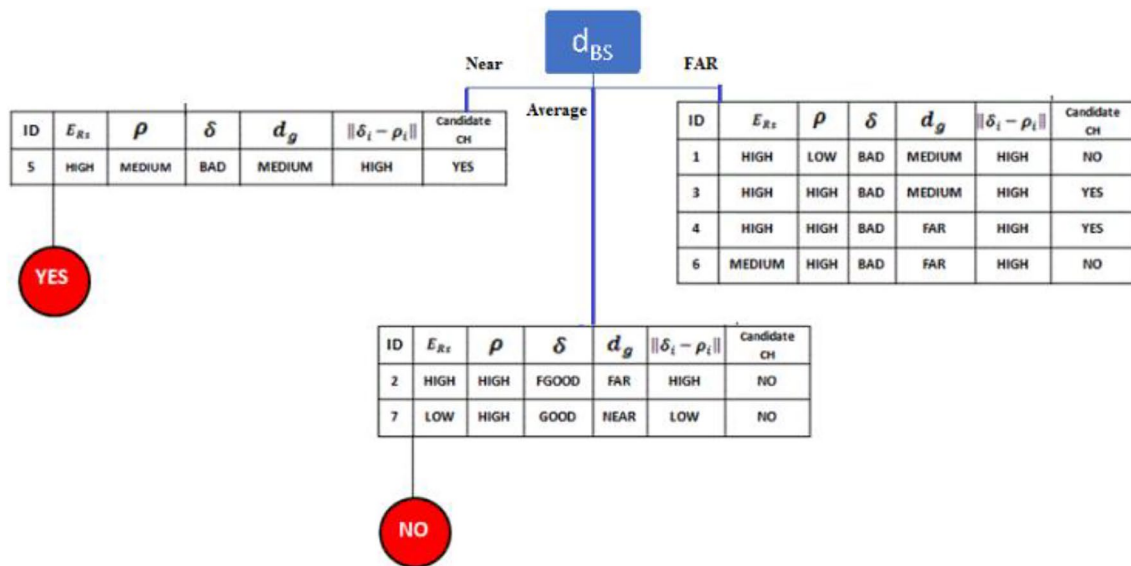


Fig. 14 Tree structure with three branches in example 2

Fig. 15 Tree structure with three branches in example 2 based on Fig. 14

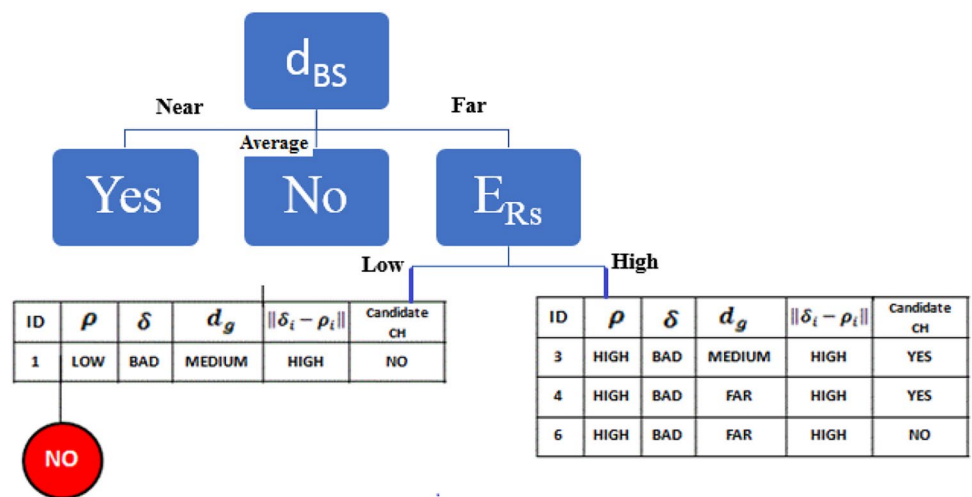
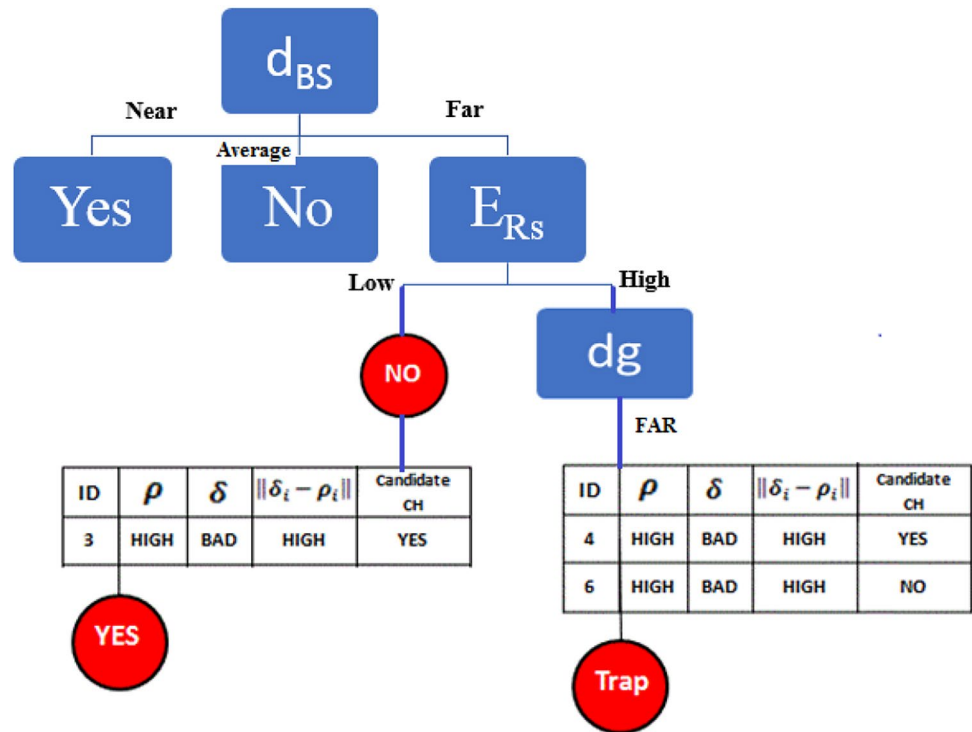


Fig. 16 Tree structure with three branches in example 2 based on Fig. 15



Example 2 Assume that five nodes with coordinates $1=(8,7)$, $2=(20,10)$, $3=(4,15)$, $4=(15,1)$, $5=(16,15)$, $6=(3,20)$ and $7=(13,17)$ with the residual energy of 0.8 J, 0.6 J, 0.5 J, 0.9 J, 0.7 J, 0.2 J and 0.3 J are in a cluster, respectively. Also, assume the BS is located at (100,100). As in Example 1, after forming the decision and normalization tables, the labeling table will be in the form of Table 8.

Other parameters are defined as follows:

$$k = \frac{1}{\log_2 5} = 0.4307, \text{Entropy}(D) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{4}{5} \log_2 \frac{4}{5} = 0.9852$$

$$\text{Entropy}(E_{RS}) = 0.6935 \quad \text{Entropy}(d_{BS}) = 0.5714$$

$$\text{Entropy}(\rho) = 0.6935$$

$$\text{Entropy}(\delta) = 0.6935 \quad \text{Entropy}(d_g) = 0.7871$$

$$\text{Entropy}(\|\delta_i - \rho_i\|) = 0.6935$$

Figures 14, 15 and 16 illustrate the tree structure in example 2.

As shown in Fig. 16, the TRAP mode has occurred. So, the algorithm does not continue. Therefore, nodes 5 and 3 are selected as main and successor CHs, respectively. Note that there may be more than one successor CH. We assume that the first proper node in the tree structure is selected as

the main CH. The rest of the nodes are considered as the successor CH.

This mode happens in high-density clusters. In other words, if the cluster density is higher, then the number of the successor CHs will also be higher. Perhaps at first glance, the reader of this article believes that with the initial look at the labeled table or decision table, it would be possible to find the main and successor CHs. It should be noted that with the initial look at the labeled table in example 2, node 4 should be selected as the original CH because it has the most residual energy. This is a false choice because node 4 just has better residual energy and is worse than nodes 1 or 5 in other parameters such as density or distance to the center of gravity. This is one of the important differences between our proposed method with other proposed methods. Most of the proposed methods choose the CH only based on the residual energy.

Funding Not applicable.

Data Availability The present study is based on synthesized data generated randomly by the authors based on some parameters mentioned in the text.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

- Ahmadi, H., F. Viani, and R. Bouallegue, *An accurate prediction method for moving target localization and tracking in wireless sensor networks*. Ad Hoc Networks, 2018. **70**: p. 14–22.
- Jan, M.A., et al., *A Sybil attack detection scheme for a forest wildfire monitoring application*. Future Generation Computer Systems, 2018. **80**: p. 613–626.
- Oracevic, A., S. Akbas, and S. Ozdemir, *Secure and reliable object tracking in wireless sensor networks*. Computers & Security, 2017. **70**: p. 307–318.
- Ozdemir, S. and H. Çam, *Integration of false data detection with data aggregation and confidential transmission in wireless sensor networks*. IEEE/ACM Transactions on Networking (TON), 2010. **18**(3): p. 736–749.
- Wang, X., et al., *Prediction-based dynamic energy management in wireless sensor networks*. Sensors, 2007. **7**(3): p. 251–266.
- Awad, F.H., *Optimization of Relay Node Deployment for Multi-Source Multi-Path Routing in Wireless Multimedia Sensor Networks Using Gaussian Distribution*. Computer Networks, 2018.
- Sun, E., et al., *Adaptive Deployment Scheme and Multi-path Routing Protocol for WMSNs*. Indonesian Journal of Electrical Engineering and Computer Science, 2014. **12**(2): p. 1454–1461.
- Anand, V., et al., *An energy efficient approach to extend network life time of wireless sensor networks*. Procedia Computer Science, 2016. **92**: p. 425–430.
- Asha, G., *Energy efficient clustering and routing in a wireless sensor networks*. Procedia computer science, 2018. **134**: p. 178–185.
- Barekattain, B., S. Dehghani, and M. Pourzaferani, *An energy-aware routing protocol for wireless sensor networks based on new combination of genetic algorithm & k-means*. Procedia Computer Science, 2015. **72**: p. 552–560.
- Elshrkawey, M., S.M. Elsherif, and M.E. Wahed, *An enhancement approach for reducing the energy consumption in wireless sensor networks*. Journal of King Saud University-Computer and Information Sciences, 2018. **30**(2): p. 259–267.
- Dwivedi, A.K., et al., *EETSP: Energy-efficient two-stage routing protocol for wireless sensor network-assisted Internet of Things*. International Journal of Communication Systems, 2021. **34**(17): p. e4965.
- Ramtin, A., V. Hakami, and M. Dehghan, *A Perturbation-Proof Self-stabilizing Algorithm for Constructing Virtual Backbones in Wireless Ad-Hoc Networks*. in *International Symposium on Computer Networks and Distributed Systems*. 2013. Springer.
- Hoang Kha, H., *Optimal precoders and power splitting factors in multiuser multiple-input multiple-output cognitive decode-and-forward relay systems with wireless energy harvesting*. International Journal of Communication Systems, 2021: p. e5047.
- Fanian, F. and M.K. Rafsanjani, *Memetic fuzzy clustering protocol for wireless sensor networks: Shuffled frog leaping algorithm*. Applied Soft Computing, 2018. **71**: p. 568–590.
- Liao, Y., H. Qi, and W. Li, *Load-balanced clustering algorithm with distributed self-organization for wireless sensor networks*. IEEE sensors journal, 2013. **13**(5): p. 1498–1506.
- Wang, T., et al., *Genetic Algorithm for Energy-Efficient Clustering and Routing in Wireless Sensor Networks*. Journal of Systems and Software, 2018.
- Biswas, S. and R. Morris, *ExOR: opportunistic multi-hop routing for wireless networks*. ACM SIGCOMM Computer Communication Review, 2005. **35**(4): p. 133–144.
- Fabbri, F., C. Buratti, and R. Verdone, *A multi-sink multi-hop wireless sensor network over a square region: Connectivity and energy consumption issues*. in *GLOBECOM Workshops, 2008 IEEE*. 2008. IEEE.
- Macit, M., V.C. Gungor, and G. Tuna, *Comparison of QoS-aware single-path vs. multi-path routing protocols for image transmission in wireless multimedia sensor networks*. Ad hoc networks, 2014. **19**: p. 132–141.
- Patil, M. and R.C. Biradar, *A survey on routing protocols in wireless sensor networks*. in *Networks (ICON), 2012 18th IEEE International Conference on*. 2012. IEEE.
- Sahin, D., et al., *Quality-of-service differentiation in single-path and multi-path routing for wireless sensor network-based smart grid applications*. Ad Hoc Networks, 2014. **22**: p. 43–60.
- Singh, O., V. Rishiwal, and M. Yadav, *Multi-objective lion optimization for energy-efficient multi-path routing protocol for wireless sensor networks*. International Journal of Communication Systems, 2021. **34**(17): p. e4969.
- Akhtar, A., A.A. Minhas, and S. Jabbar, *Energy aware intra cluster routing for wireless sensor networks*. International Journal of Hybrid Information Technology, 2010. **3**(1).
- Gong, B., et al. *Multihop routing protocol with unequal clustering for wireless sensor networks*. in *Computing, Communication, Control, and Management, 2008. CCCM'08. ISECS International Colloquium on*. 2008. IEEE.
- Min, X., et al., *Energy efficient clustering algorithm for maximizing lifetime of wireless sensor networks*. AEU-International Journal of Electronics and Communications, 2010. **64**(4): p. 289–298.
- Abdullah, M., et al., *Density grid-based clustering for wireless sensors networks*. Procedia Computer Science, 2015. **65**: p. 35–47.
- Bhakare, K.R., R. Krishna, and S. Bhakare, *An energy-efficient grid based clustering topology for a wireless sensor network*. International Journal of Computer Applications, 2012. **39**(14): p. 24–28.
- Jannu, S. and P.K. Jana, *Energy efficient grid based clustering and routing algorithms for wireless sensor networks*. in *Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on*. 2014. IEEE.
- Meng, X., et al., *A grid-based reliable routing protocol for wireless sensor networks with randomly distributed clusters*. Ad Hoc Networks, 2016. **51**: p. 47–61.
- Chang, D.-X., X.-D. Zhang, and C.-W. Zheng, *A genetic algorithm with gene rearrangement for K-means clustering*. Pattern Recognition, 2009. **42**(7): p. 1210–1222.
- Hoang, D., R. Kumar, and S. Panda, *Fuzzy C-means clustering protocol for wireless sensor networks*. in *Industrial Electronics (ISIE), 2010 IEEE International Symposium on*. 2010. IEEE.
- Heinzelman, W.R., A. Chandrakasan, and H. Balakrishnan, *Energy-efficient communication protocol for wireless microsensor networks*. in *System sciences, 2000. Proceedings of the 33rd annual Hawaii international conference on*. 2000. IEEE.
- Bara'a, A.A. and E.A. Khalil, *A new evolutionary based routing protocol for clustered heterogeneous wireless sensor networks*. Applied Soft Computing, 2012. **12**(7): p. 1950–1957.
- Su, S. and S. Zhao, *An optimal clustering mechanism based on Fuzzy-C means for wireless sensor networks*. Sustainable Computing: Informatics and Systems, 2018. **18**: p. 127–134.

36. Mirzaie, M. and S.M. Mazinani, *MCFL: An energy efficient multi-clustering algorithm using fuzzy logic in wireless sensor network*. *Wireless Networks*, 2018. **24**(6): p. 2251-2266.
37. Darabkh, K.A., et al., *C-DTB-CHR: centralized density-and threshold-based cluster head replacement protocols for wireless sensor networks*. *The Journal of Supercomputing*, 2017. **73**(12): p. 5332-5353.
38. Mirzaee, F., M. Keshavarz, and M. Vazifedoust, *Development of a new algorithm (SM-SEBAL) to evaluate Evapotranspiration based on remote sensing data*. 2018.
39. Sert, S.A., H. Bagci, and A. Yazici, *MOFCA: Multi-objective fuzzy clustering algorithm for wireless sensor networks*. *Applied Soft Computing*, 2015. **30**: p. 151-165.
40. Huamei, Q., et al., *An energy-efficient non-uniform clustering routing protocol based on improved shuffled frog leaping algorithm for wireless sensor networks*. *IET Communications*, 2021. **15**(3): p. 374-383.
41. El Khediri, S., et al., *Improved node localization using K-means clustering for Wireless Sensor Networks*. *Computer Science Review*, 2020. **37**: p. 100284
42. Tunca, C., et al. *Ring routing: An energy-efficient routing protocol for wireless sensor networks with a mobile sink*. in *Signal Processing and Communications Applications Conference (SIU), 2012 20th*. 2012. IEEE.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



sensing.

Amir Abbas Baradaran received the B.Sc. (2007), M.Sc. (2011), and Ph.D. (2019) degrees in computer engineering from Shahid Beheshti University (SBU), Tehran, Iran. He is currently a postdoctoral researcher at Sharif University of Technology, Tehran, Iran. He has completed several research projects at the National Elite Foundation (IRAN) and has been received the scholarship and talent award. His research interests include Wireless sensor networks, Fuzzy logic and systems, and Remote



Fahimeh Rabieefar received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Kashan, Kashan, Iran, in 2015 and 2018, respectively. Her fields of interest are Graphene Nano-ribbon Transistors, Fuzzy systems, Sensor networks, and Digital circuits design.