A

Major Project-II Report

On

Optimized Clustering Approach for Wireless Sensor Network

Submitted in Partial Fulfillment of the Requirement for the Degree of

Master of Technology
In
Computer Science and Engineering

By
Joginder Singh
2K15/CSE/07

Under the Guidance of
Mr. R.K. Yadav
(Assistant Professor)

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahabad Daulatpur, Main Bawana Road,

Delhi-110042

June 2017

CERTIFICATE

This is to certify that Major Project-II Report entitled "**Optimized Clustering Approach for Wireless Sensor Network**" submitted by **Joginder Singh, Roll no. 2k15/CSE/07** for partial fulfillment of the requirement for the award of degree Master Of Technology (Computer Science and Engineering) is a record of the candidate work carried out by him under my supervision.

Project Guide
Mr. R.K. Yadav
Assistant Professor
Department of Computer Science & Engineering
Delhi Technological University

DECLARATION

I, hereby declare that the Major Project-II work entitled "**Optimized Clustering Approach for Wireless Sensor Network"** which is being submitted to Delhi Technological University, in partial fulfillment of requirements for the award of the degree of Master of Technology (Computer Science and Engineering) is a bonafide report of Major Project-II carried out by me. The material contained in the report has not been submitted to any university or institution for the award of any degree.

Joginder Singh

2K15/CSE/07

ACKNOWLEDGEMENT

First of all, I would like to express my deep sense of respect and gratitude to my project supervisor Mr. R K Yadav for providing the opportunity of carrying out this project and being the guiding force behind this work. I am deeply indebted to him for the support, advice and encouragement he provided without which the project could not have been a success.

Secondly, I am grateful to Dr. Rajni Jindal, HOD, Computer Science & Engineering Department, DTU for his immense support. I would also like to acknowledge Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out.

Last but not the least I would like to express sincere gratitude to my parents and friends for constantly encouraging me during the completion of work.

Joginder Singh

Roll No – 2k15/CSE/07

MTech(Computer Science & Engineering)

Delhi Technological University

ABSTRACT

Wireless sensor network play an important role in today's networks which are mainly focusing on automating the computing needs of the world. These networks can work better in an environment which is unattended and inaccessible. As such networks have constrained potentials such as battery power, radio transmission range, memory and processing hardware of the sensor nodes which makes them less robust in case of node failures, inefficient, poor scalability. One of the main unique features of WSN is that they have wireless connectivity among the small size sensor nodes which makes them work even in harsh conditions with limited human support. WSNs is a type of an ad-hoc networks that involves the large number of small, lightweight sensor nodes. Here, the energy of nodes is a critical issue in WSN because of limited battery life. There are some of the challenging issues in WSN like energy consumption, traffic load balancing, data aggregation, data redundancy, limited wireless communication range.

In this thesis, we have worked towards designing of new energy efficient clustering algorithms to achieve higher energy performance along with longer life span of the network in WSNs. we have proposed two new improved clustering algorithms for the purpose of efficient clustering analysis of data sets to offer high energy efficiency and reduced communication overheads. The first algorithm will solve the difficulty of finding the optimal number of clusters with their cluster centroids and arbitrarily shaped clusters set by uniting model and density based algorithms. It also presents an idea for estimating input parameters to minimize intra and inter-cluster distance between sensor nodes. The proposed hybrid algorithm reduces wireless data transmissions and increases the speed of transferring the accumulated data to the mobile sink node located at the cluster centroid that patrols by following the shortest path. The second algorithm presents a new idea for cluster formation phase and cluster head selection phase based on density of the sensor node. It not only finds the optimal number of clusters in every round, but also gives an idea of electing cluster head based on density using residual energy, average energy and distance from the location of cluster centroid. After cluster formation phase, all the member nodes convey their data to selected CH node during each round. The CH node aggregates the received data to remove duplicate number of packets so that the it can be communicated to static or mobile sink node in less time with higher transmission speed when the distance between them is minimal. This suggested algorithm guarantees uniform distribution of CH in the sensing field to preserve the energy of sensor nodes.

PROBLEM STATEMENT

WSNs is a type of an ad-hoc networks that involves the large number of small, lightweight sensor nodes. In WSN, the energy of nodes is a critical issue because of constrained battery life. As such networks are constructed for bring together the large amount of data from the millions of sensors. Therefore, they suffer from problems such as higher energy depletion, traffic load balancing, data redundancy and limited wireless communication range. Moreover, there is no proper clustering technique which can find variable sized clusters, an optimal number of clusters, shortest multi-hop path and not able to minimize intra and inter-cluster distance. Even, there is no efficient method for electing cluster head of the cluster. So, we have worked towards the optimization of clustering technique to address above issues.

List of Figures

List of Abbreviations

BS	Base Station
CH	Cluster Head
CC	Cluster Centroid
EMGM	Expectation Maximization Gaussian Mixture Model
DBSCAN	Density-based spatial clustering of applications with noise
ECDEN	Energy-efficient Clustering Based on Density

List of Publications

Accepted

- R.K.Yadav and Joginder Singh “Energy Efficient Clustering Based Data Gathering Using Hybrid DB-EMGM In Distributed Sensor Networks”, IEEE International Conference on Computing, Communication and Automation (ICCCA), May 2017

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Wireless Sensor Networks

Wireless Sensor Networks (WSN) can be perceived as the most prominent emerging technology because of its increasing usage in the spacious range of inherent disciplines including health monitoring, environment monitoring, military applications, habitat monitoring and video surveillance. Moreover, such networks are able to operate in tough environment conditions in which monitoring and human admittance cannot be simply made or may not be worthwhile at all. One of the main unique peculiarities of WSN is that they have wireless connectivity among the small size sensor nodes which makes them work even in rigid conditions with insufficient human support. Also, many sensor nodes are stationed randomly in the monitored region with the help of numerous means, and they form an ad hoc network with wireless links connectivity for sensing entire area. The large numbers of deployed sensors are mostly energy constrained and have a short duration of battery life. Therefore, efficient data gathering protocols and techno-scientific energy-aware routing should be used to enhance network lifetime of sensor nodes in such environments.

WSNs is a type of an ad-hoc network which consists of a vast number of small, lightweight sensor nodes. All nodes are stationed in desired locations for sensing the environments in variety of zones such as in urban cities, schools, buildings, hospitals and mountains. They can sense and gather the mixed type of data frequently such as pressure, humidity, temperature, speed, location and they must handover the sensed data to the base station for conducting the remote tasks. All sensor nodes have the capability of sensing, computing, routing and communicating with other nodes or with the base station. The sink node or base station is supposed to be either static or mobile for conveying the assembled data to the internet or an existing communication infrastructure where the user can access sensed data for specific operations. All sensor nodes have a small size, low energy, limited storage, inadequate memory and restricted communication capabilities. Hence, such nodes produce the immense amount of information in spacious dimensions such as audio, video, images, test etc. Here, the energy of nodes is a critical issue in WSN because of insufficient battery life. There are some of the challenging issues in WSN like energy consumption, traffic load balancing, data aggregation, data redundancy, restricted wireless communication range. Usually, sensor networks are a primary source of generating big amounts of data about the real world in real time.

Therefore, a new work is required to use the well-organized data aggregation approaches for accumulating valuable sensed data received from all sensor nodes to lessen the energy loss. Formerly, an extensive work was performed on clustering techniques which isolate WSN network into diverse clusters to lengthen the life span of network. Hence, clustering is a standard approach for reducing energy consumption and traffic load to address these challenges. Grouping similar nodes produce different clusters as the single group. Each cluster or group has a cluster head (CH) which takes the relevant data from its member nodes and then forwards to the base station by aggregating to eliminate redundancy. The characteristics of Wireless Sensor Networks are the limited power of networks, handle with node failures, heterogeneity of nodes, deployment over the large scale, nodes mobility, communication failures, dynamic network topology, unreliable wireless communication, self-configuration of sensor nodes. The architecture of the wireless sensor networks is shown in figure 1.

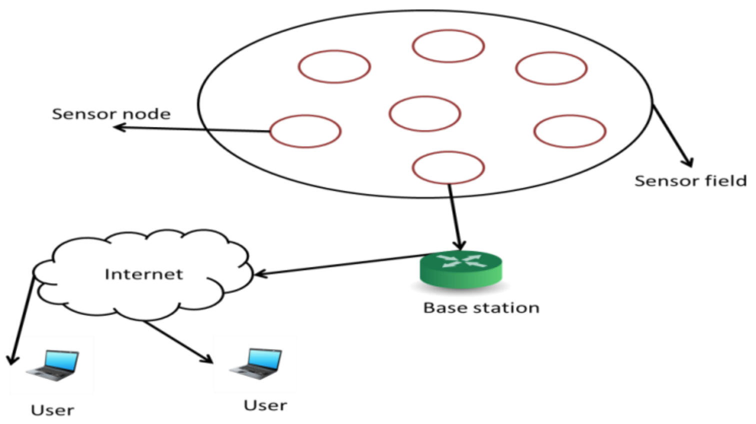

Fig.1 Architecture of the Sensor Network

1.2 Design Challenges in WSN

WSNs pose several challenges related to design and deployment of sensor nodes in the monitored territory. As such networks have constrained potentials such as battery power, radio transmission range, memory and processing hardware of the sensor nodes which makes them less robust in case of node failures, inefficient, poor scalability. Moreover, lengthening network lifetime as one of the main restraint in such systems cannot be easily met by existing traditional techniques. Therefore, it is desirable to use alternative methods to solve critical issues. So, the primary focus is towards designing of new energy efficient clustering algorithms for WSNs.

These are the following design challenges:

- **Heterogeneity**: such networks consist of different types of sensor nodes like pressure, temperature, sound.
- **Network Lifetime Maximization**: all nodes have limited power, so current research is going on to increase the lifetime of overall sensor networks.
- **Energy Efficient Routing**: There is need to efficient routing techniques to transmit aggregated data to the base station in the short interval to time.
- **Cluster Formation**: Generate an optimal number of clusters and use the best clustering algorithm to select Cluster Head (CH) so that wireless data transmission messages can be decreased along with the independent growth of the network.
- **Data Aggregation**: it optimizes the energy consumption and lessens the number of data messages exchanged between nodes and base station.
- **Robustness**: networks must be working if a node failure occurs.
- **Scalability**: a lifetime of the network should not decrease faster even in the case of increasing size of networks.
- **Low power consumption**: Efficient load balancing depends upon on good set of clusters that can preserve the energy of nodes.
- **Secure communication**: there should be a secure way of exchanging the packets among the nodes when used for military applications.
- **Application dependency**: Designed clustering and routing protocols for WSNs should be able to conform to evolving requirements of applications.

1.3 Applications of WSNs

In the current scenario, Wireless sensor network plays a significant role in today's systems which are mainly centering on automating the computing needs of the world. These networks can work better in an environment which is abandoned and remote terrain. As we know that WSN networks has become a hot area of research due to its potential use of small sensor nodes that unite physical world to virtual world. Such large-scale WSNs have high usage in the domains including medical diagnosis, environmental monitoring, home security, military operations, industrial machine monitoring and surveillance as shown in figure 2.

Following are the various applications in WSN:

- **Environmental/Habitat monitoring**

  The monitoring environmental parameters share challenges of harsh environments like

  - Air pollution monitoring
  - Forest fire detection
  - Landslide detection
  - Water quality monitoring
  - Natural disaster prevention
  - Military surveillance

  It includes deployment of multiple networked sensors (acoustic, seismic, video) disseminated across an area such as a battlefield. A user-defined application is designed on top of this network to provide useful information about battlefield to commander sitting at the remote end. It can detect an ongoing event occurring in the monitored territory like an enemy intrusion, firing, etc. It ensures that quality of service should be maintained to elongate network lifetime of the system so that such systems can work for years without maintenance.

- **Medical monitoring**

  There are different types of sensor nodes (implanted, wearable, and environment-embedded) used for aligning the orientation of the body, locations of persons, overall monitoring of ill patients in hospitals and at homes. These nodes can measure the physical state of a person for endless health diagnostics. WSN is regarded as Body Area Networks in medical fields which can gather information about an individual's health, fitness, and energy expenditure and can allow treatment of patients remotely. This type of application requires efficient, reliable, scalable and secure networking protocols.

- **Acoustic detection**
- **Seismic Detection**
- **Inventory Tracking**
- **Smart Grid and Energy Control Systems**
- **Security and Monitoring Smart Buildings**

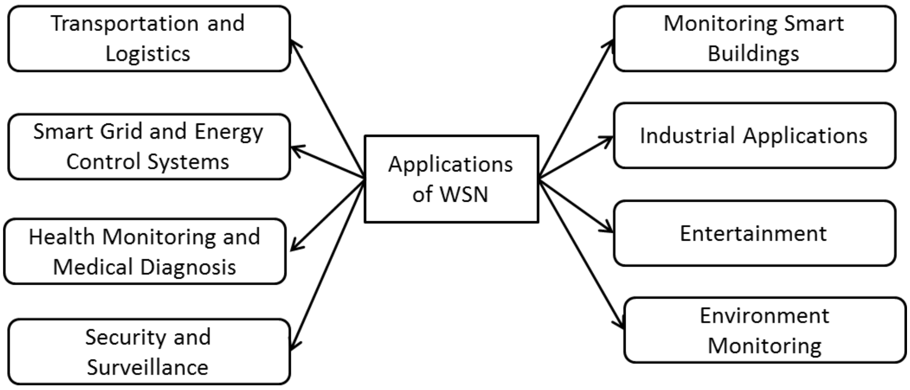

Fig. 2 Applications of Wireless Sensor Networks

1.4 What is Clustering

Clustering provisions network scalability, resource sharing, efficient use of constrained resources and lengthens network lifetime of wireless sensor networks. Clustering is used to group similar nodes into disjoint sets designated as the cluster. Moreover, this technique can be employed in various disciplines such as pattern recognition, machine learning, information retrieval and image processing. Normally, there are numerous clustering schemes which contribute high energy efficiency, efficient resource allocation, and fewer communication overheads thus decreasing energy consumption. Moreover, a significant number of clusters lead to congestion in a network with too many cluster heads and a minimal number of clusters lead to less cluster head handling most of the messages from all group members. Therefore, the designing of efficient clustering technique in wireless sensor network is required to achieve higher energy performance and extend the lifetime of the network in densely deployed WSNs.

One of the most traditional clustering procedure is hierarchical network structure in which cluster formation process points to a two-level hierarchy in which each cluster has a leader termed as Cluster Head (CH) at the top level that conducts data gathering and aggregation responsibilities. At a lower level, there are cluster members who regularly dispatch their sensed information to the corresponding cluster head node. The CH node receives the data from its members and aggregates the data to reduce data redundancy and then communicate them to the base station (BS) either directly or through multi-hop communication. However, All CH nodes convey all the aggregated data over a long distance and thus spend energy at higher rates. So, a standard solution to adjust the power loss among all the sensor nodes is the cyclic rotation of CH role over time in each cluster. Here, BS is data processing node which can be static or mobile in nature which accumulates the data from all CH nodes. Hence, numerous types of clustering techniques are used to isolate the whole WSNs network into distinct clusters to elongate network lifetime of WSN as shown in Figure 3.

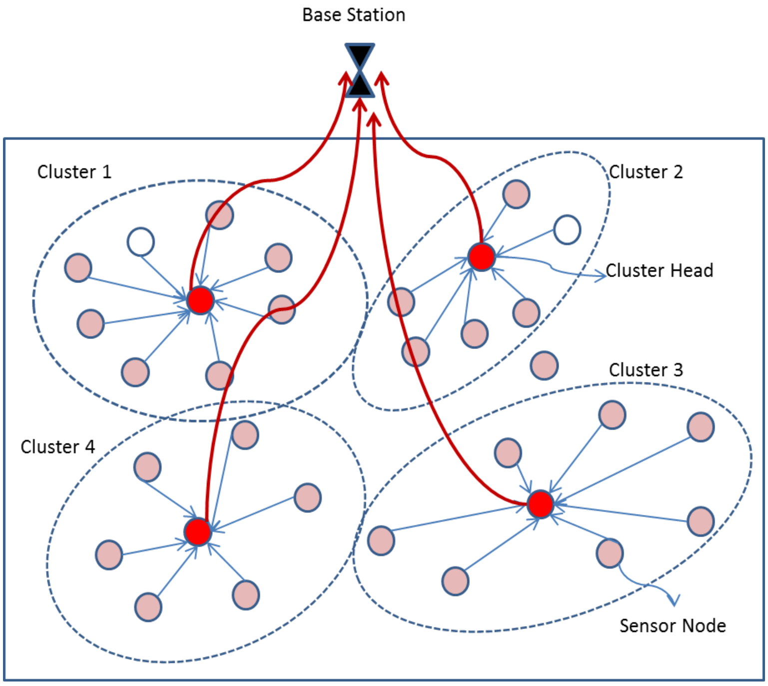

Fig. 3 Clustering in WSNs

1.5 Classification of Clustering Algorithms

1.5.1 Clustering Parameters

Before classifying clustering algorithms, it is worth to mention about the significant clustering parameters required to form clusters in WSNs network. These are the key parameters required for designing of clustering algorithms:

A number of clusters: Currently, most of the existing algorithms locate the diverse number of clusters. So, research is going on to discover an optimal number of clusters which points to higher efficiency routing protocol.

Intra and Inter-cluster communication: it is the distance between cluster head and its member nodes as well as between cluster heads. The focus is on to lessen this distance to elongate network lifetime.

Mobile CH Node: if CH node and its member nodes are stationary then, it provides stable clusters with predefined estimated distance. Whereas, mobile member nodes create dynamic group frequently over time.

Cluster-head selection: CH nodes are pre-assigned in the case of heterogeneous networks as well as in homogenous networks; CH nodes are elected based on defined parameters such as residual energy, connectivity, etc.

Density based: multiple clustering algorithms generate groups as areas of high density and low density by analyzing closeness of nodes stationed in the area.

1.5.2 Types of Clustering Algorithms

In light of above clusters parameters, various types of available clustering techniques are diverse approaches in the literature. There are different ways to classify such algorithms such as

1. Homogeneous or Heterogeneous clustering algorithms
2. Clustering algorithms related to centralized or distributed in nature

Apart from this, a number of clustering algorithms can be arranged as partitioning, hierarchical, model-based, grid-based and density-based methods as shown in figure 4.

Partitioning algorithms use an iterative way to deal with isolating the dataset into various diverse clusters. K-means [1] and k-mediod [2] is considered as the most famous partition-based clustering algorithm, but the issue is that number of clusters, K should be known a priori, and only circularly shaped clusters can be resolved due to convergence towards the cluster focus.

Hierarchical algorithms construct a hierarchical collection of clusters and relate nodes to form clusters based on their distance.

There are two types of Strategies for this clustering

Agglomerative or bottom-up approach: it starts with each object as one cluster and then merges the clusters while moving up the hierarchy until all the clusters are combined to form a single big cluster.

Divisive or top-down approach: it starts with all the objects considered as a single cluster, and several splits are performed recursively while moving down the hierarchy until each object is in one cluster itself. The issue of this algorithm is that they are computationally costly and only circularly shaped clusters can be acquired. The samples of this hierarchical clustering method are LEACH [3], HEED [4], BIRCH [5] and CURE [6].

Density-based algorithms like DBSCAN [7] and OPTICS [8] find the clusters in view of the density of nodes stationed in a region. It finds core points at first and grows clusters set by searching for a point within a radius of a given point. The good of these algorithms is that they can subjectively discover arbitrarily shaped clusters and can determine noisy nodes.

Grid-based algorithms split the monitored boundary into a limited number of cells known as grids and after that carry out the required operations on the grids. The advantage of this approach is its fast processing time. Some of the conventional grid-based algorithms are STING [9], Wave-Cluster [10], and CLIQUE [11].

Model-based algorithms attempt to discover best inexact model parameters that best fit the given data. This algorithm follows Gaussian distribution model that corresponds to distinct clusters in the data set. Initially, fixed number of clusters determined a priori is expected during execution. A standout amongst the most popular model-based clustering methods is EM [23].

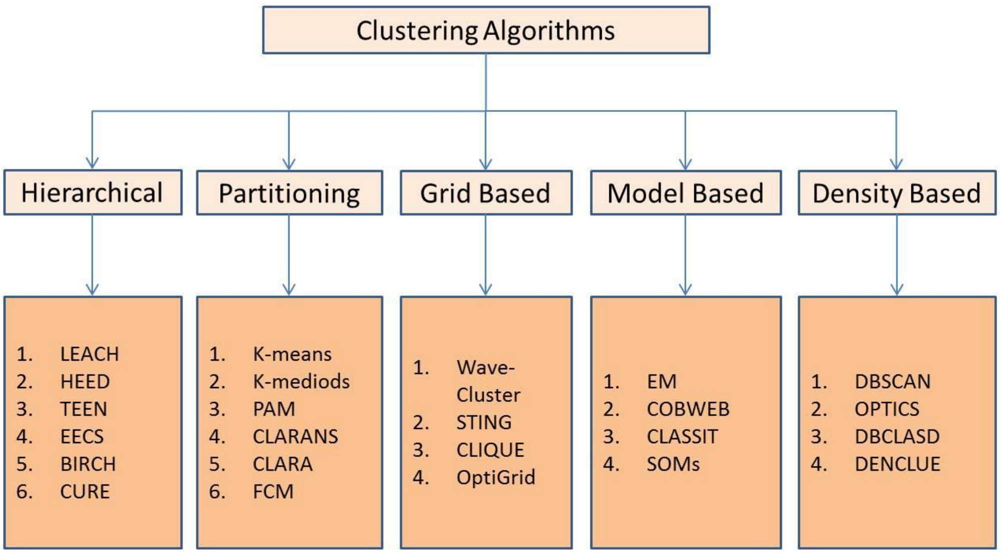

Fig. 4 Types of Clustering Algorithms

CHAPTER 2: LITERATURE REVIEW

In this chapter, we have conducted the literature survey on the several available clustering techniques recommended so far to meet the design challenges of WSNs. Currently, one of the most extensively used first hierarchical cluster-based protocols was Low Energy Adaptive Clustering Hierarchy (LEACH) that was proposed by Heinzelman et al., [3]. LEACH protocol is a distributed clustering technique for homogeneous WSNs networks that considers all nodes with same initial energy. It uses a probabilistic method to elect the cluster head (CH). Each non-CH node determines its cluster by choosing minimum distance CH node. Here, there is a periodic rotation of CH node in every round to balance the energy load among all nodes. The rotation of CH role is based on selecting a random number T(n) between 0 and 1. The node that became CH recently cannot be CH again in next round. In LEACH protocol, CH receives and aggregates relevant sensed data from it member nodes as per their TDMA slot and export aggregated data to the sink node. CH may be situated far from the BS, so it utilizes the vast amount of its power for transmission, and because it is continuously on, it will expire then quicker than other nodes. The weakness of LEACH protocol is that it cannot find a suitable number of clusters and performs CH selection based on a random number without considering the initial or residual energy of nodes. There is an irregular arrangement of clusters due to varying position and number of CH in the network.

Several authors extended LEACH protocol to improve its performance by considering various parameters for CH selection to avoid randomness such as LEACH-C [12], E-LEACH [13], TL-LEACH [14], A-LEACH [15]. As far as improvement in LEACH is concerned, Heinzelman et al. proposed LEACH-C [4] to overcome the problems of LEACH by shifting cluster formation phase and CH selection of nodes to Base Station. In this protocol, all nodes convey their information (energy, location) to BS. The BS computes the average energy of network to select CH node those having residual energy more than average energy. Hence, LEACH-C enhances the lifetime of the network as compared to LEACH.

As WSNs are heterogeneous conscious networks that consider all nodes with different initial energy to extend stability period region as compared to LEACH protocol. So, the author in [16] proposed Stable Election Protocol (SEP). It utilizes the remaining energy of each node for CH selection which is based on weighted election probabilities of each node. In this approach, there are two categorizes of nodes as advanced nodes having more energy as compared to the normal nodes. So, such advanced nodes have more probability to become CH node.

Authors in [4], Younis and Fahmy proposed the hybrid energy-efficient distributed (HEED) clustering protocol which is intended for multi-hop networks. There are two levels used for CH selection. First one is based on the residual energy of node and second depends on node degree, distance to neighbours and least intra-cluster power dissipation. The results show that HEED is far better than LEACH for lengthening the lifetime of the network. But, it employs more overhead communication via iterative approach. A novel clustering technique by Ye et al. in [17] for single hop communication introduced a new cluster formation phase based on distance from the node to CH and the distance from the CH the BS. The CH is selected based on higher residual energy between its all nodes. It evenly scatters load among all CHS and hence providing a longer life of WSN and also demonstrates that it is 32% more energy efficient than LEACH.

A one more homogeneous network protocol PEGASIS proposed by S. Lindsey at al. [18] uses chain based cluster approach for cluster formation and select one node as CH for routing aggregated data to the sink node. It applies greedy approach for chain formation and is centralized clustering algorithm

Apart from homogeneous and heterogeneous clustering approach, clustering is also performed on the basis of the density of deployed sensor nodes in the area to be monitored. This class of clustering algorithm comes under the density-based algorithms.

Density-based Spatial Clustering of Application with noise (DBSCAN) Algorithm suggested by Ester et al. in 1996 [7], was the first approach which considers density as a parameter for clustering. It requires two input parameters, first is the radius of clusters (Eps) and second is the minimum number of points (MinPts) to build the cluster. But, it is hard to guess these parameters initially. So, estimating these parameters is major shortcomings of this algorithm. DBSCAN beats both partition and hierarchical clustering algorithms as it can find the arbitrary shape of clusters as well as identify noisy points. It can create clusters as high point low point density. However, it cannot handle data containing varying densities for large dataset. The quality of clusters in DBSCAN depends on a good set of parameters.

OPTICS [8] is an enhanced version of DBSCAN that is proposed to handle data with varying density. It maintains ordered set of clusters in the data.

Currently, the research is going on to design a new approach for clustering which can eliminate the drawbacks of all types of clustering algorithms. Many researchers have developed combined clustering algorithms with meta-heuristic and nature-inspired techniques to enhance the lifetime, stability period of networks. Some of the hybrid algorithms described in the literature as follows. Soft-DBSCAN [19] introduced by Abir and Eloudi improves DBSCAN by linking with fuzzy set theory (FCM algorithm). First DBSCAN is executed to produce the arbitrary shape of the cluster and then calculates the degrees of fuzzy membership to represent the data near to cluster centroids. This method does not only outperform Fuzzy C-means clustering but also able to create more dense clusters.

DBCSVM [20] suggested by Santosh Kumar Rai and Nishchol Mishra is a Density-Based Clustering Using Support Vector Machines. This method applied for feature extraction and calculation of description of a feature from a large dataset containing features. The previous DBSCAN takes more time for clustering process and does not produce a better result. However, the proposed method is very efficient for the separation of farther and nearer points because farther points help in clustering. Hence, it takes less time as compared to DBSCAN.

Smiti and Elouedi [21] merge DBSCAN and Gaussian-Means (GM) algorithm to estimate the input parameters of DBSCAN. However, GM produces circularly shaped cluster, not nased on density, and it is not effective against noise. Still, there is a requirement to know the inout parameter required for the Gaussian distribution. Such current algorithms have their individual shortcomings which lead to a bad clustering. D. Raja Kishor and Venkateswarlu [22] offered hybridization of Expectation-Maximization (EM) and K-Means methods to improve the speed of the clustering process. In this method, Clustering fitness and Sum of Squared Errors (SSE) are determined to measure the clustering performance. It takes less execution time with tolerable Clustering Fitness value and less SSE than the conventional EM algorithm and hence yields better clustering results.

Hence, after surveying about available clustering algorithm, it comes to know that there is no proper research work to eliminate the shortcomings of existing algorithms. Therefore, I have worked towards optimization of better clustering algorithms by joining model and density based algorithms. This optimized developed method will assist to enhance the network lifetime, scalability, stability period, low power loss, faster transmission data rate along with finding out an optimal number of clusters so that there is a uniform arrangement of CHs in WSNs to adjust the energy load among all sensor nodes..

CHAPTER 3: CLUSTERING ALGORITHMS

3.1 Low-Energy Adaptive Clustering Hierarchy (LEACH)

Currently, one of the most widely used first hierarchical cluster-based protocols was Low Energy Adaptive Clustering Hierarchy (LEACH) that was proposed by Heinzelman et al., [3]. LEACH protocol is a distributed clustering technique for homogeneous WSNs networks that considers all nodes with same initial energy. One of the main focus of this protocol is a random selection of CH node out of the available node. The process of LEACH protocol is performed in several rounds in which each round can be into two steps: the set-up phase and the steady state phase as shown in figure 2. In this protocol, cluster formation, CH selection and broadcasting of TDMA schedule by CH node to all member nodes are performed.

CH selection: this is followed by generating a random number between 0 and 1. If this number is less than the threshold value T(n) then that node becomes CH. T(n) is calculated using equation 1.

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

Where P denotes the desired percentage of sensor nodes to become CHs among all sensor nodes. r implies the current round number and G is the set of sensor nodes that have not participated in CH election in previous 1/P rounds. There is a periodic rotation of CH node in every round for consistent spreading of energy load. After CH node selection, it disseminates a message to all other nodes. Upon receiving this message, nodes decide to select its own CH based on minimum distance. After cluster formation, each CH creates a TDMA schedule and transmits these schedules to their members within the cluster to avoid the collision of data sent by member nodes. The set-up phase is completed if every sensor node knows its TDMA schedule. In the steady state phase, TDMA schedule is responsible for carrying sensed data from member nodes to the CH and CH to the BS, and hence, this concept of LEACH reduces intra-cluster collision among member nodes as shown in figure 5.

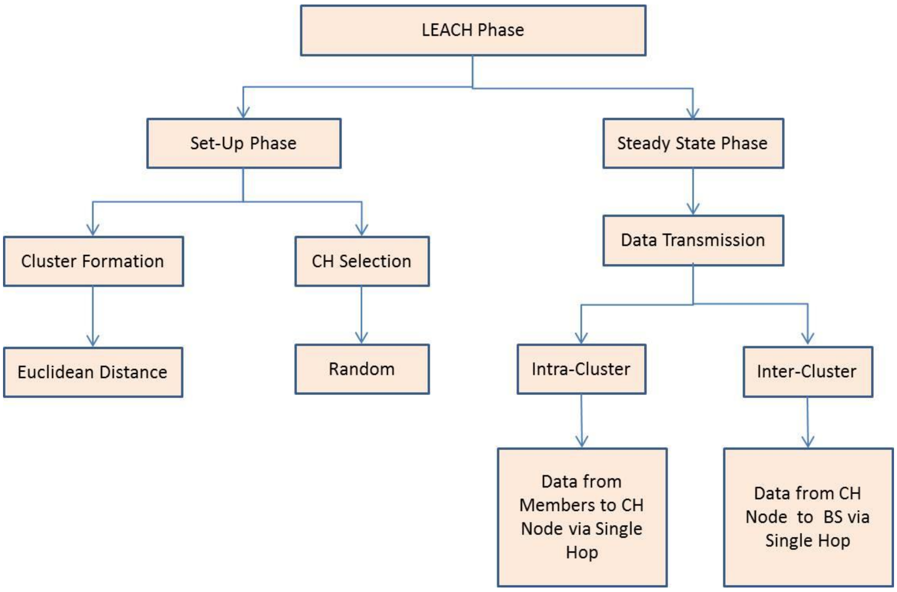

Fig. 5 Operation of LEACH Protocol

Benefits of LEACH:

LEACH is a distributed and hierarchical routing protocol for homogenous networks. These are some of the main features of LEACH are as follows:

- Exchange of less number of packets between sensor nodes and BS which prolongs the network lifetime.
- Data aggregation is performed by CH that reduces the number of data packets transmitted to BS.
- The Intra and Inter-Cluster collisions are avoided to enhance the life of sensor nodes.
- The CH selection is random in every round that leads to energy balance in the network.

Drawbacks of LEACH:

There are several problems of LEACH are as follows:

- There is no efficient CH selection procedure because CH is chosen out of high energy or less energy node randomly. Moreover, it does not consider the residual energy of node as a parameter for CH selection that leads to irregular dispersion of CH across the whole network.

- LEACH does not produce exact position and number of CHS in each round due to randomness nature which leads to irregular dispersion of clusters in the network. Hence, it causes tremendous loss of energy within the cluster because of more intra-cluster distance and thus decreases the efficiency of the network.
- The performance of network degrades when CH is far away from the BS, then it spends more energy compared to CHs that are near to the BS. It follows single hop communication.

3.2 Power-Efficient Gathering in Sensor Information Systems (PEGASIS)

PEGASIS) [18] is one of the centralized clustering algorithms that is a near optimal chain-based protocol for data gathering in WSNs. It mainly focus on building the chain of different clusters by applying the greedy approach so that valuable data can be dispatched to the nearest neighbour node. This method operates better than LEACH by decreasing overhead of dynamic cluster formation and CH selection. In this, node take turns to dispatch the aggregated data to the BS to overcome energy depletion of the network.

3.3 Hybrid Energy Efficient Distributed Clustering (HEED)

HEED [4] clustering protocol is principally intended for multi-hop networks. There are two parameters used for CH selection. First one is based on the residual energy of node and second depends on node degree, distance to neighbours and least intra-cluster power loss. The results reveal that HEED is far better than LEACH for lengthening the lifetime of the network. But, it consumes much overhead communication via iterative approach.

However, this algorithm is suffering from the energy hot-spot problem because of unbalanced energy loss. It sometimes generates more number of clusters than expected.

CH Selection: CH node is selected based on the following parameters as

$$\text{CHprob} = \text{CHprob} * \frac{\text{Eresidual}}{\text{Emax}} \tag{2}$$

Where CHprob the probability of the becoming a cluster head, Eresidual is the current residual energy of the node and Emax is the initial maximum energy of node.

3.4 Expectation-Maximization Gaussian Mixture (EMGM)

The EMGM [23] applies the concept of model-based clustering approach that tries to find out the best model on the input data. This is used to creates the diverse cluster sets by applying Gaussian mixture model distribution function on the given data.

$$P(x) = \sum_{k=1}^{K} \pi_k \; N(x|\mu_k, \Sigma_k) \tag{3}$$

Where K signifies the total number of clusters and π_k is mixing coefficients for each cluster.

$$N(x|\mu, \Sigma) = \frac{1}{(2\pi)\sqrt{|\Sigma|}} \exp\left(-\tfrac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right) \tag{4}$$

EMGM tries to maximize the log-likelihood function for parameters (Means, μ_k, Covariance, Σ_k and Mixing Coefficients, π_k).

The steps of EMGM are described as follows:

1. Set the initial values for μ_j, Σ_j and π_j and log likelihood (LL)

2. E step: computes the responsibilities among the nodes for each cluster via equation below

$$\gamma_k(x) = \frac{\pi_k \, N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \; N(x|\mu_j, \Sigma_j)} \tag{5}$$

3. M step: - Updates computed parameters via E step values.

4. Estimates the new value for log-likelihood using below equation

$$\ln P(x|\mu, \Sigma, \pi) = \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k \; N(x_n|\mu_k, \Sigma_k)\right) \tag{6}$$

If there is no convergence, then, go back to step 2

This algorithm cannot obtain an optimal number of clusters and also cannot find out the noisy data existing in the test dataset.

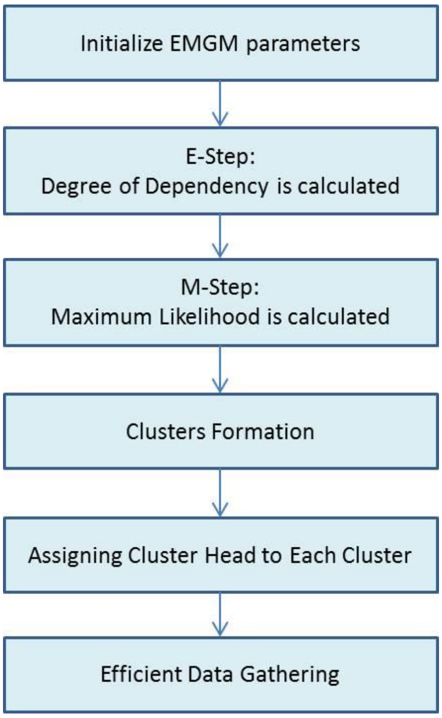

Fig. 6 Steps of EMGM Algorithm

3.5 Density-based spatial clustering of applications with noise (DBSCAN)

DBSCAN is a density-based clustering algorithm that determines clusters as low-point and high-point density zones along with the detection of an arbitrary shaped cluster set as well as has the ability to locate noisy data. It expects two input parameters, first is the radius of clusters (Eps) and second is the minimum number of points (MinPts) to build the cluster. But, it is hard to guess these parameters initially. So, estimating these parameters is major drawbacks of this algorithm. DBSCAN beats both partition and hierarchical clustering algorithms as it can obtain the arbitrary shape of clusters as well as recognize noisy points. However, it cannot handle data containing varying densities for large dataset. The quality of clusters in DBSCAN depends on a good set of parameters. The process of DBSCAN clustering is shown in figure 7.

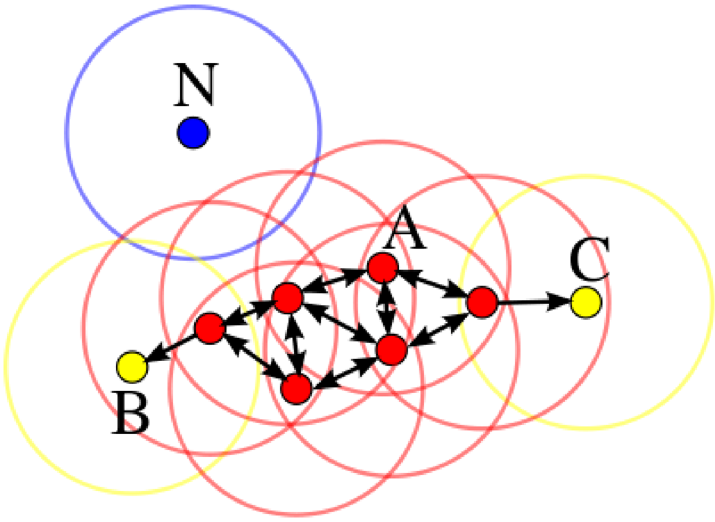

Fig. 7 Clusters Process in DBSCAN

The pseudocode of DBSCAN clustering algorithm as follows:
Function DBSCAN (Dataset D, Eps, MinPts) 1. Choose any point P in D 2. Obtain all points from P which are within Eps and have count of at least MinPts 3. **if** P is found to be as core point, then assign the all points in same cluster 4. **if** P is found to be as border object, then DBSCAN picks up the next point from D 5. **else** mark P as noise object 6. Repeat from Step 1 **End**

CHAPTER 4: PROPOSED WORK ON CLUSTERING

In this chapter, I have described two new improved clustering algorithms for the purpose of dynamic clustering analysis of data sets to offer high energy efficiency and reduced communication overheads. The first algorithm known as hybrid DB-EMGM will solve the problem of finding the optimal number of clusters and arbitrarily shaped clusters set produced. It also presents an idea about how mobile sink node reduces wireless data transmissions and data redundancy. The second algorithm gives the idea for selecting cluster head for each group based on density of nodes to produce better clustering results. Therefore, the designing of efficient clustering technique in wireless sensor network is required to achieve high power efficiency to extend the lifetime of the network in densely deployed WSNs.

4.1 First Proposed Clustering Algorithm

In this section, we will describe a new algorithm known as hybrid DB-EMGM that unites both EMGM and DBSCAN clustering methods for finding an optimal number of clusters and arbitrarily shaped clusters set along with input parameters estimation for DBSCAN. It also presents an idea about how mobile sink node discovers the shortest path over cluster centroid location and thus reduces wireless data transmissions and data redundancy.

4.1.1 Problem Statement

As discussed in introduction part, it comes to know that clustering is an appropriate method for data gathering in distributed wireless sensor networks. Although existing clustering algorithms cannot find variable sized clusters, an optimal number of clusters, multi-hop path and not able to minimize intra and inter-cluster distance. Hence, to solve such issues, an improved clustering method is proposed that aim to find the optimal set of clusters of arbitrary shape along with their cluster centroid out of the given data set as input. Moreover, it is assumed that sink node is considered as mobile which will help to reduce wireless data transmissions and multi-hop path communication among all nodes. The shortest path of mobile sink node is decided based on the location of cluster centroids and then sink node moves around the sensing zones.

4.1.2 Proposed Hybrid Algorithm (DB-EMGM)

We will explained our suggested clustering algorithm termed as hybrid DB-EMGM that unites the both EMGM and DBSCAN algorithm. This type of work mainly joins the both model and density based algorithms to partition the distributed type of WSN network into diverse sets of cluster for finding an optimal number of clusters out of given data sets. Moreover, the concept of mobile sink node is utilized that moves all over the place in the sensing zones. The proposed hybrid algorithm transmits the position of cluster centroid to the sink node. When movable sink node comes at the position of the cluster centroid for receving the aggregated data, then all the node does not convey their sensed data over a long communication distance. Hence, data sending and receiving rate is enhanced between the mobile sink node and all sensor nodes and thus saving the energy spent during data transmission in WSN.

We illustrate in a nutshell in relation to EMGM and DBSCAN algorithm that creates the diverse clusters set in WSNs. We suggested our enriched hybrid clustering method for collecting the valuable information extracted from the duplicate number of packets received from all the member nodes. The basic steps of proposed method has been clarified using the flowchart depicted in figure 8.

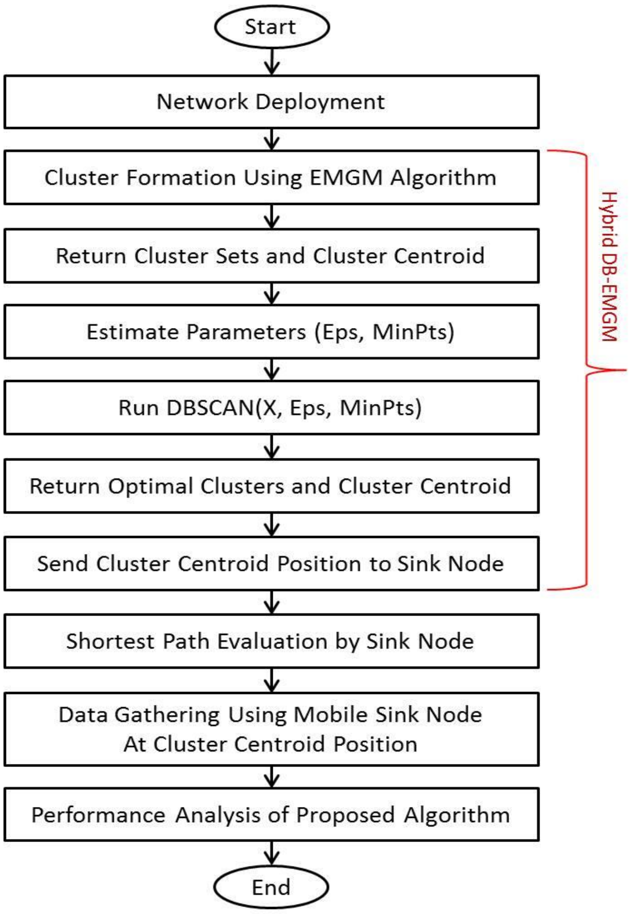

Fig. 8 Flowchart of Proposed Algorithm

4.1.3 Network Deployment

A distributed WSN type of network is considered in which the lots of sensor nodes are stationed in number of zones such as in schools, hospitals, buildings, mountains, urban cities. In this considered model, the sink node is located at the center and mobile in nature. The network is geographically dispersed and partitioned into a number of sub-networks as displayed in figure 9. Once, the mobile sink node comes at the location of cluster centroid by patrolling over the shortest route, then it starts assembling the benefical information from all the member nodes to preserve the transmission energy of node spent during transmission.

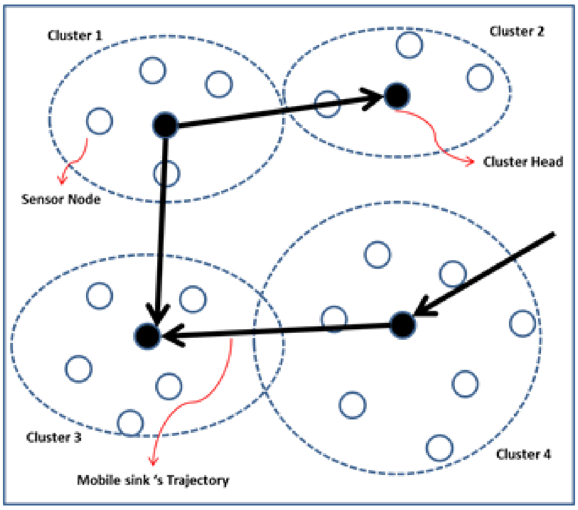

Fig. 9 Deployed WSN Network Model

4.1.4 Overview of Model Based: EMGM Algorithm

The EMGM uses the concept of model-based clustering method that applies Gaussian mixture model to create the diverse sets of clusters by searching the best fit of the model in the input data. Moreover, the nodes deployed in the sensing zones exhibit the distribution according to Gaussian model. There are some advantages of EMGM algorithm such that it is centralized algorithm for finding more efficient clusters than distributed clustering algorithm. Moreover, it can efficiently minimize the distance between every member node and cluster centroid and produces somewhat good set of clusters from the data set. It does not consider density while creating clusters and performs poorly when node density is low. One of the main limitation of EMGM is its inability to find noisy nodes present in the test dataset. Moreover, there is no way for opbtaining the optimal number of clusters. Hence, to solve above problem, this algorithm is united with DBSCAN for considering density condition for producing better set of clusters.

4.1.5 Overview of Density Based: DBSCAN Algorithm

Density-based Spatial Clustering of Application with noise (DBSCAN) is a density-based clustering algorithm which has the ability to generate clusters as the different zones of low-point and high-point density on the input dataset and also can discovers the arbitrary shaped clusters of wide variety along with the considering some of the nodes as noisy. It can find an optimal number of clusters which depend on guessing the good set of values for two input parameters, first is the radius of clusters (Eps) and second is a minimum number of points (MinPts). However, it is hard to guess these parameters initially. So, estimating these parameters is major limitations of this algorithm. DBSCAN beats both partition and hierarchical clustering algorithms. The quality of clusters in DBSCAN depends on a good set of parameters.

Algorithm 1: The Pseudocode of DBSCAN algorithm

```
DBSCAN ( X, Eps, MinPts)
{
        ClustNo = 0   //cluster ID number
        For each unvisited point P in Dataset, X
        ClustMem = Find All points within radius, Eps
        if  ClustMem < MinPts
                label P as noise
        else
        {
         ClustNo = ClustNo +1;   // increase cluster count
        FormBigCluster(P, ClustMem, ClustNo, Eps, MinPts)
        {
          Assign P to ClustNo
          For each unvisited point Z in ClustMem
            {
              ClustMemNew = Find All points within radius, Eps
              if ClustMemNew >= MinPts
                 ClustMem = Merge ClustMem with ClustMemNew
          }
         if Z is not  member of any cluster
              Assign Z to ClustNo
         }
    Return Set of Clusters
 }
```

4.1.6 Estimation of Parameter Eps of DBSCAN

It is difficult to guess the proper value for the parameter Eps of DBSCAN. So, determining the good estimate for the radius of the cluster, Eps is a crucial problem because if we select the smaller value for Eps, then many points will be considered as noisy. Consequently, a large number of small clusters will be produced and leads to more energy consumption. Similarly, if we select a large value of Eps, then it is not possible to find noisy points, and the lesser number of clusters are produced. Hence, we must choose right value of Eps to produce better clustering result as shown in figure 10.

We have proposed a method to select the good value of Eps by doing the analysis of the results obtained from EMGM algorithm.

Find radius, R which is the average of all distances from all member nodes to cluster centroid for each cluster and take the minimum radius, R as the global value for Eps.

$$Eps = \min(R1, R2, R3 \ldots \ldots) \quad (7)$$

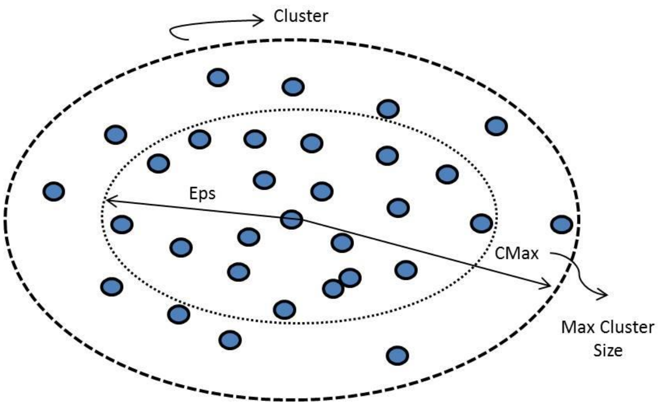

Fig. 10 Estimation of Radius, Eps for Each Cluster

4.1.7 Estimation of Parameter MinPts of DBSCAN

To calculate the minimum number of points required to form the clusters, we proposed a method to estimate the global value for MinPts.

Find the number of points, C within cluster which comes under the radius, R of each cluster and then choose smallest number of points, C as global value of MinPts.

$$MinPts = \min(C1, C2, C3 \ldots \ldots) \quad (8)$$

4.1.8 Data Gathering by Mobile Sink Node

Our proposed hybrid algorithm returns the optimal number of clusters and cluster centroids for each cluster. This algorithm transmits the position of cluster centroids to the sink node along with the residual energy of all member nodes. The sink node decides the shortest path using the location of cluster centroids. When mobile sink node moves all over the network region along the defined path, and then sensor nodes send their sensed data to sink node and hence thereby decreasing wireless data transmissions as well as fewer overheads for relaying data to other nodes. It is assumed that sink node adorned with high energy.

When mobile sink node comes near to cluster centroid of each group, then it broadcasts a data request message to all member nodes of that cluster. All member nodes receive the message and send their data to sink node. Following receipt of the required information from all nodes, sink node performs data aggregation to eliminate the duplicate number of packets. This algorithm increases the transmission and receiving rate to sink node because the aggregated data is not transmitted to sink node over a long distance and hence thereby working far better than existing clustering algorithms such as EMGM, LEACH and SEP.

This new approach for clustering achieves network scalability, high energy efficiency by reducing energy spent in transmitting data packets among all nodes. Hence, the validity of this algorithm is determined through experiments in the results section of this report.

4.1.9 The pseudocode of DB-EMGM Algorithm

Our recommended (DB-EMGM) method tries to diminish the traffic flow of excess data in the network by getting minimized Intra and Inter-cluster distance among the nodes in WSN. This also discovers the the location of cluster centroid using the estimated mean and cluster-covariance by creating diverse sets of cluster and thus save the energy of sensor nodes spent during data transfer.

The pseudocode of proposed Algorithm as follows:
1. Set the initial value of Gaussian components, k by analyzing the data set, X 2. Initialize Randomly the values for (μ, π, ∑) 3. Calculate the Intra and Inter distances among the nodes, *D* 4. Finds responsibility(R) of node and log-likelihood (LL) using (5) 5. **while** |new LL – old LL |< small number, ϵ 6. **For** 1 to n **do** // n is number of nodes 7. Update μ, π, ∑ using R with finding number of nodes of each cluster **8. End for** 9. Estimate new log-likelihood (LL) using (6) **10. End while** 11. Return number of nodes and cluster centroid for each cluster by EMGM 12. Estimate Input Parameters (Eps,MinPts) via **step 11** 13. Execute DBSCAN (X, Eps, MinPts) using **Algorithm 1** 14. Return optimal number of clusters, C and Cluster Centroid Position 15. **For** r = 1 to Round // Round is the maximum round number 16. **For** i = 1 to C // C is the optimal number of clusters 17. Transmits the Cluster Centroid Position to Sink Node 18. Sink Node calculates the distance from all cluster centroids 19. Calculation of the Shortest path using Minimum Spanning Tree 20. Mobile Sink Node visits Cluster Centroid over the shortest path 21. Sink Node Broadcast Data Request Message within Cluster 22. Member Nodes Transmits the sensed data to sink node 23. Data aggregation performed by the sink node **24. End For** **25. End For**

4.2 Second Proposed Clustering Algorithm

In this section, we have explained a new method known as ECDEN that gives the idea for selecting cluster head for each cluster based on the density of deployed sensor nodes in the field. Moreover, it also finds the optimal number of cluster set. For performance evaluation of this algorithm, sink node can be viewed as static or mobile in nature.

4.2.1 Problem Statement

As discussed in introduction part, it comes to know that clustering is a relevant method for data gathering in distributed wireless sensor networks. As we know that current clustering algorithms suffers from problems such as limited communication range, wireless data transmissions due to its distributed nature. Moreover, there is no proper clustering technique for electing cluster head of the clusters. Hence, an improved clustering method is suggested that will solve the problem of finding appropriate cluster head for each cluster in every round based on density using residual energy, average energy and distance so that there will be regular distribution of CH node in the field. It also aim to find the optimal number of cluster set using density concept of stationed sensor nodes in the field and reduce wireless data transmissions by decreasing intra-cluster distance among all nodes.

4.2.2 Proposed Energy-efficient Clustering Based on Density (ECDEN)

We recommend energy-efficient clustering based on Density (ECDEN) that divides the network into diverse cluster set. This developed algorithm mainly focuses on cluster formation and cluster head selection phase in the network field. Moreover, it not only finds the optimal number of clusters in every round but also gives an idea of electing cluster head based on density using residual energy, average energy and distance from the cluster centroid location of each cluster obtained by results of the proposed algorithm. It is assumed that base station is static or mobile in nature for data collection. After cluster formation phase, all member nodes send their data to selected cluster head during each round. The CH node aggregates the received data to drop the duplicate number of packets so that the data can be transmitted to static or mobile sink node in less time with more transmission speed. If the sink node is static, then all CH nodes convey the aggregated data over the distance between them. If the sink node is mobile in nature, then, CH node transmit the aggregated data to the sink node when distance between them is minimal. Here, the mobile sink node follows the shortest path over CH location. So, this suggested algorithm guarantees uniform distribution of CH in the network field and thereby conserving the energy of sensor nodes. Our density-based clustering method is illustrated using the steps shown in figure 11.

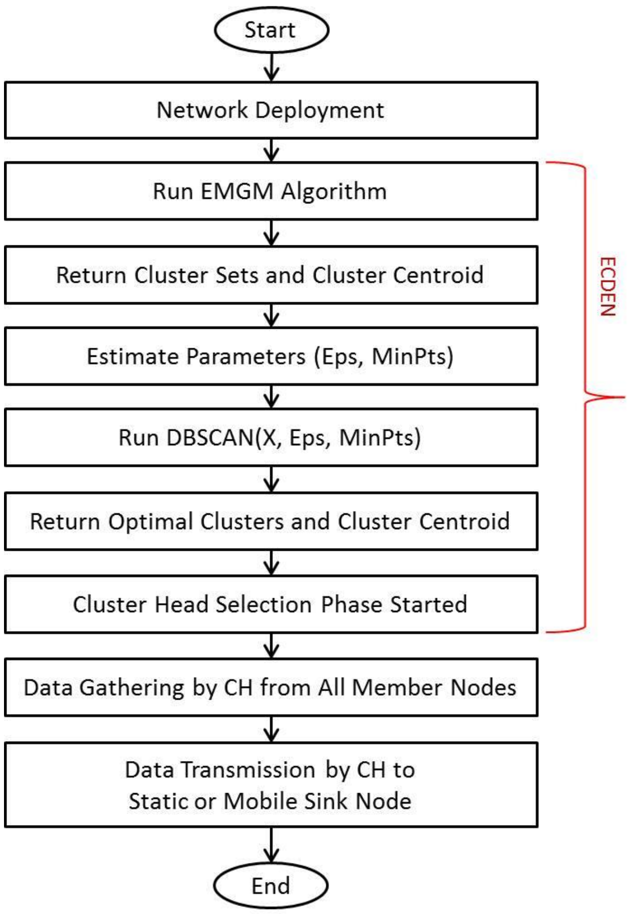

Fig. 11 Flowchart of ECDEN Proposed Method

4.2.3 Network Deployment

We have assumed that WSNs network consists of N sensor nodes that are deployed randomly in M × M region. All nodes observe the surrounding environment and communicate the data to the base station which is placed far away from these nodes. It is assumed that base station can be static or mobile in nature. The network is geographically dispersed and divided into various sub-networks as shown in Figure 12. After cluster formation phase, the sensed data is transmitted to cluster head. Later, the aggregated data is communicated to the remote base station while maintaining the energy of nodes.

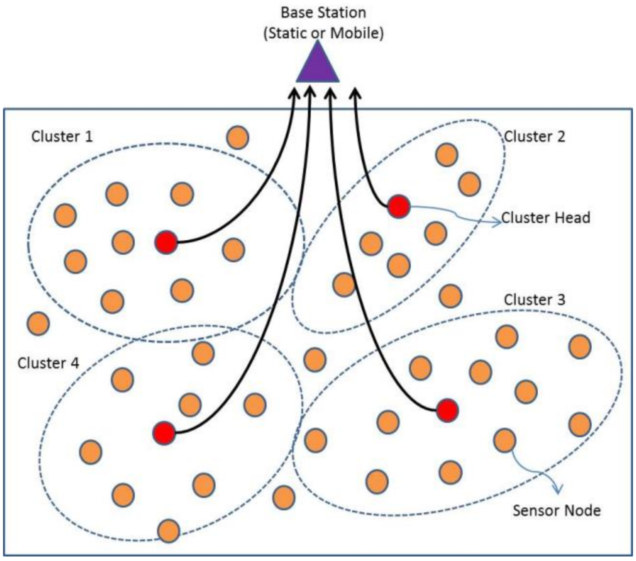

Fig. 12 Deployed Network Model

4.2.4 First Order Radio Energy Model

We assess our approach using the first order radio energy model [3] and the parameters are listed in TABLE II. As shown in figure 13, the energy spent by the transmit amplifier $\mathrm{E_{Tx}(m, d)}$ to transmit an m-bit packet between a transmitter and receiver along the distance, d is

$$\mathrm{E_{Tx}(m, d)} = \begin{cases} \mathrm{m * E_{elec} + m * E_{fs} * d^2 \ : \ if\, d < d_o} \\ \mathrm{m * E_{elec} + m * E_{mp} * d^4 \ : \ if\, d \geq \ d_o} \end{cases} \quad (9)$$

Where $\mathrm{d_o} = \sqrt{\mathrm{E_{fs}/E_{mp}}}$ signifies the threshold distance, $\mathrm{E_{elec}}$ signifies the energy loss in the electronics for forwarding or receiving one bit, and $\mathrm{E_{fs} * d^2}$and $\mathrm{E_{mp} * d^4}$ signifies amplifier energy loss for free space and multipath fading channel, respectively.

To obtain an m-bit message, the energy $\mathrm{E_{Rx}(m)}$ wasted by the receiver is given by

$$E_{Rx}(m) = \ m * E_{elec} \quad (10)$$

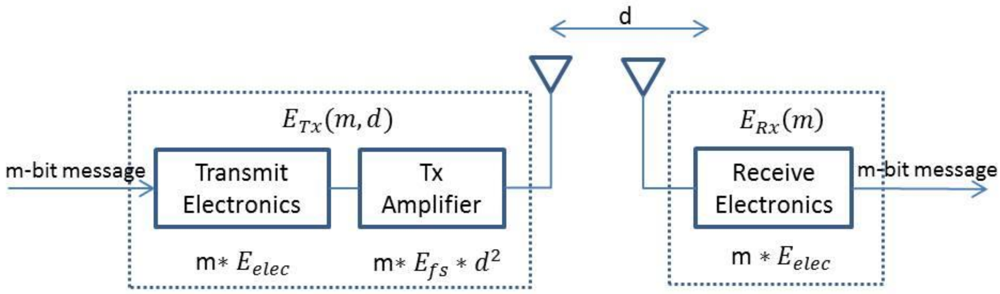

Fig. 13 First Order Radio Energy Model

4.2.5 Cluster Head Selection Phase

After the cluster formation phase, the proposed clustering algorithm (ECDEN) returns the optimal number of cluster and cluster centroid. Now, cluster head selection phase is started to select original cluster head of each cluster. The each cluster head (CH) is picked on the basis of the distance from the cluster centroid. The sensor node nearest to the cluster centroid with its residual energy larger than or equal to the average energy of the cluster is accepted as the cluster head for current round. If two or more sensor nodes are at a same distance from the cluster centroid, then anyone of them can be selected randomly by comparing residual energy with the average energy of the cluster. This recommended method for cluster head selection maintains the balanced distribution of CH node in the network and thus balance the energy among the nodes spent during data transmission phase. Once CH node is elected in the current round, then it will notify all the member nodes within the cluster that it is decided as the leader for collecting the data. In this case, CH node disseminated the HELLO message within the cluster by computing broadcasting range calculated by function $BRmax$.

$$BRmax = Average\ of\ Sum\ of\ all\ distances\ from\ member\ nodes\ to\ cluster\ Head$$

The steps of cluster head selection phase are shown in the flowchart in figure 14.

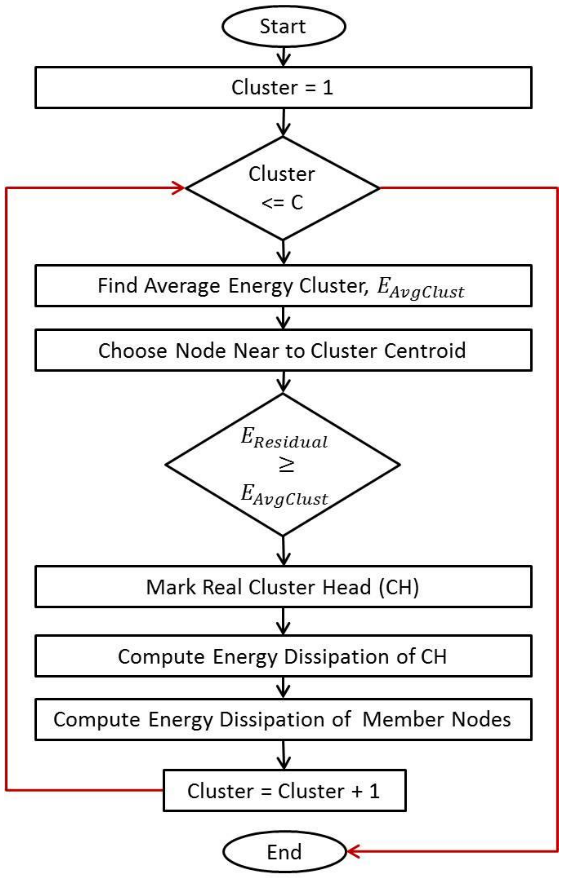

Fig. 14 Flowchart of Cluster Head Selection Phase

4.2.6 Data Gathering Phase

The suggested density based clustering algorithm (ECDEN) conveys the cluster head information (position, energy) to the sink node. If the sink node is static, then all member nodes forward the sensed data to their CH node. After collecting the information, All CHs starts aggregating the data to lessen the duplicate number of packets to eliminate data redundancy problem. The CH node conveys the valuable data to the sink node directly. If the sink node is mobile in nature, then, the sink node chooses the shortest path using the location of the cluster head. When mobile sink node moves over the region of deployed network along the defined path, then sensor nodes send their sensed data to cluster head. The cluster head performs data aggregation and transmits the same to the mobile sink node. Hence, it decreases wireless data transmissions as well as fewer overheads for relaying data to other nodes. This algorithm increases the transmission and receiving rate to sink node because the aggregated data is not transmitted to sink node over a long distance and hence thereby working far better than existing clustering algorithms such as LEACH, EMGM and SEP. The performance analysis is done through experiments in the results section.

4.2.7 The pseudocode of proposed Algorithm

Our proposed (ECDEN) work balances the energy of all nodes inside the WSNs network by minimizing inter and intra-cluster distance and finds the optimal number of clusters with their cluster centroids. The pseudocode of the proposed algorithm is shown as:-

The pseudocode of proposed Algorithm as follows:

Set the initial values for the Gaussian components, k by analyzing data set, X
For 1 to N **do** // N is number of nodes
Estimate random values for μ, π and ∑ via EMGM
Finds the Inter-cluster distance of each cluster, D_{inter}
Execute E step to find nodes responsibility (R)
Execute M step to find log likelihood (LL)
Update μ , ∑ and number of nodes using R
End for
Cluster centroids (μ) and the number of nodes belongs to each cluster returned by EMGM
Parameters Estimation of Eps and MinPts for DBSCAN
Run DBSCAN (X, Eps, Mints)
Return optimal cluster count (C) and cluster centroid position
For r = 1 to Round // Round is the maximum round number
For j = 1 to C // C is the optimal number of clusters
Determine real CH node based on minimum distance from the cluster centroid, μ
Computes average energy E_{avg} of each cluster
Choose the suitable node as CH whose residual energy, E_{resi} is larger than, E_{avg}
Elected CH node broadcast HELLO message to all member nodes
All member nodes send their received data to CH node
CH node performs the data aggregation to reduce the number of packets
If (Sink node is static)
CH node transmits the compressed data directly
Else (Sink node is mobile)
Mobile sink node calculates the shortest path over CH node Location
All CH Node pass on the data to Mobile sink node when distance between them is minimal
End For
End For

CHAPTER 5: SIMULATION AND RESULT

5.1 Result Analysis for First Proposed Clustering Algorithm

We have briefly exhibited the clustering analysis of the suggested algorithm through simulation using MATLAB. We evaluate our developed clustering algorithms from various aspects like a number of alive nodes, data sending and receiving rate, residual energy over the time. The simulation parameters are listed in Table 1.

TABLE 1 DB-EMGM Initial Energy Parameters

Parameter	**Value**
Distribution of Node	Uniformly Random
Sensor field region	100m *100 m
Number of Nodes, n	100
Initial Node Energy, E_o	0.5 J
Data Message Size	2000 bits
Control Packet Size	50 bits
Simulation Platform	MATLAB

The figure 15 shows that 100 sensors nodes are stationed randomly in the region of size of 100 × 100 m2. The mobile sink is situated at the center of the monitored region.

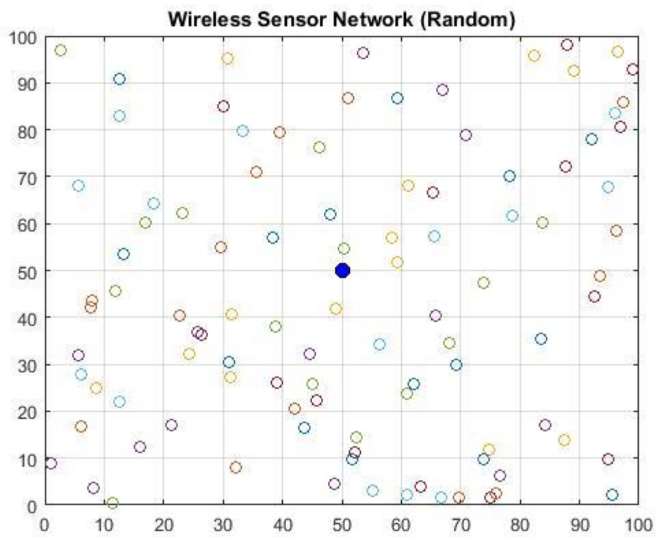

Fig.15 Wireless Random Network

After the executing the proposed DB-EMGM algorithm, the various shape of clusters are created along with the optimal number of clusters in the figure 16. It shows the estimated values of input parameters (Eps, MinPts) for DBSCAN along with noisy nodes that are not the part of any clusters.

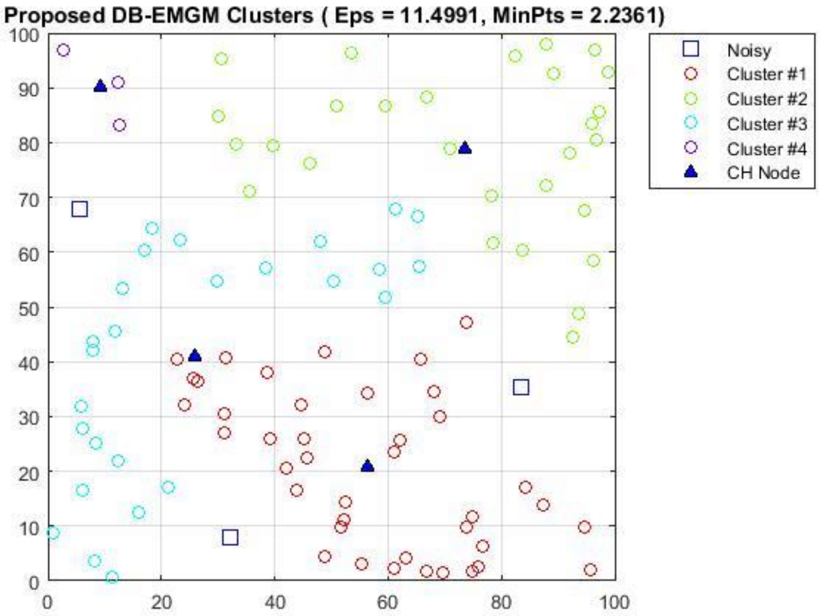

Fig. 16 Clusters found by proposed DB-EMGM

As it is assumed that the sink node is mobile in nature which is positioned initially at the center of the monitoring area. The proposed algorithm transmits the position of the cluster centroid to the mobile sink. The mobile sink node finds the shortest path based on minimum spanning tree for visiting the cluster centroid. Once the sink node comes at cluster centroid, it broadcasts a data request message to all member nodes. All member nodes transmit the data to the sink node. The sink node collects the data and moves on to another cluster centroid along the defined path. The sink node waits for a certain period of time for data collection. The figure 17 shows the path of mobile sink node below.

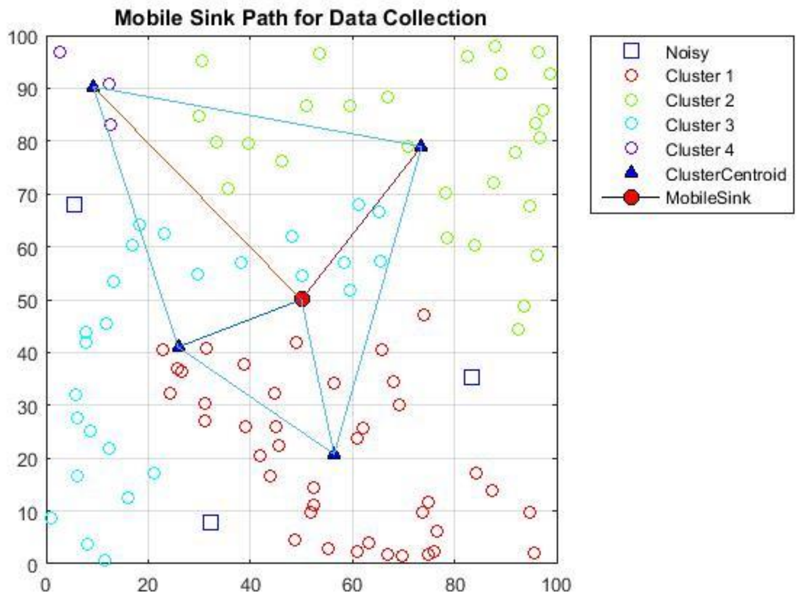

Fig. 17 Path Defined by Mobile Sink Node

Our recommended clustering algorithm has been compared with respect to earlier proposed EMGM method. Figure 18 and 19 shows the graphical results about the sending and receving data rate in regard to the time. The results confirms that DB-EMGM method has been found to be efficiently working better with respect to EMGM because good sets of cluster cannot be created in case of the low or varied density network. The one more limitation of the EMGM algorithm is its inablity to find arbitrarily shaped clusters and noisy data. When movable sink node visits cluster centroid for receving the aggregated data, then all the node does not convey their sensed data over a long communication distance. Hence, data sending and receiving rate is enhanced between the mobile sink node and all sensor nodes and thus saving the energy spent during data transmission in WSN. Therefore, with this new clustering technique, the proposed approach can easily adapts according to variation in node density and thus diminishes data transmission energy.

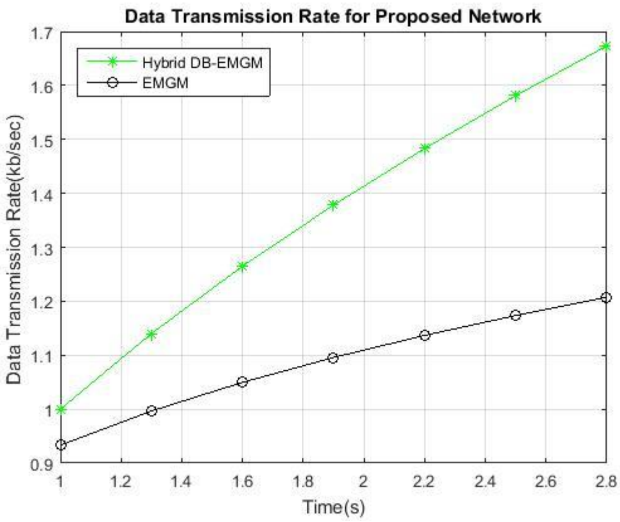

Fig. 18 Data Transmission Rate with Time

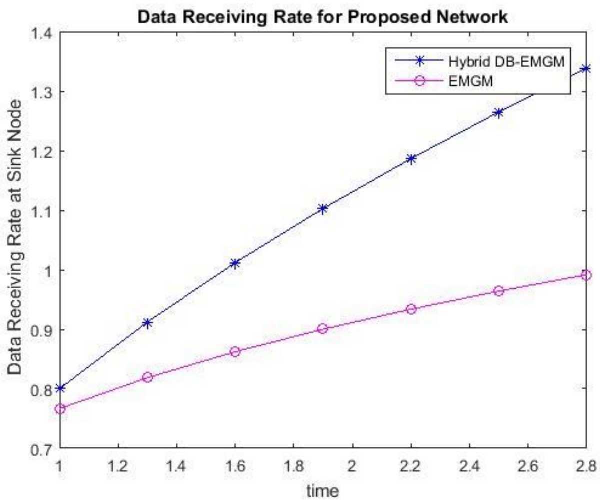

Fig. 19 Data Receiving Rate with Time

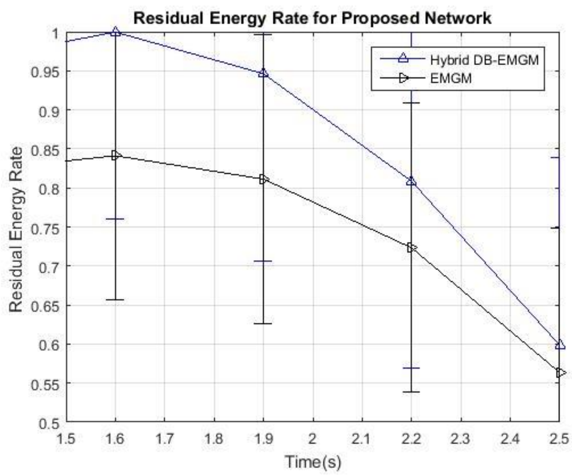

Fig. 20 Residual Energy Rate with Time

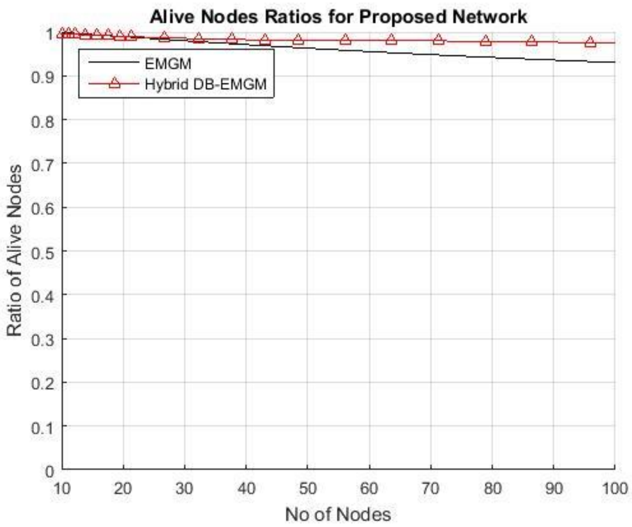

Fig. 21 Ratio of Alive Nodes with No of Nodes

The proposed clustering approach enhances the rate of the clustering process. This method tries to lessen the traffic flow of excess data in the network by getting minimized Intra and Inter-cluster distance among the nodes in WSN. This also discovers the the location of cluster centroid using the estimated mean and cluster-covariance by creating diverse sets of cluster. After the cluster formation phase, the concept of portable sink node is presented to gather the data by patrolling around the sensing zones. The suggested method send the position of cluster centroid to the portable sink node which search for the shortest path to collect the data from all the sensor nodes by staying at the cluster centroid for the certain duration. Upon doing analysis of the improved method, it is pointed out to be computationally much faster as compared to EMGM with reduced Intra and inter-cluster communication distance and boosts the sending rate to the sink node. The validity of the enhanced method is authorized with the help of the results displayed above.

5.2 Result Analysis for Second Proposed Clustering Algorithm

In this part, we judge the our proposed solution performance together with LEACH and SEP protocol. The our suggested algorithm (ECDEN) mainly tries to increase the reliability of the clustering technique used for sensor network with the help of longer stability region so that there is a more time taken by the first node to die as per previvous protocols. For this, static sink node is placed at the center of the network region in which many nodes are deployed. We have carried out the analysis in MATLAB to validate the proposed ECDEN algorithm and to compare its effectiveness with LEACH and SEP protocol. The first order energy model [3] is exploited for computing the energy spent by nodes during packets transmission. An experiment is exercised for deployed WSN network in which 100 nodes are organized randomly in a 100 × 100 m2 as shown in figure 22. The simulation parameters are noted in Table 2.

TABLE 2 ECDEN Initial Energy Parameters

Parameter	**Value**
WSN Node Distribution	Uniform Random Distribution
Sensor Node Area	100m *100 m
Base Station Position	Center
Number of nodes, n	100
Initial energy of nodes, E_o	0.5 J
Message packet size	4000 bits
Control packet size	50 bits
Simulation platform	MATLAB

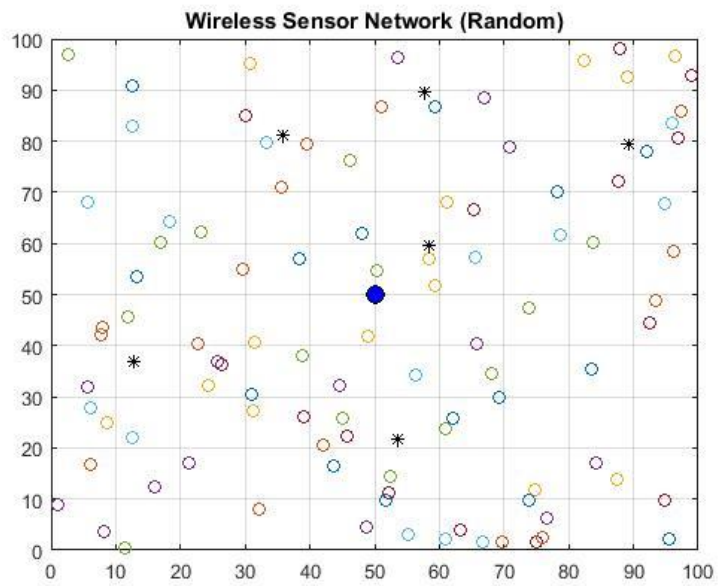

Fig. 22 Wireless Sensor Network (Random) before Clustering

Proposed ECDEN is executed on the deployed network of 100 nodes and generates clusters set as presented in figure 23. This returns six different clusters set along with noisy nodes. The square shaped nodes named as noisy nodes are not allocated to any clusters set because they does not comes in the range of the radius, Eps of each clusters. The such noisy nodes passes their sensed data to nearest cluster heads regurlay.

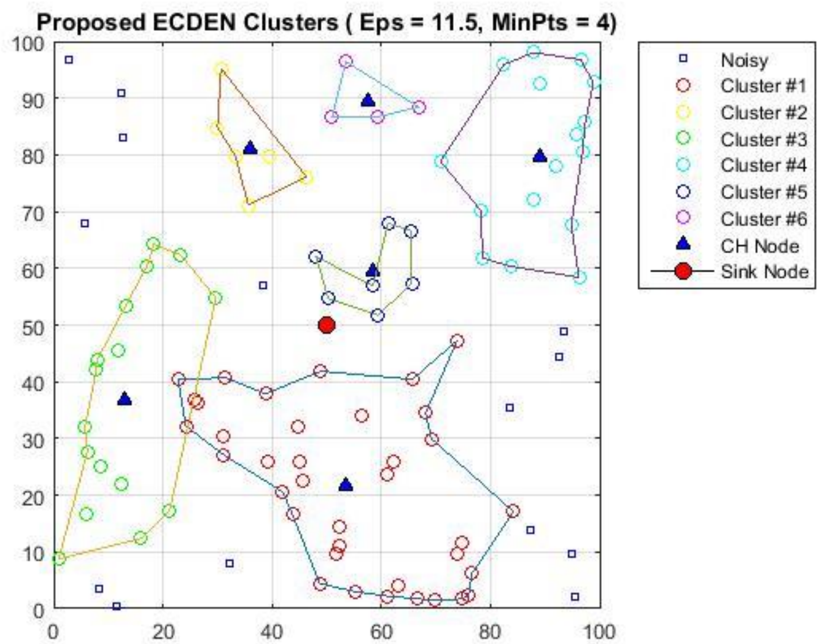

Fig. 23 Network Clustering Using Proposed ECDEN

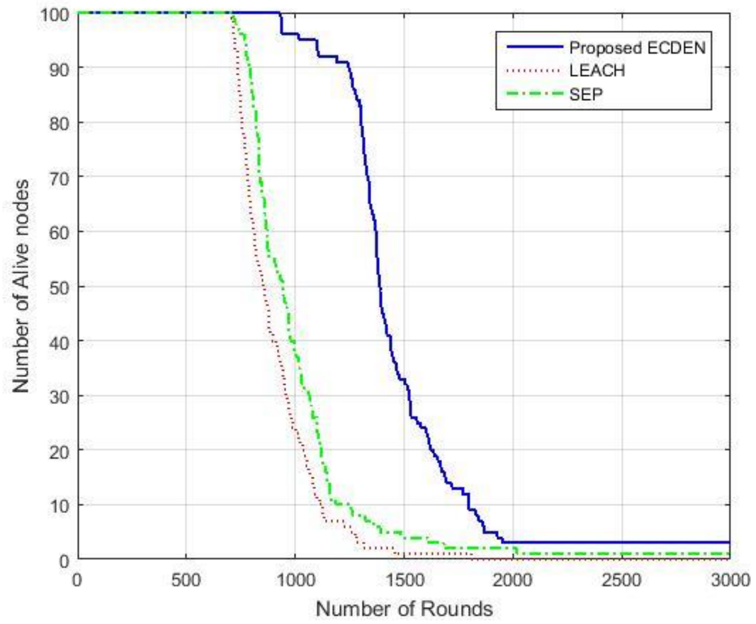

Fig. 24 Number of Alive Nodes Per Round

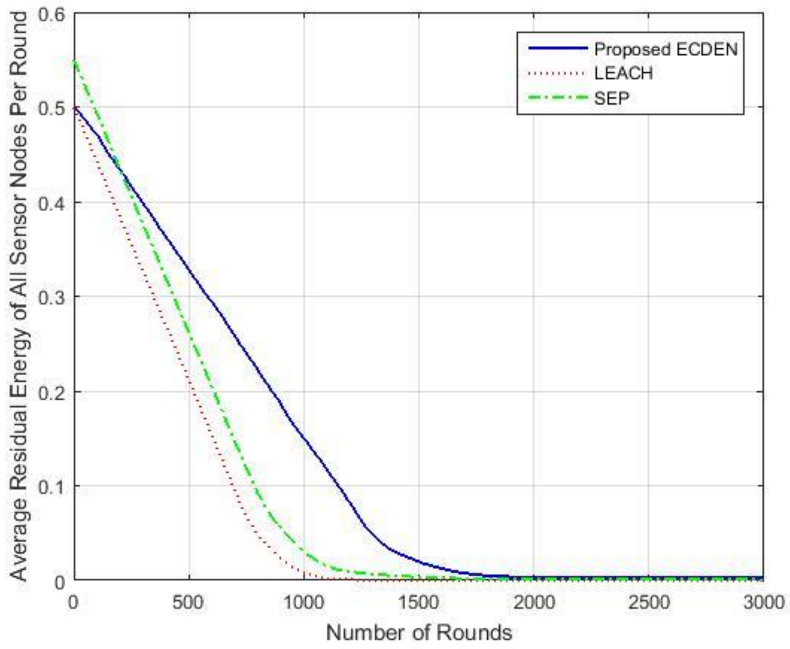

Fig. 25 Average Residual Energy Per Round

The figure 24 and 25 show that introduced algorithm increases the lifetime of network and efficient utilization of the remaining power for all nodes as compared to LEACH and SEP protocol. From the figure 24, it can be observed that the first node died after 703 rounds and all sensor nodes died after 1490 rounds in LEACH protocol and after 717 and 1710 in the case of SEP protocol. However, in Proposed ECDEN, the first node died after 933 rounds, and the last node died after 2700. Hence, it understood that proposed method extends the reliability of clustering process and also lengthens the lifetime of WSN as compared to LEACH and SEP protocol. Some of the nodes are determined as noisy by proposed method, and noisy nodes forward their data to nearest CHs. Apart from this, this method also saves the remaining energy of all nodes. The reason for this is due to that proposed ECDEN always selects the CHs from nodes which are near to cluster centroid and have their residual energy greater than average energy of each cluster. As clustering of the network is realized by considering the node density variation through the recommended above approach to reduce data transmission energy.

Hence our second proposed algorithms offers an novel cluster head selection method. This balances the energy among all nodes by choosing CHs based on distance and residual energy. The MATLAB simulation results showed that the proposed method is more energy efficient and reliable in clustering process as compared to LEACH and SEP protocol. The results confirm that it is computationally faster than LEACH as well as SEP and can work better even in the case of node density variation. This method extends the transmission rate to the sink node and hence maintains longer stability period. Moreover, there is a lot of scope in future for the proposed density based clustering algorithm in different application such as extracting useful data from medical diagnosis data.

CHAPTER 6: CONCLUSIONS

6.1 Conclusion

In this thesis work, we have described the two novel proposed clustering approach for solving the challenges related to clustering for accummulating data received for the nodes. The suggested algorithm focus on generating arbitrary shaped cluster sets so that compressed data can be transmitted by appropriate cluster head to the static or mobile sink node in geographically distributed WSNs. As such networks are constructed for bringing together a large amount of data from the millions of sensors. Our recommended hybrid clustering technique solves the problem of the optimal number of clusters along with getting variable sized cluster sets. Both the proposed algorithm utilizes the concept of static or mobile sink node so that data transmission rate is increased. By considering the energy factor of WSN, we have proposed an efficient cluster head selection method based on density of nodes for solving clustering hurdle and maximize the lifetime with enhanced stability period in WSN. The proposed algorithm ECDEN clustering method balances the energy among all nodes by selecting appropriate CHs based on distance and residual energy and is more energy efficient and reliable in clustering process as compared to LEACH and SEP protocol. Hence, both the clustering approach is found to be optimized approach as per the presented analysis and proves to be enhancing the lifetime of WSN.

6.2 Scope for Future Work

Moreover, there is a lot of scope in future because sensor network will be used everywhere in any fields in order to make future technologies (Healthcare, Smart homes thorough sensors, environment monitoring and control, security, IOT) as smart as possible. Even, the recommended density based clustering algorithm can be used in different application such as extracting useful data from medical diagnosis data, image processing, etc.

REFERENCES

1. J. MacQueen et al., “Some methods for classification and analysis of multivariate observations,” in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 281- 297. California, USA, 1967, p. 14.
2. L. Kaufman and P. J. Rousseeuw, Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, 2009, vol. 344.
3. W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy-efficient communication protocol for wireless microsensor networks,” in System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on, Jan 2000, p. 10 vol.2.
4. Ossama Younis, Sonia Fahmy," HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks," IEEE Transactions on Mobile Computing, vol. 03, no.4, pp. 366-379, Oct.,2004.
5. T. Zhang, R. Ramakrishnan, and M. Livny, “Birch: an efficient data clustering method for very large databases,” in ACM SIGMOD Record, vol. 25, no. 2. ACM, 1996, pp. 103–114.
6. S. Guha, R. Rastogi, and K. Shim, “Cure: an efficient clustering algorithm for large databases,” in ACM SIGMOD Record, vol. 27, no. 2. ACM, 1998, pp. 73–84.
7. Martin Ester, Han-peter Kriegel, Jorg Sander, Xiaowei Xu,”A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226 – 231, 1996.
8. M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, “Optics: Ordering points to identify the clustering structure,” in ACM SIGMOD Record, vol. 28, no. 2. ACM, 1999, pp. 49–60.
9. W. Wang, J. Yang, and R. Muntz, “Sting: A statistical information grid approach to spatial data mining,” in VLDB, vol. 97, 1997, pp. 186–195.
10. G. Sheikholeslami, S. Chatterjee, and A. Zhang, “Wavecluster: a wavelet-based clustering approach for spatial data in very large databases,” The VLDB Journal, vol. 8, no. 3-4, pp. 289–304, 2000.
11. R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, Automatic subspace clustering of high dimensional data for data mining applications ACM, 1998, vol. 27, no. 2
12. W. Heinzelman, A. Chandrakasan and H. Balakrishnan, "An Application-Specific Protocol Architecture for Wireless Microsensor Networks," IEEE Transactions on Wireless Communications, vol. 1, no. 4, (2002) October, pp. 660-670.
13. W. Liu and L. Wang, “An improved algorithm based on LEACH protocol”, Journal of Applied Mechanics and Materials, vol. 347, (2013), pp. 2725-2727.
14. V. Loscrì, G. Morabito and S. Marano, “A Two-Level Hierarchy for Low-Energy Adaptive Clustering Hierarchy”, IEEE 62nd Vehicular Technology Conference, (2005) September 25-28, pp. 1809-1813.
15. K. Go. Vijayvargiya and V. Shrivastava, “An Amend Implementation on LEACH protocol based on Energy Hierarchy”, International Journal of Current Engineering and Technology, vol. 2, no. 4, (2012) December, pp. 427-431.

16. G. Smaragdakis, I. Matta and A. Bestavros, “SEP: A Stable Election Protocol for clustered heterogeneous wireless sensor networks”, Second International Workshop on Sensor and Actor Network Protocols and Applications (SANPA), (2004).
17. M. Ye, C. Li, G. Chen, and J. Wu, "EECS: an energy efficient clustering scheme in wireless sensor networks," in PCCC 2005. 24th IEEE International Performance, Computing, and Communications Conference, 2005.,2005, pp. 535-540.
18. S. Lindsey and C. S. Raghavendra, "PEGASIS: Power-Efficient Gathering in Sensor Information Systems," in the Proceedings of the IEEE Aerospace Conference, Big Sky, Montana, Vol. 3, PP 1125 – 1130, March 2002.
19. Smiti, Abir, and Zied Eloudi. "Soft DBSCAN: Improving DBSCAN Clustering method using fuzzy set theory." In the IEEE 6th International Conference on Human System Interaction (HIS- 2013), pp. 380-385, 2013.
20. Santosh Kumar Rai, Nishchol Mishra “DBCSVM: Density Based Clustering Using Support Vector Machines”, IJCSI International Journal of Computer Science Issues, ISSN (Online): 1694-0814, Vol. 9, Issue 4, No 2, pp. 223 – 230, July 2012
21. Abir Smiti and Zied Eloudi “DBSCAN-GM: An improved clustering method based on Gaussian Means and DBSCAN techniques”, IJCSI IEEE 16th International Conference on Intelligent Engineering Systems (INES June 2012).
22. D. Raja Kishor, N. B. Venkateswarlu “Hybridization of Expectation-Maximization and K-Means Algorithms for Better Clustering Performance”, Published in Journal Cybernetics and Information Technologies, Volume 16 Issue 2, 6 2016, pp. 16-34.
23. D. Takaishi, H. Nishiyama, N. Kato, and R. Miura “Toward Energy Efficient Big Data Gathering in Densely Distributed Sensor Networks” IEEE transactions on emerging topics in computing, Volume 2, No. 3, September 2014.
24. T. K. Moon, “The expectation-maximization algorithm,” Signal Processing Magazine, IEEE, vol. 13, no. 6, pp. 47–60, Nov. 1996.
25. E. Chandra and V. P. Anuradha, “A survey on clustering algorithms for data in spatial database management systems,” International Journal of Computer Applications, vol. 24, no. 9, pp. 19–26, June 2011, published by Foundation of Computer Science.
26. Sagiroglu and D. Sinanc, ``Big data: A review," in Proc. Int. Conf. CTS, 2013.
27. S. Wen-Zhan, H. Renjie, X. Mingsen, B. A. Shirazi, and R. Lahusen, ``Design and deployment of sensor network for real-time high fidelity volcano monitoring," IEEE Trans. Parallel Distrib. Syst., vol. 21, no. 11, pp. 1658-1674, Nov. 2010.
28. K. Baumgartner, S. Ferrari, and A.V. Rao, ``Optimal control of an underwater sensor network for cooperative target tracking," IEEE J. Ocean. Eng.,vol. 34, no. 4, pp. 678697, Oct. 2009.
29. R. C. Shah, S. Roy, S. Jain, and W. Brunette, ``Data MULEs: Modeling and analysis of a three-tier architecture for sparse sensor networks," Ad Hoc Netw., vol. 1, nos. 23, pp. 215-233, 2003.
30. H. Nakayama, N. Ansari, A. Jamalipour, and N. Kato, ``Fault-resilient sensing in wireless sensor networks," Comput. Commun., vol. 30, nos. 1112, pp. 2375-2384, Sep. 2007.
31. M. Youssef, A. Youssef, and M. Younis, ``Overlapping multi-hop clustering for wireless sensor networks," IEEE Trans. Parallel Distrib. Syst., vol. 20, no. 12, pp. 1844-1856, Dec. 2009.

32. D. Baum and CIO Information Matters. (2013). *Big Data, Big Opportunity* [Online].
33. Ramaswamy, V. Lawson, and S. Gogineni, ``Towards a quality centric big data architecture for federated sensor services," in *Proc. IEEE Int. BigData Congr.*, Jul. 2013, pp. 86-93.
34. Swati Dewangan, S. D. Mishra "A Survey on Energy Optimization in WSN using Distributed Clustering Approach" International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 11, November 2016.
35. Mohammed T. H. Elbatta and Wesam M. Ashour, "A Dynamic Method for Discovering Density Varied Clusters", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.-6 (1), 2013.
36. Hongfang Zhou, Peng Wang, Hongyan Li, "Research on Adaptive Parameters Determination in DBSCAN Algorithm", Journal of Information & Computational Science Vol.-9, pp. 1967–1973, 2012.
37. Manisha Naik Gaonkar, Kedar Sawant, "AutoEpsDBSCAN:DBSCAN with Eps Automatic for Large Dataset", International Journal on Advanced Computer Theory and Engineering (IJACTE), Vol.-2, pp. 2319–2526, 2013.
38. Amin Karami, Ronnie Johansson, "Choosing DBSCAN Parameters Automatically using Differential Evolution", International Journal of Computer Applications, Vol.-91, 2014.