Research Progress Report

Botao Zhu

August 21, 2019

1 Reading and Research Activities

An improved soft K-means clustering protocol for balancing energy in non-uniform distributed wireless sensor networks

Abstract
Energy load balance is the essential issue in designing the wireless sensor networks (WSNs). Clustering technique is utilized as an energy-efficient routing to balance the network energy and prolong the lifetime, which is influenced by the cluster head selection and clustering method. In this paper, we propose an approach based on the idea clustering by fast search and ind of density peaks (CFSFDP) and kernel density estimation (KDE) to optimize the selection of the initial cluster heads of clustering algorithm. And then, we use reassigning member nodes and multi-heads scheme to balance the energy consumption of all nodes in the whole network. Simulation results demonstrate that the proposed approach can postpone the first node death by almost 3 times and the half of nodes death by 2 times compared to LEACH. Also, smoother energy distribution curve of all nodes in different round and smaller energy variance are obtained by our proposed algorithm.

1.1 Introduction

Energy limitation is the key challenges in WSNs, since the batteries of nodes are not rechargeable, which is finite and draining of the batteries may make sensing are uncovered. Further more, it is impossible to replace the sensor nodes when their energy is exhausted. So, energy consumption is the most important issue for WSNs [1, 2]. Clustering has been well researched for a long time in WSNs, which group the sensor nodes into distinct clusters with a head in each and each sensor nodes belongs to only one cluster. All member nodes sense the data and send it to the cluster head (CH), at the same time, the CHs collect and process the data and send it to the base station (BS) via single-hop or multi-hop [3]. A hierarchical based WSN [4] has many advantages: increasing scalability of the network, efficient data aggregation and utilizing channel bandwidth efficiently.

At present, some research on wireless sensor networks are based on the assumption that the sensor nodes are randomly distributed throughout the entire network [5] [6], which may obtain non-uniform distribution of the sensor nodes. The main problem of non-uniform clustering is that it can lead to high energy dissipation of sensor node, total energy consumption increases, and network connectivity not being guaranteed. [7] [8] proposed W-LEACH algorithm to handle non-uniform network. In order to extend the lifetime of network, it only chooses a number of sensors in each cluster to send data to their CHs based on a fixed percentage, which can reduce the traffic load of CHs. However, it may result in the loss of important information. Hence, it is significant to study non-uniform distributed wireless networks.

K-means clustering algorithm is used by [9] firstly in LEACH for achieving better classification performance compared to the traditional LEACH. [10] also combines LEACH and K-means in the cause of creating symmetric clusters and reducing the average intra-cluster communication distance. However, there may be several drawbacks we need to consider. Firstly, K-means algorithm chooses the initial centroid randomly, which cannot ensure the clustering result is optimal. If the initial centroids are far away from the true centroid of clusters, it will cause the results that the number of iterations will increase until the algorithm converges. Secondly, the node with the highest residual energy in the cluster is elected as the CH. It can lead to a potential problem: unbalanced energy load and excessive energy expenditure. For example, node CH1 in Fig.1 has higher residual energy than other nodes in the cluster A, which means node CH1 can be elected as the CH with the highest probability. However, this way enables the other nodes in the same cluster to send data in the opposite direction

to the BS, causing higher energy consumption. Thirdly, it does not consider that some of sensor

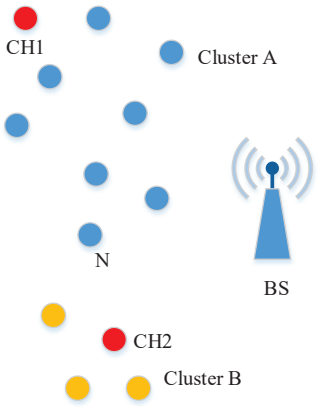


Figure 1: A example of clustering

nodes in a cluster can be reallocated to another cluster for balancing the size of different clusters. As shown in Fig.1, node N belongs to cluster A by applying hard classification algorithm, K-means. However, if we view this question from the perspective of probability, node N joins the cluster A with a certain probability, 55%. Then, it will have a 45% probability to join the cluster B. Because the density of cluster A is greater than that of cluster B, node N can be allocated to cluster B in order to reduce the stress of CH of cluster A. Fourthly, each cluster only selects one cluster head, so the energy of cluster head will be depleted quickly for aggregating, compressing and transmitting information when there are too many sensor nodes in the cluster. [11] proposed a method to prolong the lifetime of each round by balancing the energy consumption of among nodes in the network. The nodes are divided into three types in each cluster: CH, vice CH and member nodes. The vice CH plays the role of CH when the CH dies before the completion of current round, which can diminish the frequency of re-clustering and extend network lifetime. [12] also comes up with a similar concept. Three nodes in each cluster are selected as CHs to ensure the load is evenly balanced in different clusters. However, if these two methods are applied to non-uniform distributed wireless networks, the problem of energy balance is still unsolved.

In this paper, an improved soft K-means (ISK-means) clustering protocol is proposed to address above four problems. Specifically, the main contributions of this paper are as follows.
1) In order to obtain a faster convergence speed and a higher probability of the global optimum, we choose the initial centroids of soft K-means [13] clustering algorithm by using the idea of density from clustering by fast search and find of density peaks (CFSFDP), which is implemented by kernel density estimation (KDE).
2) According to the characteristics of soft K-means, reassigning member nodes scheme is employed to adjust the number of nodes in each cluster.
3) Multi-heads method is used to balance traffic load of CHs of different clusters.

The rest of this paper is organized as follows: Section 2 reviews the related work. The theories of soft K-means and kernel density estimation are described in Section 3 and Section 4 respectively. Section 5 gives the energy model and network model. Section 6 proposes the implementation of ISK-means algorithm. In section 7, we compare the performance of ISK-means with other protocols. Finally, the conclusion is presented in Section 8.

1.2 Related works

Energy efficiency is one of the main aim for maintaining WSN because the energy of each sensor nodes is limited. An efficient clustering protocol is one of the core goals for WSN, because it can reduce energy consumption. [14] proposed a modified K-means algorithm in which they take into account two factors:distance and remaining energy of nodes. The results show it can effectively reduce energy consumption and extend the lifespan compared to LEACH. In order to minimize the sum of Euclidean distances between the head and member nodes, [15] provides a novel way to choose cluster heads by using K-means method for maximizing the energy efficiency of WSN. However, this algorithm may cause different kinds of clusters in different runs based on the chosen initial centroids randomly. Balanced Parallel K-means (BPK-means) [16] is used to group the nodes and CHs are chosen depending on their distance from cluster center and residual energy. But, with the random deployment of the nodes, K-means may result in vagueness of classifying nodes near the boundary of the clusters.

Most of the deployment in WSN is random, where sensor nodes are distributed unevenly and the number of nodes is not uniform in each cluster. Some efforts are made to come up with some solutions regarding the uniform distribution of load and to achieve energy efficiency through non-uniform deployment of nodes. [17] proposed an unequal clustering size (UCS) model for network organization, which can lead to more uniform energy dissipation among the cluster head nodes, thus increasing network lifetime. Area covered by the clusters can be altered in each layer by changing radius of a layer near to BS and hence will change density of a particular cluster. The drawback of this approach is the number of nodes in each cluster may vary to a great extent. A new model for non-uniform deterministic node distribution is proposed [18] that reduces the total number of nodes to be deployed over the area and the network traffic satisfying both coverage and load balancing criteria. The simple distributed algorithm is introduced to load balanced data gathering. However, this approach may cause energy hole problem. [19] proposed fuzzy based unequal clustering protocol (FUCP), which is a novel clustering algorithm which incorporate CH selection algorithm by using fuzzy logic and relay traffic distributed to eliminate hot spot problem. It has more longer network lifetime and sends more packets to BS compared to distributed energy efficient hierarchical clustering. Energy degree distance unequal clustering algorithm (EDDUCA) [20] aims to balance energy consumption and maximize the network life, which use 'Sierpinski' method to divide the network into unequal clusters. The results indicate that EDDUCA can effectively balance the energy consumption. Energy-aware clustering algorithm (EAC) [21] considers two factors: the energy factor for cluster head selection and distance factor for non-cluster heads to select its cluster head, which achieves a good performance in terms of lifetime by balancing the energy load among all the sensor nodes in the network. An energy-aware clustering and cluster-based routing algorithm [22] uses competition range to construct clusters of even size for nonuniform deployment of sensor nodes to balance the load across the entire network. The cluster head is selected on the basis of the ratio of average remaining energy of nearby nodes and the energy of node itself, which achieves load balance among cluster heads.

1.3 Soft K-means

[13] K-means is the simplest clustering algorithm in unsupervised learning, which partitions the data set in to k clusters using some distance measurement methods, like Euclidean distance. It is a hard clustering method, that is to say the membership degree of one node has only two values 0 and 1 for a specific class. However, in some cases, there are a few data points for which it is not quite so obvious to which cluster they belong. Soft K-means clustering decides to which degree each data point belongs to , the assignments to clusters will be probabilistic. Generally speaking, soft K-means clustering can be seen as the problem of finding k cluster centroids with the aim of minimizing the

error function. Given a set of data points $X = \{x_1, x_2, ..., x_n\}$, the error function is

$$E(\mu_1, \mu_2, ..., \mu_k) = \sum_{i=1}^{k} E(\mu_i) = \sum_{i=1}^{k} \sum_{j=1}^{n} z_{ij} ||x_j - \mu_i||^2 \tag{1}$$

where μ_i is the centroid of each cluster, z_{ij} is the indicator variable.
For the traditional K-means clustering

$$z_{ij} = \begin{cases} 1, \text{if } x_j \in C_i \\ 0, \text{otherwise} \end{cases} \tag{2}$$

where C_i represents cluster i, this equation means whether x_i belongs to cluster C_i. However, z_{ij} is not a integer value for soft K-means clustering

$$z_{ij} = \frac{e^{-\beta ||x_j - \mu_i||^2}}{\sum_{l=1}^{k} e^{-\beta ||x_j - \mu_l||^2}} \tag{3}$$

where β is the stiffness parameter, its impact for clustering result will be discussed in simulation section. From equation (3), we can get $z_{ij} \in [0, 1]$ and $\sum_i z_{ij} = 1$.
How to update the centroid μ_i of each cluster until convergence? For a specific class C_i

$$E(\mu_i) = \sum_{j=1}^{n} z_{ij} ||x_j - \mu_i||^2 \tag{4}$$

The target is to minimize the error function, so the problem can be turned into the following optimization problem

$$e = \text{argmin} \sum_{j=1}^{n} z_{ij} ||x_j - \mu_i||^2 \tag{5}$$

which is a convex function and has a unique minimum value. The equation (4) can be written as

$$E(\mu_i) = \sum_{j=1}^{n} z_{ij} (x_j - \mu_i) \top (x_j - \mu_i) \tag{6}$$

$$= \sum_{j=1}^{n} z_{ij} (x_j \top x_j - 2 x_j \top \mu_i + \mu_i \top \mu_i) \tag{7}$$

deriving $E(\mu_i)$ with respect to μ_i and denoting to zero, we can have

$$\mu_i = \frac{\sum_{j=1}^{n} z_{ij} x_j}{\sum_{j=1}^{n} z_{ij}} \tag{8}$$

Each cluster updates the centroid according to equation (3) and (8) until the probabilities of the clusters in which the data points are located remain unchanged or the maximum number of iterations are reached.

1.4 Kernel density estimation

The first law of geography states that all things are interrelated. The closer they are, the stronger they are. Kernel density estimation is based on this law, its value decreases gradually with the

increase of the central radiation distance. From two-dimensional or three-dimensional surface of density, we can intuitively acquire the features of point sets [23].

For an observation value x of random variable X, the probability that it falls into the interval $[a, b]$ can be computed by

$$P = \int_a^b \hat{p}(x)\, dx \tag{9}$$

where $\hat{p}$ is the probability density function. When $|b-a| \ll \varepsilon$, $\varepsilon \to 0$, equation (9) becomes to

$$P = \hat{p}(x) \int_a^b dx = \hat{p}(x)(b-a) \tag{10}$$

Hence

$$\hat{p} = \frac{P}{b-a} \tag{11}$$

If there are k observation value of n falling into the interval $[a, b]$, the probability will be

$$P = \frac{k}{n} \tag{12}$$

the probability density function

$$\hat{p} = \frac{k}{n(b-a)} \tag{13}$$

We define the kernel function

$$K(u) = \begin{cases} 1, |u| \leq \frac{1}{2} \\ 0, \text{otherwise} \end{cases} \tag{14}$$

Assuming the center of interval $[a, b]$ is x and $h = b - a$, any sample x_i falling into the interval $[a, b]$ needs to meet the following requirement.

$$|x - x_i| \leq \frac{b-a}{2} \tag{15}$$

$$\frac{|x - x_i|}{h} \leq \frac{1}{2} \tag{16}$$

From equation (14) and (16), we can obtain

$$K\left(\frac{x - x_i}{h}\right) = \begin{cases} 1, |\frac{x-x_i}{h}| \leq \frac{1}{2} \\ 0, \text{otherwise} \end{cases} \tag{17}$$

So, the value k can be expressed by

$$k = \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{18}$$

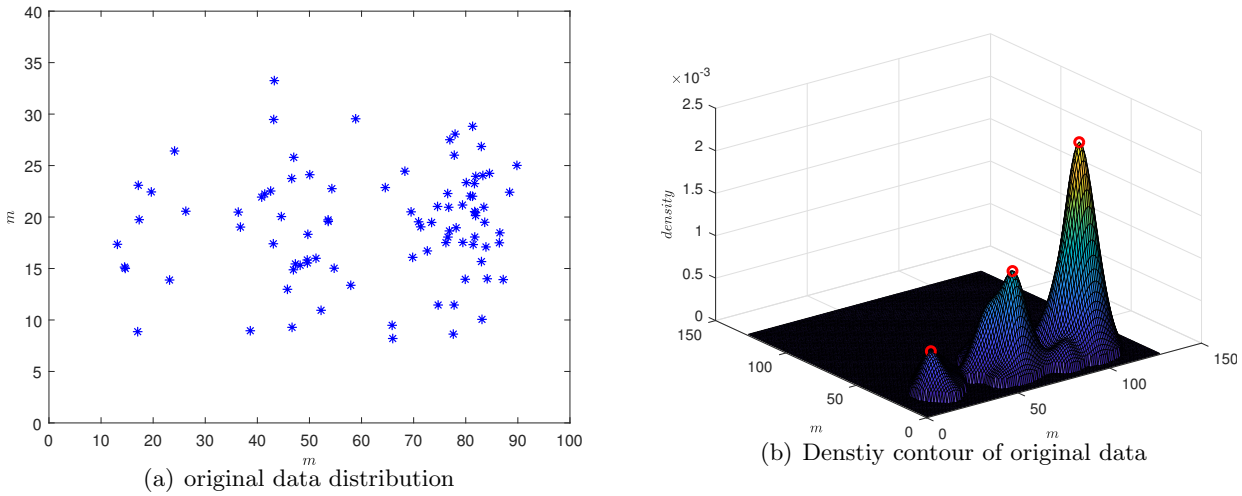
and then,

$$\hat{p}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{19}$$

which is the kernel density function.

[24] [25]Kernel density is one of the non-parametric estimation methods used to estimate the unknown density function. Each point is covered with a smooth surface. The surface value at the location of the point is the highest and decreases with the increase of distance until the value is zero when the

distance equals the search radius, which looks like the density map of Gauss distribution function. Then, the values of kernel density estimation of all points are superimposed and a set of points is transformed into a surface showing continuous density changes. Fig.2 is an example of kernel density estimation for a non-uniform distribution network. The discrete point set is transformed into a smooth density map, shown in Fig.2(b), which shows its spatial distribution. The higher the density value, the higher the aggregation degree of the point is.



(a) original data distribution

(b) Denstiy contour of original data

Figure 2: A example of kernel density estimation

1.5 Network Model and Energy Model

1.5.1 Network Model

We have following assumptions regarding the network model [26] [27]:
(a) The BS is a high energy node and is located far away from the extremities of the sensor network.
(b) All sensor nodes having the same computational and transmission capabilities. In other words, all nodes are capable of acting as CH nodes.
(c) The sensor nodes can vary the power with which they transmit signals according to the received signal strength indication of a particular node.
(d) The sensor nodes scans its environment at a fixed rate and will contain data to be sent to the BS at all intervals.
(e) The sensor nodes and BS are all static after deployment.
(f) The sensor nodes in general have location information such as a GPS support.
(g) The CHs perform data compression to reduce the amount of bits transmitted to the BS.
(h) The BS indicates all nodes to reinitiate clustering when all the CH nodes in the network have insufficient energy.

1.5.2 Energy model

Because the main energy consumption of the protocol is only for receiving and sending data, the first order radio model as the energy model is in line with the needs [28]. The consumption energy in the transmitter nodes and in the receiver node can be calculated as follows:

$$E_T = \begin{cases} l * E_{elec} + l * \varepsilon_{fs} * d^2 & d \leq d_0 \\ l * E_{elec} + l * \varepsilon_{mp} * d^4 & d > d_0 \end{cases} \tag{20}$$

$$E_R = l * E_{elec} \tag{21}$$

where E_{elec} is the dissipated energy per bit in both transmitter nodes and receiver nodes, d is the transmission distance and d_0 is defined as the distance threshold, $d_0 = \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}}$, l is the packet size. The free space ε_{fs} and multipath fading channel ε_{mp} represent the energy consumption parameters of amplifier, which one can be used depends on the distance between the transmitter and the receiver.

If the transmission distance d_{toBS} between the cluster head and the sink node is larger than d_0 and the distance between member node and the cluster head d_{toCH} is less then d_0, the energy consumption of cluster head can be calculated in this round by

$$E_{CH} = k * l * c * \left(E_T + E_{DA} + \varepsilon_{mp} * d_{toBS}^4\right) + k * l * E_R \tag{22}$$

where E_{DA} represents the dissipated energy of data aggregation and c is the data aggregation ratio. The first part of equation (22) is the energy consumption of the cluster head for sending aggregated data to the sink node and the second part is the energy consumption of receiving and aggregating data of k member nodes. The energy consumption of a non-cluster node sending data to the cluster head by

$$E_{nonCH} = l * E_T + l * \varepsilon_{fs} * d_{toCH}^2 \tag{23}$$

Hence, the residual energy of node i in each round r can be computed by

$$E_i(r) = \begin{cases} E_i(r-1) - [k * l * c * \left(E_T + E_{DA} + \varepsilon_{mp} * d_{toBS}^4\right) + k * l * E_R] & i \in CH \\ E_i(r-1) - l * E_T + l * \varepsilon_{fs} * d_{toCH}^2 & i \notin CH \end{cases} \tag{24}$$

where $E_i(r-1)$ is the residual energy for node i in the $r-1$ round and CH is the cluster heads.

1.6 Proposed Algorithm

Routing protocols can be classified many types, such as flat protocol and hierarchy protocol. The proposed protocol belongs to hierarchy protocol, which is divided into many rounds, like LEACH. Each round contains setup phase and steady phase. In this research, we only focus on the setup phase and improve some existing methods. During the setup phase, the sink node gathers the location and residual energy of each node in the whole wireless network. Also, it generates k initial centers by CFSFDP and KDE as the input of ISK-means clustering algorithm, which can avoid the local optimum by traditional K-means clustering method. And then, classification algorithm is implemented. Each node calculates the distance between itself and center points and chooses to join the nearest cluster with a certain probability. At the same time, if the nodes on the edge of two clusters have similar distance to these two cluster centers, they will preferentially join the cluster with smaller density. After clustering, the final CHs are selected according to the energy. The detailed discussion is given below.

1.6.1 Selection of Initial Cluster Heads

[29] discovered that clusters can be recognized regardless of their shape and the dimensionality of the space in which they are embedded. The basic idea is that cluster centers are surrounded by neighbors with lower local density and they are at a relatively large distance from any points with a higher local density. Hence, the larger the number of nodes in its neighborhood is, the higher the value of local density will be. The cluster heads are selected by the maximum distance σ and relatively high local density p.

In this paper, we use the similar idea of density maximum to select the initial cluster heads. However, we only consider the density parameter p, which can be obtained by KDE. The nodes with highest

local density will be chosen as the cluster heads, because the area with the larger number of nodes can easily form the maximum local density in non-uniform distribution wireless networks.

$$\mathrm{N}_{p_{max}} = \max\left(\mathrm{N}_{\hat{p}_i}\right), i \in \text{local region nodes} \tag{25}$$

According to above equation, we can obtain a maximum density node set $\{\mathrm{N}_{p_{max1}}, \ldots, \mathrm{N}_{p_{maxK}}\}$. After the number of maximum density regions, K, is determined, each normal node joins the nearest cluster head with higher density to form initial clusters by CFSFDP algorithm. Then, the nodes with the largest energy in each initial cluster are selected as the input of improved soft K-means clustering algorithm. There are several benefits through the above steps to ensure the initial cluster heads: (1) the number of clusters k is determined by maximum density principle. (2) the distances between the greatest energy nodes of each cluster are relatively large. The detailed description is shown in algorithm 1.

Algorithm 1: Selection of initial cluster heads

Input: the set of n data items $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$
Output: initial cluser heads set
for $i = 1, 2, \ldots, n$ **do**
  calculate p_i by equation (19)
end
obtain density set $P = \{p_1, \ldots, p_n\}$
get local maximum density node set $\{\mathrm{N}_{p_{max1}}, \ldots, \mathrm{N}_{p_{maxK}}\}$ by statistical tool as the initial cluster heads
each member node joins its cluster head $\mathrm{N}_{p_{maxK}}$ via CFSFDP algorithm to form K initial clusters $\{C_1, \ldots, C_K\}$
for $j = 1 : K$ **do**
  calculate the highest energy nodes of C_j
end
the set of the highest energy nodes of each cluster $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$

1.6.2 Cluster Formation

Machine learning has been widely used in wireless sensor networks for forming cluster, such as distributed K-means clustering algorithm [30], improved K-means cluster-based routing [31] and LEACH-CKM [32]. The above algorithms are based on the distances between normal nodes and cluster head nodes to form clusters, which can easily lead to a large gap in the number of different clusters in non-uniform distribution networks. Furthermore, the energy consumption of cluster head nodes is not balanced. Hence, compared with these K-means clustering algorithms, our proposed algorithm uses soft K-means clustering algorithm. Each node is assigned a probability of belonging to cluster head rather than completely being a member of just one cluster. Therefore, the nodes close to the boundary of clusters may have the similar probabilities belonging to different clusters. And, we also take one more step after the classification convergence, assigning nodes at the edge of different clusters to join the approximate cluster for balancing the number of normal nodes in clusters. We discuss this problem in two scenarios.

Scenario 1: The node N is at the edge of two clusters, shown in Fig. 3. The distance from node N to the center of cluster B is a little bigger than that from node N to the center of cluster A, which means it has a higher probability to join cluster A. However, it is a different story if we consider the densities of different clusters. Cluster A has 5 member nodes and cluster B has 10 member nodes. Assuming that all normal nodes send messages to its CH in each round, CH of cluster B will deal with more information from its member nodes. In order to balance the energy consumption of CHs,

it is better for node N to join cluster A. Here, we give a simple definition of re-assigning nodes. When the difference of probability of a node belonging to two clusters is less than 25%, the node will join the cluster with low density.

Scenario 2: The node N is at the boundary of three or more clusters, shown in Fig.4. After clustering, each node will have a probability set representing the possibilities of belonging to different clusters, $\boldsymbol{z}_i = \{z_{i1}, z_{i2}, \dots z_{iK}\}$. The cluster of the node belonging to with low probability will not be considered. So, we only choose the first two maximum probabilities and calculate their difference, D_i. If $D_i < 10\%$, the node will join the cluster with smaller density between two maximum probability clusters. As shown in Fig. 4, node N has similar probability belonging to cluster A, B and C, which finally joins cluster C because cluster C has lower density than cluster A and B.

Algorithm 2: Cluster formation

Input:
Initial cluster heads $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_K\}$ from **algorithm** 1
Data items $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$
Iterations r_{max}
Output:
K clusters

for $r = 1 : r_{max}$ **do**
  for $i = 1 : n$ **do**
    $z_{ik} = \frac{e^{-\beta||x_i - \mu_k||^2}}{\sum_K e^{-\beta||x_i - \mu_k||^2}}$
  end
  $\mathbf{z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1K} \\ \vdots & \vdots & \vdots & \vdots \\ z_{i1} & \cdots & \cdots & z_{iK} \end{bmatrix}$
  for $k = 1 : K$ **do**
    $\mu_k = \frac{\sum_{i=1}^{n} z_{ki} x_i}{\sum_{i=1}^{n} z_{ki}}$
  end
end
K clusters $\boldsymbol{C} = \{C_1, \dots, C_K\}$
for $i = 1 : n$ **do**
  check z_i and reassign the node i by scenario 1 and 2
end
K clusters $\boldsymbol{C}_{new} = \{C_{new1}, \dots, C_{newK}\}$

1.6.3 Selection of Final Cluster Heads

In non-uniform distributed wireless sensor networks, the cluster size will be different. If only one CH is selected for each cluster, CH will consume too much energy to deal with the information from its member nodes in the large cluster, which will cause its death too early. Hence, our proposed algorithm designs the scheme of multi-cluster heads. The number of cluster head nodes is not fixed in each cluster, which is determined by the number of the nodes in the cluster. The larger of the number of nodes in the cluster is, the more the number of CH will be. In a cluster, the nodes which have more remaining energy and close to the center are preferentially selected as CHs. The total remaining energy of cluster k can be computed by:

$$E(k)_{total} = \sum_{i=1}^{S_k} E_i(r) \tag{26}$$

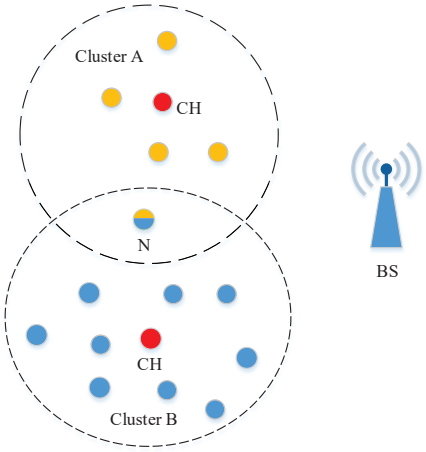


Figure 3: Node at the boundary of two clusters

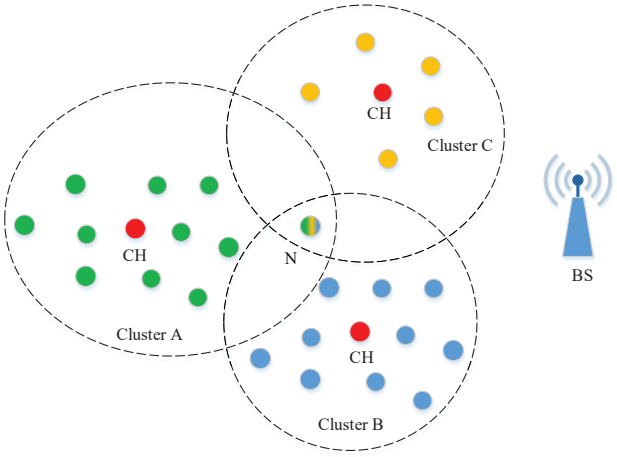


Figure 4: Node at the boundary of three clusters

where S_k is the size of cluster k, $E_i(r)$ is the residual energy of node i in current round, which can be obtained by equation (24).
The average energy of cluster k is:

$$E_{ave} = \frac{E(k)_{total}}{S_k} \tag{27}$$

This scheme can balance the energy consumption of cluster heads of different cluster in non-uniform distributed wireless sensor networks, shown in Fig.5.

1.6.4 Switching Cluster Head

In the first round, the first node in CH_k is selected as the current CH for cluster k. After CHs for the current round are selected, the sink node notifies all nodes to join the cluster to which they belong. CHs broadcast TDMA schedules to their member nodes for transmitting data in different time slots to avoid data collision. Then, the network is in a steady phase and begins to exchange data between normal nodes and their CHs. For balancing the energy consumption of CH, when the energy of current CH of any cluster is below the threshold value, the next candidate CH in that cluster is enabled. Until all CH in the cluster are executed, the algorithm starts re-clustering.

Algorithm 3: Selection of final cluster heads

Input:
K clusters $\mathbf{C}_{new} = \{\mathrm{C}_{new1}, \ldots, \mathrm{C}_{newK}\}$
Output:
K cluster heads

```
for k = 1 : K do
    Calculating the size of cluster C_newk, S_k
    Calculating average energy of cluster E_ave
    Num = S_k / constant, the number of CHs of cluster
    for i = 1 : S_k do
        Iterating x_i from near the center of cluster
        if E_i > E_ave and Num > 0 then
            Num = Num − 1
            CH_k (Num) = x_i
        end
    end
end
```

K cluster heads $\mathbf{CH} = \{\mathrm{CH}_1, \ldots, \mathrm{CH}_K\}$

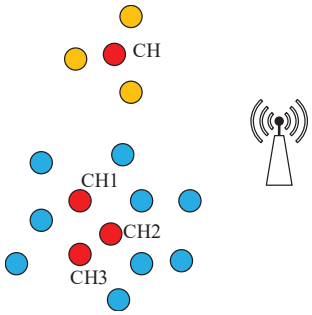


Figure 5: Multi-heads scheme

Algorithm 4: Switching cluster head mechanism

Input:
K cluster heads $\mathbf{CH} = \{\mathrm{CH}_1, \ldots, \mathrm{CH}_K\}$
Output:
Next cluster head

```
Current round
for k = 1 : K do
    ρ = (the residual energy of CH_k last round) / (the residual energy of CH_k current round)
    if ρ < Threshold then
        if CH_k has next then
            switching next CH
        else
            reclustering
        end
    end
end
```

1.7 Experiment results and analysis

1.7.1 Simulation settings

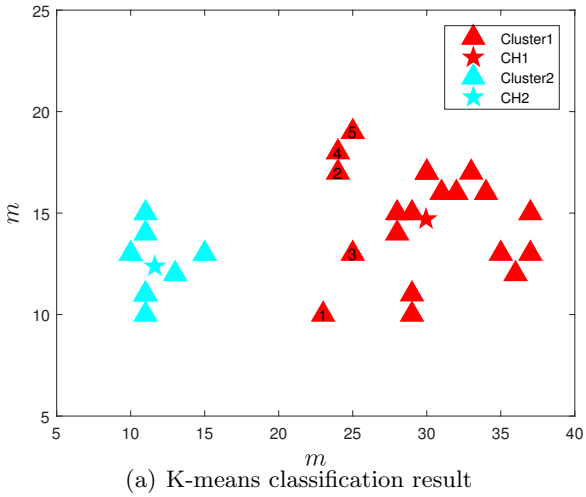
The simulation is executed in MATLAB R2017a. Sensor nodes are deployed in an area of $100 \times 100m^2$, the sink node is located at $(50m, 150m)$. The main parameters are shown in Table 1.

Table 1: Simulation parameters

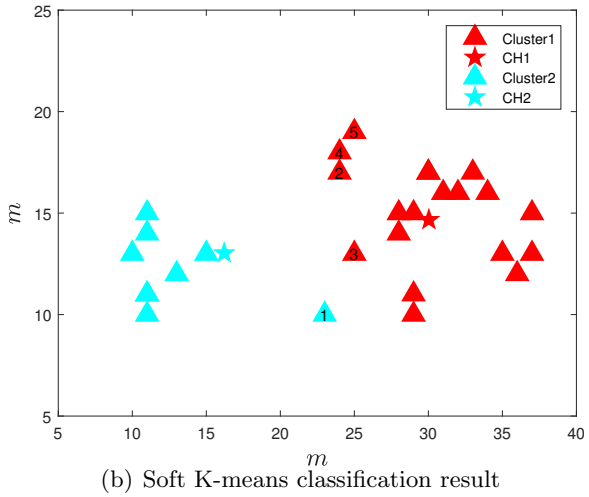
Parameter	**Value**
Area	$100m \times 100m$
Sink node	(50,150)
Initial energy	0.2J,1J
Packet length	4000bits
Control length	100bits
E_T	50nJ/bit
E_R	50nJ/bit
ε_{fs}	10pJ/bit
ε_{mp}	0.0013pJ/bit
E_{DA}	5nJ/bit
d_0	88m
Number of sensor nodes	26,100

1.7.2 Reassigning nodes of improved soft K-means analysis

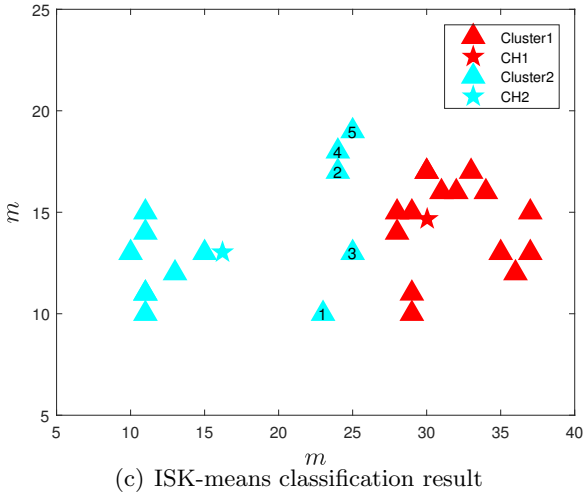
In this section, we will discuss the impact of reassigning nodes scheme of improved soft K-means for balancing energy consumption of CH. A non-uniform distributed network with 26 nodes is generated. Firstly, we use K-means classification method to classify these nodes and get two clusters, shown in Fig.6(a). It is found that cluster 1 contains 19 nodes, which is much higher than cluster 2. As a result, CH of the heavily loaded cluster, which contains 26 nodes will be exhausted much earlier than another cluster. Fig.6(b) shows the similar classification result by soft K-means method when β is 0.3, which can not balance the number of different clusters very well. However, we find that the nodes on the edge of two clusters have similar probabilities belonging to these two clusters, like node 1, node 2, node 3, node 4 and node 5. At the same time, when the value of β changes, the probabilities also change. Table 2 lists the probabilities of the above nodes belonging to different clusters and their change for different β. In section 1.2, we know β represents stiffness parameter, which is the tightness of node belong to a cluster. As can be seen from Table 2, when $\beta = 1$, all five nodes belong to the cluster with a higher probability than the case, $\beta = 0.3$. Hence, the larger the value of β, the closer the nodes is associated with a cluster. We set $\beta = 0.3$ in the simulation for using the ISK-means algorithm better. According to the principles described in scenario 1, node 2, node 3, node 4 and node 5 are reassigned to cluster 2 from cluster 1, which balances the number of these two clusters. The residual energy of CH, computed by equation (24), in each round could be used to check the result and advantage of this scheme. Fig.7(a) and Fig.7(b) show that both K-means and soft K-means result in unbalanced energy consumption of CH, however, the ISK-means method achieves an equilibrium of energy consumption in both CH.



(a) K-means classification result



(b) Soft K-means classification result



(c) ISK-means classification result

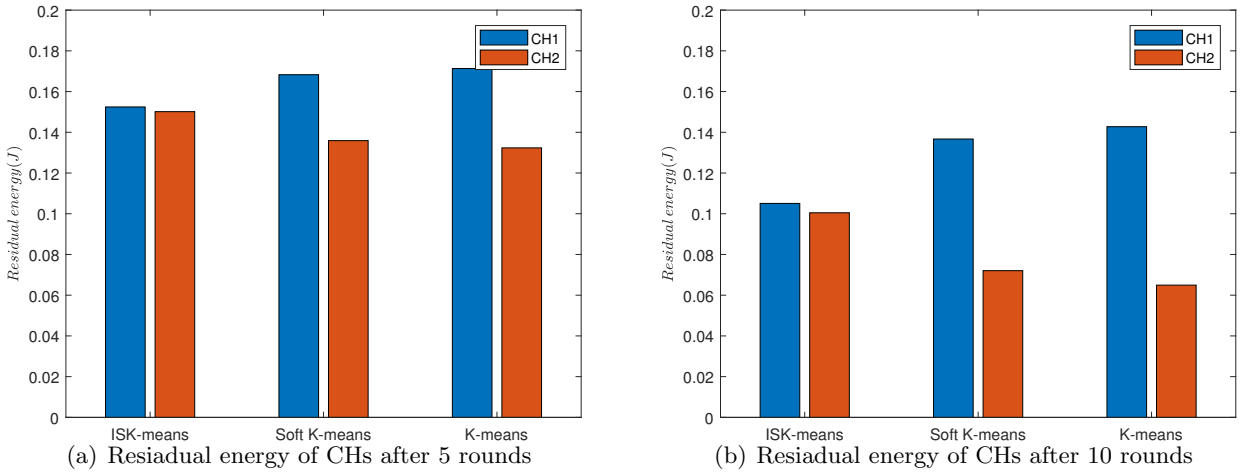
Figure 6: Comparison of different classification methods

Table 2: Probabilities comparison

Probability		Node 1	Node 2	Node 3	Node 4	Node 5
$\beta = 0.3$	Cluster 1	0.4852	0.5537	0.5684	0.6120	0.6120
	Cluster 2	0.5148	0.4463	0.4316	0.3880	0.3880
$\beta = 1$	Cluster 1	0.0438	0.9787	0.9860	0.9992	0.9992
	Cluster 2	0.9562	0.0213	0.014	0.0008	0.0008

1.7.3 Network lifetime

To test the performance of ISK-means, we compare it with LEACH [28] and K-means [9]. Fig.8 shows the first node death(FND), half of nodes death(HND) and the last node death(LND) for three methods when the number of nodes is 100 and the area is $100m \times 100m$. If the protocol can balance energy very well, the first node death will be very late. In this paper, we assume that the death of 95% nodes means all nodes dead. In view of Fig.8 and Fig.9, the sensor nodes lifetime of our proposed algorithm is better than that of LEACH and K-means. For K-means, the round of FND is 191, which is much earlier than 962 in LEACH and 2663 in ISK-means. However, the LND in K-means happens



(a) Resiadual energy of CHs after 5 rounds (b) Resiadual energy of CHs after 10 rounds

Figure 7: Comparison of residual energy of CHs

later compared with LEACH. It is obvious that the energy consumption of K-means is unbalanced, this is because K-means algorithm selects the initial cluster heads randomly, which is sensitive to noise and abnormal data, especially in non-uniform distributed wireless networks. The proposed ISK-means algorithms takes energy into account. The result shows it can effectively postpone the FND, HND and LND. The FND of ISK-means is 2663, which is about 3 times that 962 of Leach and 14 times of K-means. The HND is also around 2 times of both LEACH and K-means, which means ISK-means can balance the energy consumption of the network effectively. Obviously, ISK-means can keep most of the nodes alive in the network, so that the existence of the death of first nodes happens later. As a result,

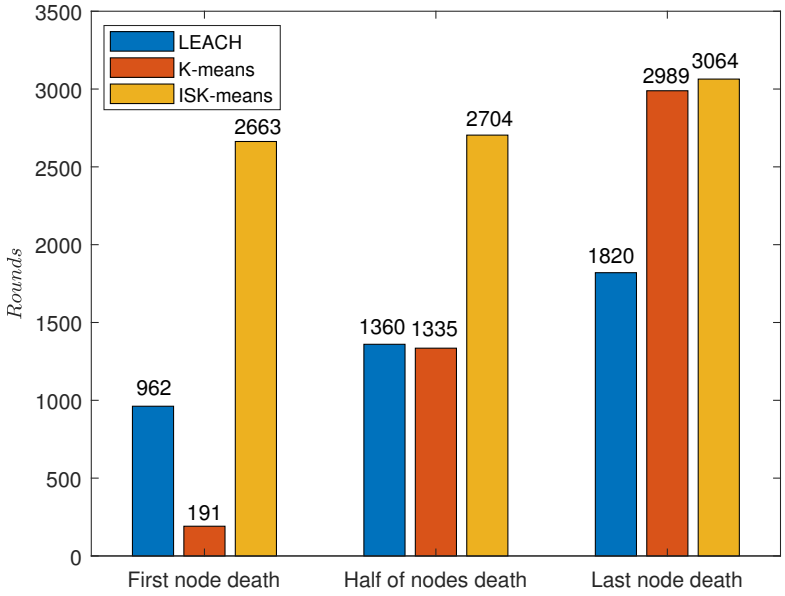


Figure 8: Comparison of network lifetime

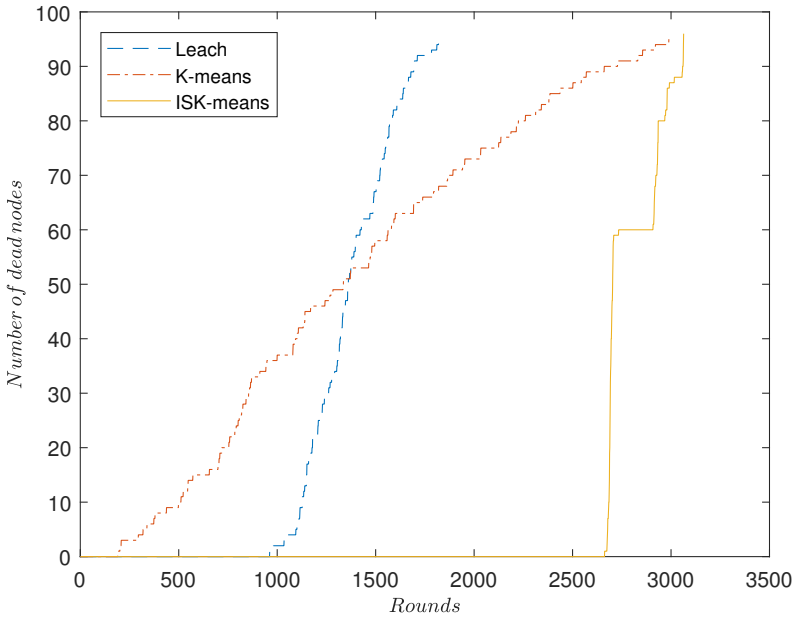


Figure 9: Comparison of node death curve

1.7.4 Received packets at BS

Fig.10 shows the received packets at BS for among three algorithms. Because of using multi-heads scheme in ISK-means, the number of re-clustering decreases, which reduces the communication cost of clustering between CHs and their member nodes. As a result, data transmission frequency increases and more packets are successfully transmitted to the BS compared to that in LEACH and K-means.

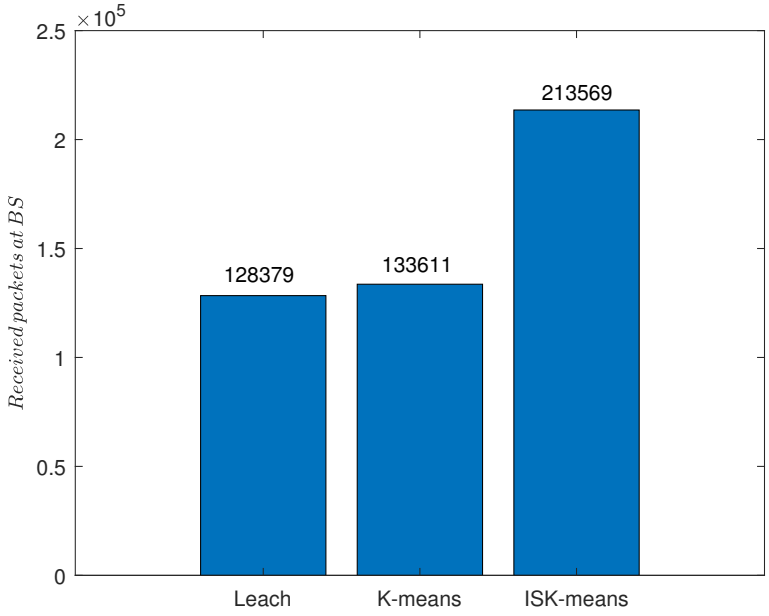


Figure 10: Received packets at BS

1.7.5 Energy Variance

Fig.11 shows the comparison of residual energy of all 100 nodes in the network for among three algorithms after different rounds. It is found that the energy distribution curve of all nodes by using ISK-means protocol is smoother than that of LEACH and K-means, and the curve of K-means is the worst. This result demonstrates that ISK-means is good at balancing energy consumption of all nodes in the whole network.
For purpose of estimating the performance of proposed algorithm, we denote a new parameter: energy variance(EV), can be expressed by

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(E\left(x_i\right) - \overline{E}\right)^2}{n} \tag{28}$$

where $E\left(x_i\right)$ is the energy of node i in current round and $\overline{E}$ is the average energy of all nodes. In table 3, it clearly reveals that ISK-means obtains smaller variances than LEACH and K-means in different rounds, which demonstrates that ISK-means can keep the energy distribution of 100 nodes in the network to be the most uniform.

Table 3: Comparison of energy variance in different rounds

	Variance							
	200 r	400 r	600 r	800 r	1000 r	1200 r	1400 r	1600 r
LEACH	0.0011	0.003	0.0062	0.0108	0.017	0.017	0.0088	0.0026
K-means	0.0345	0.0752	0.1043	0.1194	0.1158	0.1028	0.0907	0.0659
ISK-means	0.00028	0.0005	0.0007	0.00083	0.001	0.0011	0.0013	0.0016

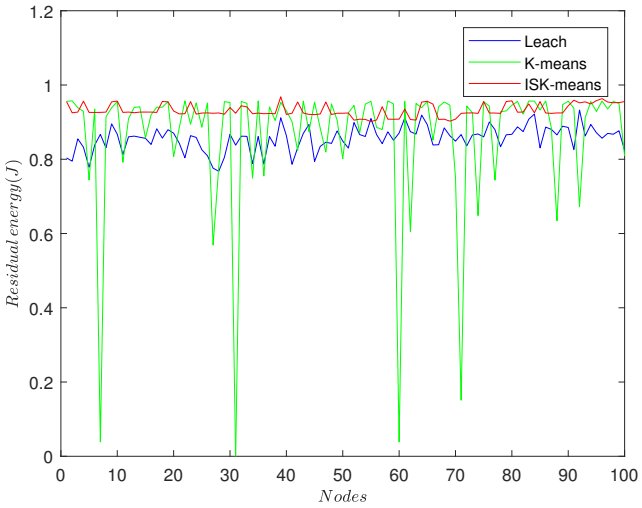
1.8 Conclusions

In this paper, we propose an energy balanced ISK-means algorithm protocol based on soft K-means for non-uniform distributed wireless sensor networks. Firstly, it optimizes the selection of initial cluster heads of soft K-means clustering method by CFSFDP and KDE algorithms and a better cluster formation is obtained. Secondly, in order to balance the size of different clusters in non-uniform networks, we use the soft classification characteristics of soft K-means to reassign some nodes locating at the boundary of different clusters to smaller size cluster. Furthermore, multi-heads scheme is used in the selection of final cluster heads, which can effectively balance the traffic load of cluster heads, reduce the number of re-clustering and save communication cost. Experiments have demonstrated that ISK-means algorithm can balance energy very well for all nodes in the network during the period of network survival and the amount of data transmitted to the BS is increased remarkably.

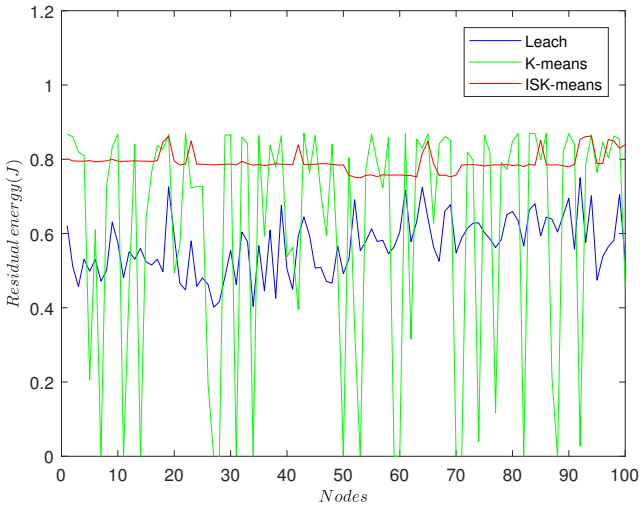
2 Objectives for the Next 2 Weeks

Simulate the algorithm proposed above.

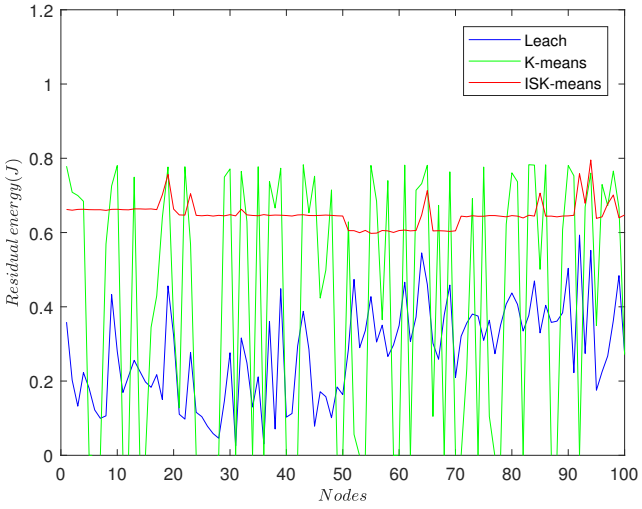
3 Advisor's Comments



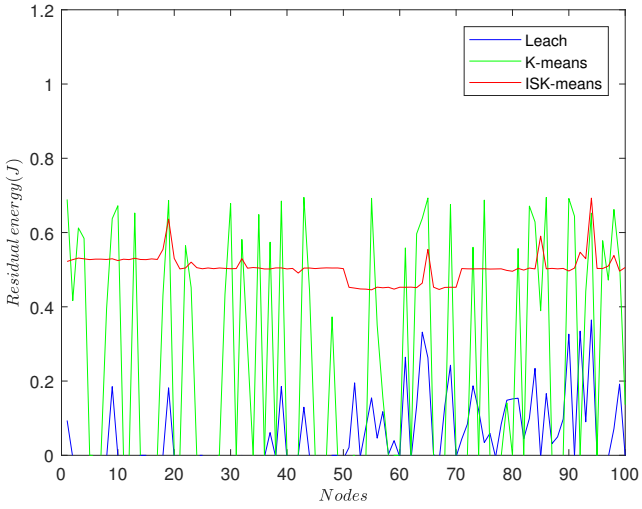
(a) Residual energy after 200 rounds



(b) Residual energy after 600 rounds



(c) Residual energy after 1000 rounds



(d) Residual energy after 1400 rounds

Figure 11: Comparison of residual energy distribution

Bibliography

[1] T. M. Behera, S. K. Mohapatra, U. C. Samal, M. S. Khan, M. Daneshmand, and A. H. Gandomi, "Residual energy-based cluster-head selection in wsns for iot application," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5132–5139, June 2019.

[2] A. Antoo and A. R. Mohammed, "Eem-leach: Energy efficient multi-hop leach routing protocol for clustered wsns," in *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, July 2014, pp. 812–818.

[3] S. K. Singh, P. Kumar, and J. P. Singh, "A survey on successors of leach protocol," *IEEE Access*, vol. 5, pp. 4298–4328, 2017.

[4] S. Zafar, A. Bashir, and S. A. Chaudhry, "Mobility-aware hierarchical clustering in mobile wireless sensor networks," *IEEE Access*, vol. 7, pp. 20 394–20 403, 2019.

[5] N. M. A. Latiff, N. N. N. A. Malik, and L. Idoumghar, "Hybrid backtracking search optimization algorithm and k-means for clustering in wireless sensor networks," in *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, Aug 2016, pp. 558–564.

[6] H. Echoukairi, A. Kada, K. Bouragba, and M. Ouzzif, "A novel centralized clustering approach based on k-means algorithm for wireless sensor network," in *2017 IEEE Computing Conference*, July 2017, pp. 1259–1262.

[7] H. M. Abdulsalam and L. K. Kamel, "W-leach: Weighted low energy adaptive clustering hierarchy aggregation algorithm for data streams in wireless sensor networks," Dec 2010, pp. 1–8.

[8] H. M. Abdulsalam and B. A. Ali, "W-leach based dynamic adaptive data aggregation algorithm for wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 9, no. 9, p. 289527, 2013.

[9] P. Sasikumar and S. Khara, "2012 fourth international conference on computational intelligence and communication networks," Nov 2012, pp. 140–144.

[10] M. Bidaki, R. Ghaemi, and S. R. K. Tabbakh, "Towards energy efficient k-means based clustering scheme for wireless sensor networks 1," 2016.

[11] A. S. D. Sasikala and N. Sangameswaran, "Improving the energy efficiency of leach protocol using vch in wireless sensor network," *International Journal of Engineering Development and Research*, vol. 3, no. 2, pp. 918–924, 2015.

[12] S. Periyasamy, S. Khara, and S. Thangavelu, "Balanced cluster head selection based on modified k-means in a distributed wireless sensor network," *International Journal of Distributed Sensor Networks*, vol. 12, no. 3, p. 5040475, 2016.

[13] C. Bauckhage, "Lecture notes on data science: Soft k-means clustering," 2015.

[14] S. Randhawa and S. Jain, "Performance analysis of leach with machine learning algorithms in wireless sensor networks," *International Journal of Computer Applications*, vol. 147, no. 2, pp. 7–12, 2016.

[15] G. Y. Park, H. Kim, H. W. Jeong, and H. Y. Youn, "A novel cluster head selection method based on k-means algorithm for energy efficient wireless sensor network," in *2013 27th IEEE International Conference on Advanced Information Networking and Applications Workshops*, March 2013, pp. 910–915.

[16] L. Tan, Y. Gong, and G. Chen, "A balanced parallel clustering protocol for wireless sensor networks using k-means techniques," in *2008 IEEE Second International Conference on Sensor Technologies and Applications (sensorcomm 2008)*, Aug 2008, pp. 300–305.

[17] S. Soro and W. B. Heinzelman, "Prolonging the lifetime of wireless sensor networks via unequal clustering," in *19th IEEE International Parallel and Distributed Processing Symposium*, April 2005, pp. 8 pp.–.

[18] P. Chatterjee and N. Das, "Coverage constrained non-uniform node deployment in wireless sensor networks for load balancing," in *2014 Applications and Innovations in Mobile Computing (AIMoC)*, Feb 2014, pp. 126–132.

[19] S. Gajjar, A. Talati, M. Sarkar, and K. Dasgupta, "Fucp: Fuzzy based unequal clustering protocol for wireless sensor networks," in *2015 39th IEEE National Systems Conference (NSC)*, Dec 2015, pp. 1–6.

[20] A. B. F. Guiloufi, N. Nasri, and A. Kachouri, "An energy-efficient unequal clustering algorithm using 'sierpinski triangle' for wsns," *Wireless Personal Communications*, vol. 88, no. 3, pp. 449–465, Jun 2016. [Online]. Available: https://doi.org/10.1007/s11277-015-3137-0

[21] M. Mohamed-Lamine, "New clustering scheme for wireless sensor networks," in *2013 8th IEEE International Workshop on Systems, Signal Processing and their Applications (WoSSPA)*, May 2013, pp. 487–491.

[22] J. Yu, Y. Qi, G. Wang, and X. Gu, "A cluster-based routing protocol for wireless sensor networks with nonuniform node distribution," *AEU - International Journal of Electronics and Communications*, vol. 66, no. 1, pp. 54 – 61, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1434841111001312

[23] W. Härdle, *Applied Nonparametric Regression*, ser. Econometric Society Monographs. Cambridge University Press, 1990.

[24] C. B, "Lecture notes on machine learning: Kernel k-means clustering," 2019.

[25] T. Buch-larsen, J. P. Nielsen, M. Guillén, and C. Bolancé, "Kernel density estimation for heavy-tailed distributions using the champernowne transformation," *Statistics*, vol. 39, no. 6, pp. 503–516, 2005.

[26] N. Srikanth and M. Ganga Prasad, "Efficient clustering protocol using fuzzy k-means and mid-point algorithm for lifetime improvement in wsns," *International Journal of Intelligent Engineering and Systems*, vol. 11, pp. 61–71, 08 2018.

[27] A. Ray and D. De, "Energy efficient clustering protocol based on k-means (eecpk-means)-midpoint algorithm for enhanced network lifetime in wireless sensor network," *IET Wireless Sensor Systems*, vol. 6, no. 6, pp. 181–191, 2016.

[28] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 660–670, Oct 2002.

[29] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014. [Online]. Available: https://science.sciencemag.org/content/344/6191/1492

[30] J. Qin, W. Fu, H. Gao, and W. X. Zheng, "Distributed k -means algorithm and fuzzy c -means algorithm for sensor networks based on multiagent consensus theory," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 772–783, March 2017.

[31] M. Lehsaini and M. B. Benmahdi, "An improved k-means cluster-based routing scheme for wireless sensor networks," in *2018 IEEE International Symposium on Programming and Systems (ISPS)*, April 2018, pp. 1–6.

[32] D. Mechta, S. Harous, I. Alem, and D. Khebbab, "Leach-ckm: Low energy adaptive clustering hierarchy protocol with k-means and mte," in *2014 IEEE 10th International Conference on Innovations in Information Technology (IIT)*, Nov 2014, pp. 99–103.