Wang *et al. EURASIP Journal on Wireless Communications and Networking* (2018) 2018:54
https://doi.org/10.1186/s13638-018-1067-8

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Spectral partitioning and fuzzy C-means based clustering algorithm for big data wireless sensor networks

Quyuan Wang[1], Songtao Guo[1*], Jianji Hu[1] and Yuanyuan Yang[1,2]

Abstract

In wireless sensor networks, sensor nodes are usually powered by battery and thus have very limited energy. Saving energy is an important goal in designing a WSN. It is known that clustering is an effective method to prolong network lifetime. Due to the development of big data, there are more sensor nodes and data needed to process. So how to cluster sensor nodes cooperatively and achieve an optimal number of clusters in a big data WSN is an open issue. In this paper, we first propose an analytical model to give the optimal number of clusters in a wireless sensor network. We then propose a centralized cluster algorithm based on spectral partitioning method. After that, we present a distributed implementation of the clustering algorithm based on fuzzy C-means method. Finally, we conduct extensive simulations, and the results show that the proposed algorithms outperform the hybrid energy-efficient distributed (HEED) clustering algorithm in terms of energy cost and network lifetime.

Keywords: Clustering, Spectral partitioning, Fuzzy C-means, Cooperative nodes, Big data wireless sensor networks

1 Introduction

In recent years, wireless sensor networks (WSNs) have been used as an important information gathering paradigm in a wide range of applications, such as environmental monitoring, target tracking, battlefield surveillance, home security, and health monitoring [1]. A wireless sensor network (WSN) is composed of hundreds or even thousands of sensor nodes which use wireless communication to perform distributed sensing tasks. Meanwhile, in many applications, the amount of sensory data manifests an explosive growth with the development of wireless sensor networks. The data generated by an individual sensor may not appear to be significant, but numerous sensors in WSNs can produce a significant portion of the big data in the above military and civilian applications. Due to the limited battery capacity and low-cost requirement, sensor nodes are usually equipped with low-end computational module and radio transceiver [2]. As it is infeasible to replace batteries once WSNs are deployed in a harsh environment, an important design principle in WSNs is to minimize the energy consumption in sensing, computing, and communication.

Big data analysis provides more convenience for our data gathering and processing; however, the number of sensor nodes and the size of gathering data grow rapidly. There are several literatures that focus on big sensory data (BSD) in WSNs [3–7]. But different with these works, this paper concentrates on how to achieve effective data clustering under the background of big data. It has been shown that clustering is an effective scheme in increasing network lifetime and scalability of WSNs. Sensor nodes are partitioned into clusters, and each cluster consists of a cluster head (CH) and a number of cluster members. Cluster members collect data from the environment and send the data to their CH. A CH is responsible for gathering data from its members and relaying the data to sink node. This method can save the energy of sensor nodes since sensor nodes do not need to upload data to the sink node directly. Based on whether sensor nodes can communicate directly with their CH, the clustered WSNs can be classified as single-hop WSNs and multi-hop WSNs.

*Correspondence: songtao_guo@163.com
[1]National & Local Joint Engineering Laboratory of Intelligent Transmission and Control Technology (Chongqing), College of Electronic and Information Engineering, Southwest University, 2 Tiansheng Road, Beibei District, 400715 Chongqing, China
Full list of author information is available at the end of the article

In a multi-hop WSN, the data from some sensor nodes need to be relayed via multiple hops to reach the CH. The clustered WSNs can also be categorized into homogeneous WSNs and heterogeneous WSNs. All sensor nodes are of the same specification in homogeneous WSNs, while in heterogeneous WSNs, a small number of powerful nodes are deployed as cluster heads and the rest of regular nodes act as cluster members. Such a two-layer hybrid network can improve the network lifetime and stability with a marginal increase in the cost of network deployment. Basically, any clustering algorithm involves cluster management which consists of determining the suitable number of clusters, selecting the cluster head for each cluster, and transmitting data within clusters and from cluster heads to the sink node [8].

For a sensor node, the dominant energy consumer is the radio unit. When the network is partitioned into some clusters, data transmission can be classified into two stages: intra-cluster transmission and inter-cluster transmission. One of disadvantages of existing clustering algorithms [9–19] is that thet do not give the energy consumption model and do not study what the optimal number of clusters should be in a network. Moreover, most of these existing algorithms do not fit well into the big data environment. Thus, in this paper, we introduce the energy consumption model to determine the optimal number of clusters. And we use spectral partitioning method and Fuzzy C-means method to partition the network into a fixed optimal number of clusters respectively. Compared with previous works, our contributions in this paper can be summarized as follows:

- We propose an analytical energy consumption model to determine the optimal number of clusters in a big data WSN. In this model, the communication between two clusters use cooperative transmission. By using cooperative nodes, we can reduce the energy consumption of cluster heads efficiently and prolong the lifetime of network.
- We propose a centralized clustering algorithm based on spectral partitioning and a distributed implementation of the clustering method based on fuzzy C-means in cluster head selection.
- We verify the performance of the proposed algorithm in terms of network lifetime and node remaining energy. In particular, our algorithm can decrease the interval between the time of the first node dying and the time of the last node dying, which implies our algorithm can efficiently balance the energy consumption among sensor nodes.

The rest of this paper is organized as follows. Section 2 describes the related work. Section 3 briefly introduces the model to predict energy consumption and the characteristics of Laplacian matrices, which will be used in the analytical model. Section 4.1 determines the optimal number of clusters. Section 4.2 presents a clustering approach based on spectral classification and the distributed implementation. Section 4.3 presents an algorithm for choosing the cooperative nodes and the cluster heads. The simulation and performance evaluations are presented in Section 5. Finally, Section 6 concludes this paper.

2 Related work

In order to utilize the energy of sensor nodes efficiently, several clustering protocols have been proposed for WSNs. One of the most concerned protocols is low-energy adaptive clustering hierarchy (LEACH) [20], an advantage of which is that it is able to balance the energy consumption of sensor nodes by randomly rotating cluster heads. However, cluster heads locating on the edge of a cluster may waste a lot of energy due to the very close distance between cluster heads. The hybrid energy-efficient distributed (HEED) [21] is another well-known clustering algorithm. In this algorithm, residual energy and node proximity to its neighbors or node degree are used to select cluster head. Compared to LEACH, HEED can evenly distribute the cluster heads in the sensing area by local competition.

However, HEED may result in longer data gathering delay. The centralized LEACH (LEACH-C) algorithm was proposed in [11], in which the base station (BS) selects cluster head and makes sure that any node with low energy does not become a cluster head. However, it is not suitable for large-scale WSNs since it leads to the increase of the delay. The FAR-Zone LEACH protocol (FZ-LEACH) [12] is an improvement to LEACH to eliminate clusters with large scale in sensor networks, by forming Far-Zone that is a group of sensor nodes placed at locations with energies less than a threshold.

In [22–27], fuzzy logic was employed in clustering algorithms to handle the uncertainties in WSNs. Bezdek [22] proposed a fuzzy logic-based clustering and data processing in WSNs. This approach incorporates energy level of each node, bandwidth, and link efficiency. The objective of the proposed work is to improve the performance of network in terms of energy consumption, throughput, time for cluster head selection, number of alive nodes, and network lifetime. In [23], Ahamad proposed an energy-efficient clustering algorithm by using the fuzzy logic system to extend WSN lifetime in probabilistic approach model. This addresses the problem of the bad utilization of residual energy of sensor nodes efficiently with the help of appropriate cluster head selection method.

In [24], cluster formation using fuzzy logic (CFFL) approach has been proposed to prolong network lifetime and reduce energy consumption in WSNs. This approach

uses fuzzy logic in the formation cluster phase, and two fuzzy parameters are used. The two parameters are residual energy which is energy level of each CH and closeness to base station (BS) which is the distance between the CH and the BS. Based on the difference in expected residual and residual energy, a fuzzy logic-based clustering algorithm with an extension to the energy predication has been proposed to prolong the lifetime of WSNs by evenly distributing the workload [25].

The fuzzy C-means (FCM) algorithm was first proposed by Bezdek [28] and has been used in cluster analysis, pattern recognition, and image processing [29–32]. This algorithm is a soft partition technique that assigns a degree of belongingness to a cluster for each sensor node. In this work, FCM algorithm is adopted to form clusters within WSNs. The goal of this algorithm is to address the issue of uneven distribution of sensor nodes related with the application of protocols like LEACH.

3 Network model and spectral classification

In this section, we present the network model and energy model and introduce briefly the Laplacian matrix and spectral classification.

3.1 Network model

In this paper, we consider a WSN where N sensor nodes are uniformly distributed in a $M \times M$ square area. We model the deployment of sensor nodes as a Poisson point process and use λ to represent the density of the underlying Poisson point process. We assume that the sensor network has the following properties and capabilities.

- The network topology keeps unchanged over time, and the base station has unlimited power, computing ability, and locates at the network center.
- Nodes are deployed uniformly, and all the nodes are homogeneous.
- Each node is aware of its own position through RSSI localization.
- All sensor nodes are static, and their battery cannot be recharged.

3.2 Energy model

In this paper, we adopt the same energy consumption model as that in [11], i.e., the energy consumed by transmitting an l-bit message by the radio transmitter is given by

$$E_{Tx}(k,d) = \begin{cases} l * \left(E_{\text{elec}} + \varepsilon_{\text{fs}} d^2\right), d < d_0 \\ l * \left(E_{\text{elec}} + \varepsilon_{\text{amp}} d^4\right), d \geq d_0 \end{cases} \quad (1)$$

However, the differences between our work and the one in [11] are that (i) we consider the relationship between cluster area size (D) and cluster head cover radius (R) in the energy consumption model and (ii) we give the new energy consumption model in the cluster head and the whole cluster, which means that the expression of optimal number of clusters is different.

The energy consumption by receiving an l-bit packet is given by

$$E_{\text{RX}}(l) = l * E_{\text{elec}} \quad (2)$$

In (1) and (2), E_{elec} is the energy dissipated per bit by the transmitter or the receiver circuits. We use $\varepsilon_{fs} d^2$ and $\varepsilon_{\text{amp}} d^4$ to present the amplifier energy consumption per bit in a free space model and a multi-path fading channel model, respectively. d denotes the distance between the transmitter and the receiver. The threshold d_0 can be defined as

$$2d_0 = \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{amp}}} \quad (3)$$

If the distance d is less than d_0, the free space model is used as the energy consumption model of the transmitter; otherwise, the multi-path model is used. As the energy consumption of data transmission is much higher than that of computing, we do not take the energy consumption for computing into consideration. We assume in each round of data collection, every cluster member sends l-bits of data to its cluster head, and then, the energy consumed by a CH in one round of data gathering can be represented by

$$E_{\text{ch}} = \frac{N}{k} * l * E_{\text{elec}} + \frac{N}{k} * l * E_{\text{DA}} + l * \varepsilon_{\text{fs}} * d_{\text{BS}}^4 \quad (4)$$

Cluster head dissipates energy by receiving signals from nodes, aggregating the signals, and transmitting the aggregated signal to the BS. k is the number of clusters, E_{DA} is the energy consumed for a CH processing a bit of data from its cluster members, and d_{BS} is the average distance between a CH and the BS. Then, the energy consumed by each cluster member $E_{\text{non-ch}}$ in one round of data collection is

$$E_{\text{non-ch}} = l * E_{\text{elec}} + l * \varepsilon_{\text{elec}} * d_{\text{ch}}^2 \quad (5)$$

where d_{ch} is the distance from the node to the cluster head.

3.3 Laplacian matrix and Fiedler vector

Spectral clustering algorithms have attracted lots of research attentions recently. They are easy to implement as they can be solved efficiently by standard linear algebra components and outperform the traditional clustering algorithms such as the k-means algorithm. Spectral methods usually involve taking the top eigenvectors of some matrix based on the distance between points (or other properties) and then using them to cluster various points [33]. Fiedler associates the second smallest eigenvalue of

the Laplacian matrix with connectivity and suggests partitioning by splitting vertices according to their eigenvalue in the corresponding eigenvector.

In this paper, we use an undirected graph $G = (V, E)$ to represent a WSN, where $V = \{v_1, ..., v_N\}$ denotes the set of sensor nodes and $E = \{e_1, ..., e_N\}$ indicates the set of wireless links. Let $A_{\text{mtx}} = A_{\text{mtx}}(G)$ be the adjacency matrix of graph G. In addition, the degree matrix $D_{\text{mtx}} \in \mathcal{R}^{N\times N}$ of G is a diagonal matrix where d_{ii} is the vertex degree of node i. We can get the Laplacian matrix of the graph G by

$$L_{\text{mtx}} = D_{\text{mtx}} - A_{\text{mtx}}. \tag{6}$$

We sort the eigenvalues of $L_{\text{mtx}}(G)$ in the order of $\lambda_0 = 0 \leq \lambda_1 \leq \lambda_1 \ldots \leq \lambda_{N-1}$. Fiedler investigated graph-theoretical properties of the eigenvector corresponding to λ_1, which is named Fiedler vector. For the second smallest eigenvector v_1, we define

$$V_- = \{i : v_1 < 0\}, V_+ = \{i : v_1 > 0\}, V_0 = V - V_- - V_+ \tag{7}$$

Then, the set of vertices will be defined by $V = V^+ \bigcup V^-$, where V^+ and V^- are taken as the vertex sets of two new subgraph obtained by spectral graph partitioning, respectively. Spectral graph partitioning is a method of partitioning a graph into two subgraphs in such a way that the subgraphs have the approximately equal number of vertices while minimizing the number of edges between the two subgraphs. In this paper, we use the second eigenvector Laplacian of the graph representing the WSN to determine the optimal bipartitions of a given graph.

4 Clustering algorithm

In this section, we propose our clustering algorithms, which consists of three steps: (1) Give the optimal number of clusters; (2) propose a centralized clustering algorithm and a distributed algorithm; and (3) present how to choose cluster heads and cooperative nodes.

4.1 The optimal number of clusters

It is of great significance to determine the optimal number k of clusters, because the amount of inter-cluster communications increases with k. On the other hand, the amount of intra-cluster communications grows significantly as k decreases. In the following, we will derive the optimal number of clusters by analyzing the energy model introduced in Section 3.2.

Suppose that each cluster area is a square of size $D \times D$. Given the Poisson distribution with density λ, there are λD^2 sensor nodes in each cluster on average. Thus, the number of clusters is $\frac{N}{\lambda D^2}$.

After the sensor nodes are partitioned into clusters, each cluster member sends the sensed data to its cluster head. The cluster head processes the data and then forwards them to its members that are located near the boundary of the cluster and are closer to the sink node while having sufficient residual energy. For example, as shown in Fig. 1, the WSN is partitioned into 16 clusters. Cluster A transmits data to the sink node through cooperative node A'.

Without loss of generality, we use cluster A to analyze the energy consumption in a cluster as depicted in Fig. 2. As for intra-cluster communication, the distances between cluster members and the CH are not very far; thus, we suppose that any node can transmit data to the CH directly. As described in Eq. (5), we can get the energy consumption of cluster members. In Fig. 2 , the relation between D and R is: $D = \sqrt{2}R$, where D denotes the length of cluster. Then, we calculate the expected squared distance from cluster members to the CH.

$$\begin{aligned} E\left[d_{\text{ch}}^2\right] &= \int_0^{\sqrt{2}R}\int_0^{\sqrt{2}R} \rho f(x,y)dxdy \\ &= \rho \int_{-\sqrt{2}R/2}^{\sqrt{2}R/2}\int_{-\sqrt{2}R/2}^{\sqrt{2}R/2} \left[u^2+v^2\right]dudv \\ &= \frac{2R^4}{3}\rho \end{aligned} \tag{8}$$

where $\rho = k/M^2$ and $f(x,y) = \left[\left(x-\sqrt{2}R/2\right)^2 + \left(y-\sqrt{2}R/2\right)^2\right]$. The total energy dissipated in

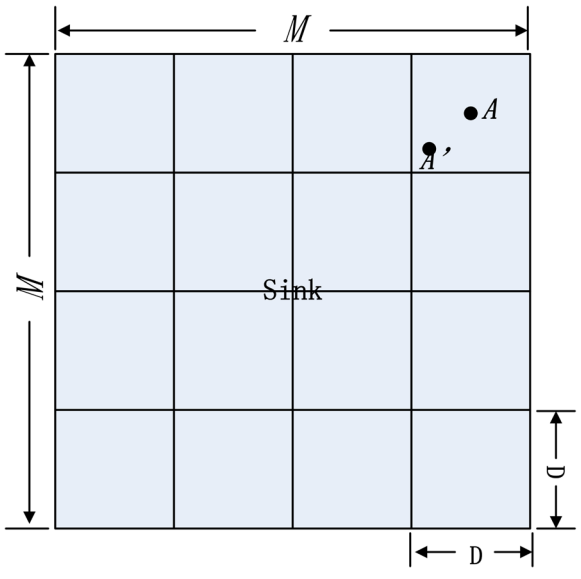

Fig. 1 Data transmission using cooperative nodes. The WSN is partitioned into 16 clusters. Cluster A transmits data to the sink node through cooperative node A'

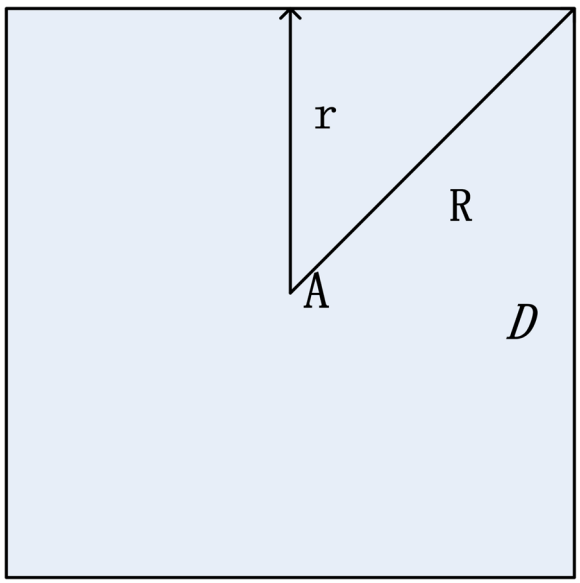

Fig. 2 The structure of cluster A. The relation between D and R is $D = \sqrt{2}R$, where D denotes the length of cluster

one round of data collection in cluster A can be calculated by

$$\begin{aligned} E_{\text{cluster}} &= E_{ch} + \left(\frac{N}{k} - 1\right) E_{\text{non-ch}} \\ &\approx E_{\text{ch}} + \left(\frac{N}{k}\right) E_{\text{non-ch}} \end{aligned} \tag{9}$$

As the cluster head does not transmit the data to another cluster head directly, E_{ch} can be derived by

$$\begin{aligned} E_{\text{ch}} &= \left(\frac{N}{k} - 1\right) * l * E_{\text{elec}} + \frac{N}{k} * l * E_{\text{DA}} \\ &\quad + l * E_{\text{elec}} + l * \varepsilon_{\text{fs}} * d_{\text{cn}}^2 \end{aligned} \tag{10}$$

where d_{cn} is the distance from the cluster head to the cooperative node. Based on Eqs. (5) and (10), we can get

$$\begin{aligned} E_{\text{cluster}} &= E_{ch} + \left(\frac{N}{k}\right) E_{\text{non-ch}} \\ &= \lambda D^2 \left(l * E_{\text{elec}} + l\varepsilon_{\text{fs}} \frac{M^2}{6k}\right) + \left(\frac{N}{k} - 1\right) * l * E_{\text{elec}} \\ &\quad + \frac{N}{k} * l * E_{\text{DA}} + l * E_{\text{elec}} + l * \varepsilon_{\text{fs}} * d_{\text{cn}}^2 \end{aligned} \tag{11}$$

and the total energy consumption for one round of data collection is

$$\begin{aligned} E_{\text{total}} &= k * E_{\text{cluster}} \\ &= \left(\frac{M^2}{D^2}\right) \times \lambda D^2 \left(lE_{\text{elec}} + l\varepsilon_{\text{fs}} \frac{M^2}{6k}\right) \\ &\quad + k \times \left(\left(\frac{N}{k} - 1\right) * l * E_{\text{elec}} \right. \\ &\quad \left. + \frac{N}{k} * l * E_{DA} + l * E_{\text{elec}} + l * \varepsilon_{\text{fs}} * d_{\text{cn}}^2\right) \end{aligned} \tag{12}$$

Then, we take the derivation of (12) with respect to k and let it be 0. We can get the optimal number of clusters, $k = \sqrt{\frac{M^2 N}{6d_{\text{cn}}^2}}$, to reach the minimum value of E_{total}.

4.2 Clustering algorithm

4.2.1 Centralized clustering algorithm

In this part, we propose a centralized clustering algorithm based on spectral partitioning method to partition the network into a fixed optimal number of clusters. We assume the sink node has full knowledge of the network topology. The sink node divides the sensor nodes into k clusters and connect all CHs.

As aforementioned, we use the second eigenvector Laplacian, also named as Fiedler vector, of the graph representing the WSN to determine the optimal bipartitions of a given graph. The process of the spectral bisection clustering is described in Algorithm 1, which consists of two phases, i.e., recursively partitioning the graph into two subgraphs and repeatedly applying the same procedure to the subgraphs. According to Algorithm 1, we can get two disjoint graphs G_1 and G_2, and the number of the nodes in G_1 and G_2 is almost the same. After the spectral partitioning, we can obtain the optimal number of clusters k. As shown in Fig. 3, 200 sensor nodes are partitioned into six clusters using our proposed spectral partitioning method, which are depicted by six colors.

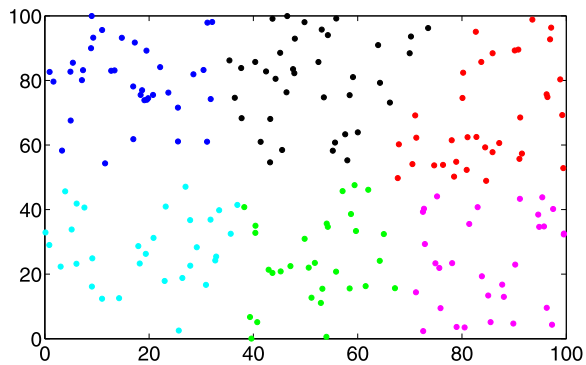

Fig. 3 The six clusters for 200 nodes by spectral partitioning method. Two hundred sensor nodes are partitioned into six clusters using our proposed spectral partitioning method

Algorithm 1 Spectral bisection algorithm

Require:
The graph $G = (V, E)$;
Ensure:
$G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$;
compute the Fiedler eigenvector v;
for each node i in G **do**
  if $v_i < 0$ **then**
    put node i in partition V_1;
    $i++$;
  else if $v_i > 0$ **then**
    put node i in partition V_2;
    $i++$;
  else
    take node i as an isolated node.
  end if
end for

4.2.2 Distributed clustering algorithm

In the previous subsection, a centralized clustering algorithm is proposed; however, it is unrealistic that each node has full knowledge of the network topology. Therefore, in this part, we would like to propose a distributed clustering algorithm by only using neighbor information of a node, which would be more applicable to WSNs. And in the simulation, we will use a centralized algorithm as a benchmark to evaluate the performance of our distributed algorithm. This is because the centralized algorithm may have the best performance by using full knowledge of the network topology. We first assume that the sink is aware of the sensing field but does not need to know the exact locations of sensor nodes. In our proposed distributed algorithm, the sink divides the field into c cluster areas by using fuzzy C-means (FCM), calculates the geographic central point of each cluster area, and broadcasts the information to all sensor nodes. The sensor nodes in each cluster elect their CH. The sensor node closest to the center of the cluster area is elected as the CH. The CHs then broadcast advertisement messages to sensor nodes to invite them to join their respective clusters.

Given a sensing field and the optimal number of clusters, the sink needs to find out the central points of c cluster areas. We adopt the clustering algorithm in [34] to divide the whole sensing field into small grids and place a virtual node at the center of each grid to represent the grid. As depicted in Algorithm 2, V' is the set of nodes in the grids and the nodes in V'' are the approximate central points of the c cluster areas in the sensing field.

After getting the geographic location of the central point of a cluster, the sensor node that is the closest to the central point will become the CH. To elect the CH, we let all nodes within the range of r from the center be the CH candidates and each candidate broadcasts a CH election message that contains its identifier and location. After a timeout, the candidate with the smallest distance to the center of the cluster among the other candidates becomes the CH node.

When a CH is elected, the CH broadcasts an advertisement message to other sensor nodes in the sensor field, to invite them to join the cluster. During this phase, each non-CH node joins the cluster with the closest CH node based on the received signal strength of the advertisement message. After that, the sensor node informs the CH node that it will be a member of the cluster by sending a short join message.

Algorithm 2 Distributed clustering algorithm

Require:
graph $G = (V, E)$, an auxiliary graph $G' = (V', E')$, a subset of nodes $V'' = c$;
Ensure:
k clusters;
for $j = 1$ && $j \varepsilon V'$ **do**
  Node j is given the coefficient u_{ij} for being a member of cluster i $u_{ij} = \frac{1}{\sum_{k=1}^{c}(d_{ij}/dkj)^{2/(m-1)}}$
end for
repeat
for $j = 1 \rightarrow k$ **do**
  compute the centroid of each cluster
  $pos(center_i) = \frac{\sum_{j=1}^{n} u_j^m pos(node)j}{\sum_{j=1}^{k} u_j^m}$
  until no change of cluster
end for
for $i = 1 \rightarrow N$ **do**
  compute distance between i and V'', d_{toch};
  if $d_{toch} < r$ **then**
    broadcast a CH *election* message to i;
  end if
  take the node with the smallest distance to v'' as cluster head;
end for
for $j = 1 \rightarrow k$ **do**
  broadcast *advertisement* messages to the sensor field
  nodes i to n decide the cluster to join into;
end for

4.3 Cooperative nodes and cluster head selection

4.3.1 Strategy to choose cooperative sensor nodes

The selection of cooperative nodes (CNs) greatly impacts the network lifetime. Thus, it is expected to design an appropriate selection strategy. During the initialization phase, the sink node broadcasts several *BEACON* messages periodically to all sensor nodes at a fixed power level.

The nodes near the sink node receive the messages and flood them to the rest of the network.

The best candidate CNs are the sensor nodes that have sufficient residual energy and receive more BEACON messages. When a sensor node v receives a BEACON message, it increases the BEACON counter n_b by one and records the signal strength s. Then, the sensor node calculates a probability of being selected as a CN, v_{chance}, based on its residual energy, the counter n_b, and the average signal strength of the received BEACON messages, i.e., v_{chance} can be calculated by

$$v_{\text{chance}} = a_1 \frac{E_{\text{res}}}{E_{\max}} + a_2 n_b + a_3 \frac{\sum s}{n_b} \tag{13}$$

where a_1, a_2, and a_3 are the weight coefficients of the residual energy, the n_b value, and the average signal strength of received BEACON messages, respectively.

Subsequently, the node v sends a candidate message containing its identification and the probability value to the cluster head. The nodes with higher probability values are more likely to be elected as CNs. The role of CNs would be rotated among cluster members when the energy level of a CN drops below an energy threshold.

4.3.2 Cluster head selection

Fuzzy logic is very suitable for implementing the heuristic clustering and its optimization like the cluster head quality classification. It is inherently robust since it does not require the precise and noise-free inputs and can be programmed to fail safely. The model of fuzzy logic control consists of a fuzzier, a fuzzy inference engine, and a defuzzier. In this paper, we use a fuzzy inference (FIS) to calculate the probability for a node to become a CH. The input variables of fuzzy inference are the residual energy E_{res} and the distance to the central node D_{tocn}, and the output is the probability of the node to be selected as a CH.

The first input variable of fuzzy logic shown in Table 1 is the distance between the sensor node and the central node, which have three values: close, medium, and far. A trapezoidal membership function is chosen for the values of close and far. On the other hand, the membership function of medium is a triangular membership function.

The second one is the residual energy of the sensor node, with the values of low, rather low, medium, rather high, and high. The values of low and high correspond to a trapezoidal membership function, while other values of the variable use a triangular membership function.

Based on the two fuzzy input variables, 18 fuzzy mapping rules are defined in Table 1. We can derive the fuzzy output of probability by the fuzzy rules. This fuzzy variable has to be transformed into a single crisp number that is a form we can use in practice. This process is called defuzzification, and we induce the center of area (COA) by the defuzzification method like Eq. (14),

Table 1 Fuzzy mapping rules

Distance to CN	Residual energy	Probability
Close	High	Very high
Close	Rather high	Rather high
Close	Medium	High
Close	Rather low	Very low
Close	Low	Very low
Close	Very low	Very low
Medium	High	High
Medium	Rather high	Rather high
Medium	Medium	Medium
Medium	Rather low	Medium
Medium	Low	Very low
Medium	Very low	Very low
Far	High	Medium
Far	Rather high	Medium
Far	Medium	Rather low
Far	Rather low	Rather low
Far	Low	Low
Far	Very low	Very low

$$\text{output} = \frac{\int x * \mu_{\text{chance}}(x) dx}{\int x dx}, \tag{14}$$

where $\mu_{\text{chance}}(x)$ denotes the membership function of the fuzzy set of probability. A node which holds more residual energy and close to the cooperative sensor node has a higher probability to become a CH.

5 Simulation results

In this section, we present the simulation results to evaluate our proposed algorithms. On the establishment of the network model, we assume that the sensor nodes are distributed in a 100 m × 100 m area. The sink node is located at a central point (50, 50). The number of sensor nodes varies according to the simulation respects. The parameters used in the simulations are described in Table 2. We compare our algorithms with HEED algorithm from four respects: the number of rounds until the first node dies, the number of alive sensor nodes over time, the evolution of the remaining energy in the network, and the impact of the initial energy quantity on the performance. This is because similar to our algorithm, HEED algorithm also considers residual energy to select cluster head. We ignore the effect caused by the transmission collisions and the interference in wireless channels.

Table 2 Configuration parameters

Parameter	Value
E_{elec}	50 nJ/bit
ε_{fs}	10 pJ/bit/m^2
ε_{amp}	0.0013 pJ/bit/m^4
E_{DA}	5 nJ/bit/message
Size of message	4000 bits
Initial energy	0.5 j

5.1 Comparison of the number of rounds for the first node dead

It is necessary that all sensor nodes stay alive as long as possible because network performance decreases once there is a node to die. Thus, it is important to know when the first node dies. The time when the first node dies in the simulations for all compared algorithms can be found in Figs. 4 and 5. In Fig. 4, the initial energy is fixed but the network model is changed (i.e., the number of sensor nodes is variable). And in Fig. 5, the initial energy is changed but the number of sensor nodes is fixed. We examine separately how network models and energy models affect our algorithm performance. Figure 4 shows the performances of these clustering algorithms by increasing the number of sensor nodes N from 100 to 300. When there are 200 nodes in the given sensing area, the first node dead occurs at the round of 700 by our distributed algorithm, while it is about at 600th round by HEED algorithm. It can be seen from Fig. 5 that when the initial energy of sensor nodes is increased from 0.5 to 1.5 J and then up to 2.5 J, the advantage of our algorithms over HEED becomes more obvious in terms of prolonging the network lifetime.

It can be observed that our proposed algorithm can greatly improve the network lifetime compared to HEED

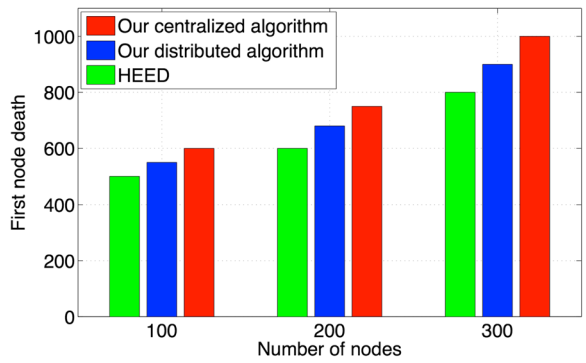

Fig. 4 Impact of the node density. It shows the performances of these clustering algorithms by increasing the number of sensor nodes N from 100 to 300

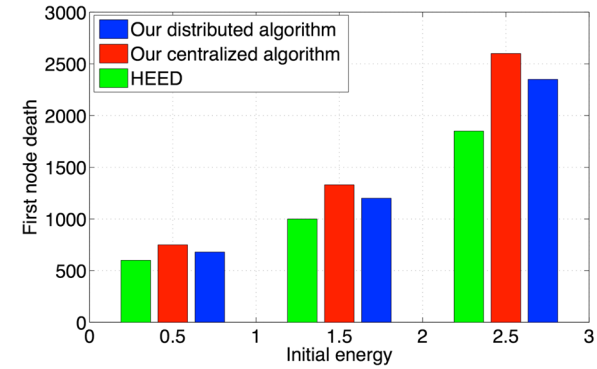

Fig. 5 Impact of the initial energy. When the initial energy of sensor nodes is increased from 0.5 to 1.5 and then up to 2.5, the advantage of our algorithms over HEED becomes more obvious in terms of prolonging the network lifetime

because they have the larger number of rounds when the first node dies. It can also be seen from the figures that our algorithms can operate efficiently as the node density and initial energy increase. As the first node's death time can reflect the robustness of algorithm, our algorithms are also shown to be robust. This is because the scalability, lifetime, and efficiency of the whole network depend upon the optimal number of clusters and the spatial position of cluster head. It is worth noting that the optimal number of clusters is calculated by the energy model (12). Compared to HEED, our algorithm can obtain the optimal number of clusters and provide an optimal cluster head selection strategy to minimize energy consumption.

Figures 4 and 5 show that our centralized algorithm outperforms our distributed algorithm. This is because the BS has global knowledge of the location and energy of all the nodes in the network, so it can produce the optimal number of clusters so that the energy consumption for data transmission is minimized. Moreover, we adopt the spectral partitioning method in our centralized algorithm to divide the network into clusters before the process of the cluster head selection. Furthermore, the simple implementation of our algorithm based on the computation of the matrix eigenvectors can make it easier to divide the network into clusters and reduce the energy consumption of clusters in the formation phase.

5.2 Comparison of the nodes' lifetime

In this subsection, we set the number of sensor nodes to 200, and the initial of each node is 0.5 J. Figure 6 shows the number of nodes alive over time for all compared algorithms. It is clear that our algorithms can improve the network lifetime (the difference between the time that the first node dies and the time that the last node dies)

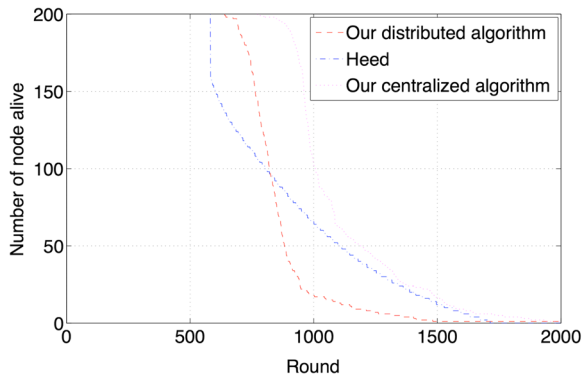

Fig. 6 Number of nodes alive over time. Our algorithm can improve the network lifetime (the difference between the time that the first node dies and the time that the last node dies) compared to HEED

compared to HEED. In HEED, tentative cluster heads are randomly selected based on their residual energy. Therefore, the sensors with low residual energy could become cluster heads because they use the intra-cluster communication cost to select final cluster heads. Furthermore, the energy consumption of cluster heads is not well-balanced. In our algorithms, we adopt two strategies to balance the energy consumption among nodes. We first choose cooperative nodes to reduce the energy consumption of cluster heads and then choose cluster heads based on FIS. The formation of clusters based on the remaining energy of nodes allows the nodes with low energy levels to have a lower probability of choice. Thus, we can efficiently prolong the lifetime of network.

5.3 Comparison of the remaining energy

Figure 7 depicts the total residual energy in the network after a large number of rounds of data collection. The network residual energy decreases rapidly in HEED than our algorithms. We can see that after 1000 rounds of data collection, approximately 95% of the total energy is consumed in the HEED. However, only 82% of the total energy is consumed in our centralized algorithm. This improvement is attributed to the consideration of the distance to central nodes and the strategy to choose cooperative nodes. It can efficiently reduce energy consumption in both intra-class and inter-class formation. As we have obtained the central node of each cluster before cluster formation, it is quite easy to get the distance between central nodes and candidate nodes.

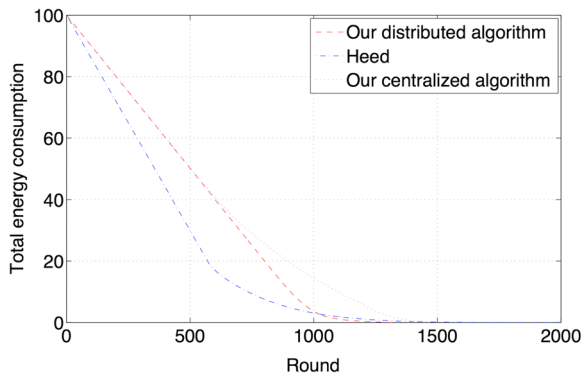

Fig. 7 Remaining energy in the network. It shows the changes of the total residual energy in the network after a large number of rounds of data collection. The network residual energy decreases rapidly in HEED than in our algorithms

In fact, the consideration in cluster head selection can ensure the nodes close to the central nodes have higher probability to become cluster head. The proposed algorithm can also make the nodes distributed uniformly for the different clusters. Besides, our proposed scheme succeeds to rotate the cooperative nodes based on the distance to BS and residual energy. Clearly, cooperative nodes can save the energy consumption of inter-class data transmission. The simulation results confirm that our algorithms give a significant performance improvement in terms of energy and lifetime, compared to HEED protocols.

6 Conclusions

In this paper, we aim to get an efficient way to prolong the lifetime of wireless sensor network in the background of big data. In order to reduce the energy consumption of cluster head, we use cooperative nodes to relay data to the sink node. Based on the energy model, we get a reasonable number of clusters which can balance the energy consumption of inter-cluster communication and intra-cluster communication. Furthermore, we propose a centralized algorithm and a distributed algorithm to divide the network into clusters which can extend the lifetime of network. To balance the energy consumption sensor node, we propose an efficient strategy to choose cooperative nodes and cluster heads. It has been shown by simulations that our proposed algorithm achieves a significant performance improvement.

However, our algorithm has also some limits. One limit is that the network topology is required to keep unchanged over time, and sensor nodes are deployed uniformly. Another limit is that all the nodes are assumed to be homogeneous and have the same energy consumption model and each node is aware of its own position through RSSI localization. In particular, the proposed fuzzy logic-based clustering algorithm is heuristic, which may lead to the failure of clustering.

Abbreviations

BS: Base station; BSD: Big sensory data; CFFL: Cluster formation using fuzzy logic; CH: Cluster head; CN: Cooperative node; COA: The center of area; FCM: Fuzzy C-means; FIS: Fuzzy inference; HEED: Hybrid energy-efficient distributed;

LEACH: Low-energy adaptive clustering hierarchy; RSSI: Received signal strength indication; WSN: Wireless sensor network

Acknowledgements
The authors acknowledged the anonymous reviewers and editors for their efforts in valuable comments and suggestions.

Funding
The work described in this paper was partially supported by the Fundamental Research Funds for the Central Universities (XDJK2016A011, XDJK2015C010, XDJK2015D023, XDJK2016D047, XDJK201710635069), the National Natural Science Foundation of China (61772432,61772433,61503309), and the Natural Science Key Foundation of Chongqing (cstc2015jcyjBX0094).

Availability of data and materials
The datasets supporting the conclusions of this article are included within the article.

Authors' contributions
Both QW and JH presented the research subject, designed the model and algorithm, performed the data analyses and numerical simulation, and wrote the draft; SG and YY provided the main idea of this work, reviewed the draft, and revised the final manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]National & Local Joint Engineering Laboratory of Intelligent Transmission and Control Technology (Chongqing), College of Electronic and Information Engineering, Southwest University, 2 Tiansheng Road, Beibei District, 400715 Chongqing, China. [2]Department of Electrical and Computer Engineering, Stony Brook University, 100 Nicolls Road, Stony Brook, 11794, NY, USA.

Received: 12 December 2017 Accepted: 26 February 2018
Published online: 06 March 2018

References

1. Z Taghikhaki, N Meratnia, PJM Havinga, in *Proc.2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*. A reliable and energy-efficient chain-cluster based routing protocol for wireless sensor networks. (IEEE, Melbourne, 2013), pp. 248–253
2. G Asada, M Dong, TS Lin, F Newberg, G Pottie, WJ Kaiser, HO Marcy, in *Proc. 24th European Solid-State Circuits Conference*. Wireless integrated network sensors: low power systems on a chip. (IEEE, The Hague, 1998), pp. 9–16
3. S Cheng, Z Cai, J Li, H Gao, Extracting kernel dataset from big sensory data in wireless sensor networks. IEEE Trans. Knowl. Data Eng. **29**(4), 813–827 (2017)
4. X Zhang, Z Cai, Real-time big data delivery in wireless networks: a case study on video delivery. IEEE Trans. Ind. Inform. **13**(4), 2048–2057 (2017)
5. T Zhu, S Cheng, Z Cai, J Li, Critical data points retrieving method for big sensory data in wireless sensor networks. EURASIP J. Wirel. Commun. Netw. **2016**(1), 1–18 (2016)
6. S Cheng, Z Cai, J Li, Curve query processing in wireless sensor networks. IEEE Trans. Veh. Technol. **64**(11), 5198–5209 (2015)
7. S Cheng, Z Cai, J Li, X Fang, in *2015 IEEE Conference on Computer Communications (INFOCOM)*. Drawing dominant dataset from big sensory data in wireless sensor networks (IEEE, Kowloon, 2015), pp. 531–539
8. D Kumar, Performance analysis of energy efficient clustering protocols for maximising lifetime of wireless sensor networks. Wirel. Sens. Syst. **4**(1), 9–16 (2014)
9. P Nayak, B Vathasavai, Energy efficient clustering algorithm for multi-hop wireless sensor network using type-2 fuzzy logic. IEEE Sensors J. **17**(14), 4492–4499 (2017)
10. F AL-Obaidy, H Zereshkian, FA Mohammadi, in *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*. A energy-efficient routing algorithm in zigbee-based cluster tree wireless sensor networks. (IEEE, Windsor, 2017), pp. 1–5
11. WB Heinzelman, A Chandrakasan, H Balakrishnan, An application-specific protocol architecture for wireless microsensor networks. Proc. IEEE Trans. Wirel. Commun. **1**(4), 660–670 (2002)
12. V Katiyar, N Chand, GC Gautam, A Kumar, in *Proc. 2011 International Conference on Emerging Trends in Electrical and Computer Technology*. Improvement in leach protocol for large-scale wireless sensor networks. (IEEE, Nagercoil, 2011), pp. 1070–1075
13. M-W Park, J-Y Choi, Y-J Han, T-M Chung, in *INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference On*. An energy efficient concentric clustering scheme in wireless sensor networks. (IEEE, Seoul, 2009), pp. 58–61
14. ASK Mammu, A Sharma, U Hernandez-Jayo, N Sainz, in *2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA)*. A novel cluster-based energy efficient routing in wireless sensor networks. (IEEE, Barcelona, 2013), pp. 41–47
15. JS Lee, TY Kao, An improved three-layer low-energy adaptive clustering hierarchy for wireless sensor networks. IEEE Internet Things J. **3**(6), 951–958 (2016)
16. A Bazregar, A Movaghar, A Barati, MRE Nejhad, H Barati, in *3rd International Conference on Information and Communication Technologies: From Theory to Applications (ICTTA 2008)*. Notice of violation of ieee publication principles a new automatic clustering algorithm via deadline timer for wireless ad-hoc sensor networks. (IEEE, Damascus, 2008), pp. 1–6
17. A Manjeshwar, DP Agrawal, in *Proc. 15th IEEE International Conference on Parallel and Distributed Processing Symposium*. Teen: a routing protocol for enhanced efficiency in wireless sensor networks. (IEEE, San Francisco, 2001), pp. 2009–2015
18. A Manjeshwar, DP Agrawal, in *Proc. International Conference on Parallel and Distributed Processing Symposium (IPDPS 2002)*. Apteen: a hybrid protocol for efficient routing and comprehensive information retrieval in wireless. (IEEE, Ft. Lauderdale, 2002), pp. 8–19
19. C Liu, S Guo, Y Shi, Y Yang, Deterministic binary matrix based compressive data aggregation in big data WSNs Telecommunication Systems. **66**(3), 345–356 (2017)
20. WR Heinzelman, A Chandrakasan, H Balakrishnan, *Energy-efficient communication protocol for wireless microsensor networks*, (2000), pp. 10–20
21. O Younis, S Fahmy, Distributed clustering in ad-hoc sensor networks: a hybrid, energy-efficient approach. Proc. 23th Annu. Joint Conf. IEEE Comput. Commun. Soc. **1**, 640 (2004)
22. LB Bhajantri, AV Sutagundar, in *2016 International Conference on Computing, Analytics and Security Trends (CAST)*. Fuzzy logic based cluster head selection and data processing in distributed sensor networks. (IEEE, Pune, 2016), pp. 282–288
23. F Ahamad, R Kumar, in *2016 IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*. Energy efficient region based clustering algorithm for WSN using fuzzy logic, (2016), pp. 1020–1024
24. HE Alami, A Najid, in *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*. Cffl: Cluster formation using fuzzy logic for wireless sensor networks. (IEEE, Marrakech, 2015), pp. 1–6
25. J Lee, W Cheng, Fuzzy logic based clustering approach for wireless sensor networks using energy predication. IEEE Sensors J. **12**(9), 2891–2897 (2012)
26. SJ Dastgheib, H Oulia, MRS Ghassami, An efficient approach for clustering in wireless sensor network using fuzzy logic. Proc. 2011 Int. Conf. Comput. Sci. Netw. Technol. **3**, 1481–1486 (2011)
27. Y Hu, X Shen, Z Kang, in *Proc. 5th International Conference on Wireless Communications, Networking and Mobile Computing*. Energy-efficient cluster head selection in clustering routing for wireless sensor networks. (IEEE, Beijing, 2009), pp. 1–4
28. JC Bezdek, Pattern recognition with fuzzy objective function algorithms. Adv. Appl. Pattern Recog. **22**(1171), 203–239 (1981)
29. D Han, J Ji, Y Dai, G Li, W Fan, H Chen, in *2016 International Conference on Progress in Informatics and Computing (PIC)*. Improved fuzzy C-means algorithm and its application to classification of remote sensing image in Chengdu City, China. (IEEE, Shanghai, 2016), pp. 437–443
30. M Li, L Zhang, Z Xiang, E Castillo, T Guerrero, in *2016 International Conference on Progress in Informatics and Computing (PIC)*. An improved

fuzzy C-means algorithm for brain MRI image segmentation. (IEEE, Shanghai, 2016), pp. 336–339
31. JK Parker, LO Hall, JC Bezdek, in *2012 IEEE International Conference on Fuzzy Systems (IEEE FUZZ)*. Comparison of scalable fuzzy clustering methods. (IEEE, Brisbane, 2012), pp. 1–9
32. M-C Hung, D-L Yang, in *Proc. IEEE International Conference on Data Mining (ICDM 2001)*. An efficient fuzzy C-means clustering algorithm. (IEEE, San Jose, 2001), pp. 225–232
33. S Ding, L Zhang, Y Zhang, Research on spectral clustering algorithms and prospects. Proc. 2th Int. Conf. Comput. Eng. Technol. **6**, 149–153 (2010)
34. R Xie, X Jia, Transmission-efficient clustering method for wireless sensor networks using compressive sensing. IEEE Trans. Parallel Distrib. Syst. **25**(3), 806–815 (2014)