Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-02

A Review: An Improved K-means Clustering Technique in WSN

NAVJOT KAUR JASSI, SANDEEP SINGH WRAICH

ABSTRACT: **A wireless sensor network (WSN) consists of spatially distributed autonomous sensors to monitor physical or environmental conditions and to cooperatively pass their data through the network to a Base Station. Due to the increase in the quantity of data across the world, it turns out to be very complex task for analyzing those data. Categorize those data into remarkable collection is one of the common forms of understanding and learning. This leads to the requirement for better data mining technique. These facilities are provided by a standard data mining technique called Clustering. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data. This paper reviews four types of clustering techniques- K-Means Clustering, LEACH, HEED, and TEEN. K-Means clustering is very simple and effective for clustering. It is appropriate when the large dataset is used for clustering. Simulated study and the experiment results are also presented in this paper.**

Keywords: **WSN, cluster, k-means, LEACH, HEED, TEEN, Centroid, Clusterhead.**

I. INTRODUCTION

Sensor networks can be applied in a variety of areas such as target tracking, environment monitoring, military surveillance, medical applications, etc. with the advances in wireless communication made it possible to develop wireless sensor networks (WSN) consisting of small devices, which collect information by cooperating with each other. These small sensing devices are called nodes and a typical node of a WSN consist of four components:

Navjot Kaur Jassi, Guru Nanak Dev University Amritsar,India
Sandeep Singh Wraich, Assistant Professor, GNDU,Amritsar, India

A *sensor* that performs the sensing of required events in a specific field, a *radio transceiver* that performs radio transmission and reception, a *microcontroller*: which is used for data processing and a *battery* that is a power unit providing energy for operation. The size of each sensor node varies with applications.
The nodes in WSNs are powered by batteries, it is expected that these batteries lasted for years before they can be replaced. Due to the cost and small size of the sensor nodes, they have been equipped with small batteries with limited energy source [4]. This has been a major constraint of wireless sensor nodes which limits their lifetime and affects utilization of the wireless sensor networks.

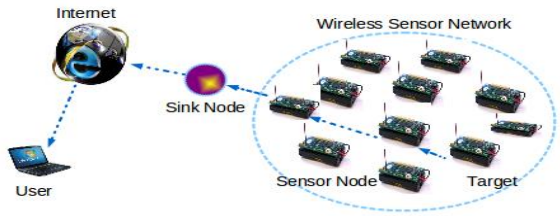

Fig1. Wireless sensor network

To extend batteries' lifetime and networks utilization, constant changing the batteries when they run out of energy may not be practical, since these nodes in most cases are many (tens to thousands of sensor nodes), recharging the weaken batteries at all time may not be feasible. Therefore, there is a need to minimize energy consumption in WSNs. The following steps can be taken to save energy caused by communication in wireless sensor networks [1].
a) To schedule the state of the nodes (i.e. transmitting, receiving, idle or sleep).
b) Changing the transmission range between the sensing nodes.
c) Using efficient routing and data collecting methods.
d) Avoiding the handling of unwanted data as in the case of overhearing.
In many cases (e.g. surveillance applications), it is undesirable to replace the batteries that are depleted or drained of energy. Many researchers are therefore trying to find power-aware protocols for wireless sensor networks in order to overcome such energy efficiency problems as those stated above.

Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-02

II. Data clustering

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering [9]. A *cluster* is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. In each cluster, sensor nodes are given different roles to play, such as cluster head, ordinary member node, or gate way node. A cluster head (CH) is a group leader in each cluster that collects sensed data from member nodes, aggregate, and transmits the aggregated data to the next CH or to the base station [6]. Clustering is an effective technique for reducing energy consumption and extending sensor network lifetimes [3]. Compared to other methods its advantages include reducing the size of the routing table in member nodes, conserving communication bandwidth, prolonging the battery life of member sensors, cutting the overhead from topology maintenance, and reducing redundant packets. A clustering algorithm attempts to find natural groups of components (or data) based on some similarities. The clustering algorithm also finds the centroid of a group of data sets. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster.

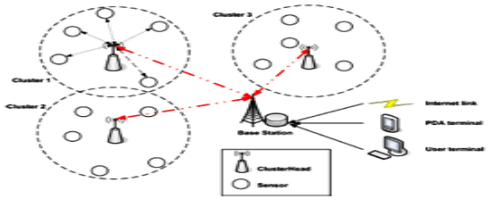

Fig2: clustered based wireless sensor network

The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters. Clustering algorithms are often useful in applications in various fields such as visualization, pattern recognition, learning theory, computer graphics, neural networks, Artificial intelligence and statistics. Practical applications [6] of clustering include pattern classification under unsupervised learning, proximity search, time series analysis, text mining and navigation.

a) Clustering Advantages

Clustering has numerous advantages. Some of these are:

1. Clustering reduces the size of the routing table stored at the individual nodes by localizing the route set up within the cluster (Akkaya 2005).
2. Clustering can conserve communication bandwidth since it limits the scope of inter-cluster interactions to CHs and avoids redundant exchange of messages among sensor nodes.
3. The CH can prolong the battery life of the individual sensors and the network lifetime as well by implementing optimized management strategies (Younis 2003).
4. Clustering cuts on topology maintenance overhead. Sensors would care only for connecting with their CHs (Hou 2005).
5. A CH can reduce the rate of energy consumption by scheduling activities in the cluster.

A. Data Clustering Techniques

The main target of hierarchical routing or cluster based routing is to efficiently maintain the energy usage of sensor nodes by involving them in multi-hop communication within a particular cluster. With clustering in WSNs, energy consumption, lifetime of the network and scalability can be improved. Various routing techniques are as follows:

a) Low-Energy Adaptive Clustering Hierarchy (LEACH) [12] is a popular clustering protocol for WSNs. It has inspired many subsequent clustering protocols. LEACH's operation is comprised of many rounds, where each round includes two phases, a setup phase and a steady phase.

i. In the setup phase, LEACH randomly selects sensor nodes as Cluster Heads (CHs), subsequently rotating roles in each round in order to evenly distribute energy dissipation to all sensor nodes of the network.

ii. In the steady-state phase, data are delivered from sensor nodes to CHs and from CHs to the BS per a TDMA/CDMA schedule.

LEACH doesn't need global information for the network, and is a distributed and adaptive approach. It achieves various benefits from balancing loads via rotation, such as reducing conflicts by TDMA schedule and saving energy through assignment of on-or-off states by time slot [5].

However, there are also disadvantages. LEACH assumes sensor nodes are homogeneous but in reality sensor node energy distribution occurs in a heterogeneous manner, especially over a period of time. The CH selection in LEACH is based on probabilities without considering residual energy of sensor nodes, and there is the possibility that CHs are overwhelmed in some areas and infrequently assigned in others. Therefore, CHs are not uniformly distributed in the network and load balancing cannot be guaranteed. Moreover, LEACH needs re-clustering in each round, which may diminish any energy savings gain.

Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-02

b) Hybrid Energy-Efficient Distributed clustering (HEED) [8] is an energy-efficient protocol. It selects CHs periodically based on residual energy and on intra-cluster communications costs among sensor nodes. HEED is fully distributed and adaptive. CHs can be uniformly distributed throughout the network thereby balancing the load. However, like LEACH, HEED requires re-clustering each round which brings significant overhead and diminishes energy gains. Furthermore, HEED can't avoid the hot spot issue.

c) TEEN [15] is a cluster based hierarchical routing protocol and it is based on LEACH. Using this protocol the data can be sensed continuously but transmission of data is not done frequently. TEEN uses LEACH's approach to form clusters. Hard threshold (HT) and soft threshold (ST) are two types of threshold mode used in this algorithm. In HT mode the sensed attribute will be within the range of interest in order to send the data. In ST mode any changes in the value of the sensed attribute will be transmitted. The nodes sense their environment frequently then store the sensed value for transmission. If it satisfies the below conditions then only the node transmits the sensed value:

i. If sensed value greater than the hard threshold (HT).

ii. If sensed value is not hard threshold and greater than or equal to soft threshold (ST).

In TEEN, cluster head always waits for their time slot for data transmission. Suppose node has no data then time slot may be wasted.

B. Parameters

Selecting appropriate parameter for any given WSN is another challenge. But here we wanted to put forward those parameters which are common for almost all WSN which are given as follows:

a) Nodes Energy

W. heinzelmann *et al*. used node energy level 2J where other research has used range varying from 1J to 5 J [6] .For experiment the energy of node is varied between 1J to 3J.

b) Distance

One of the most significant factors for WSN is distance. It is appropriate to use Euclidian distance from the base station (BS) to each node. It varied from 85 m to 175 m. The higher the distance more energy is required to data interchange [6].

$$\text{Euclidian distance} = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}$$

c) Cluster Head Selection

After the centroid positions are finalized in the clustering process, we consider nodes which are at the nearest distance and also the next nearest distance from the centroid. The node with highest energy is considered as Cluster Head. If more than one node in the two levels has the highest energy then the node nearest to the centroid is selected as cluster head

d) Latency

Clustering in WSN should take into account the latency [6]. The latency is a function of the number of communication hops between the source and the gateway (CH). Latency is also a dependent on distance. As the distance increases the latency increases and vice versa.

III. K-means clustering

K-means is a commonly used partitioning based clustering technique that tries to find a user specified number of clusters (k), which are represented by their centroids, by minimizing the square error function developed for low dimensional data, often do not work well for high dimensional data and the result may not be accurate most of the time due to outliers. There are two simple approaches to cluster center initialization i.e. either to select the initial values randomly, or to choose the first k samples of the data points. As an alternative, different sets of initial values are chosen (out of the data points) and the set, which is closest to optimal, is chosen [11].

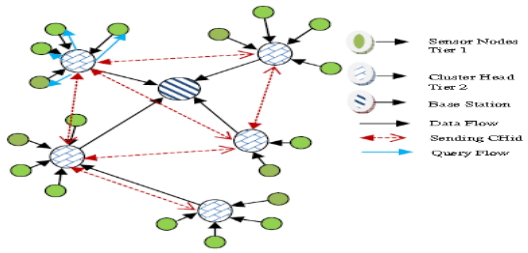

Fig3: Sensed Data forwarding with clustering and aggregation

Data Aggregation

In Sensor Network, data aggregation removes the redundancy among data collected by different sensors and consequently aims at load reduction over the network.

The computational complexity of original K-means algorithm is very high, especially for large data sets. In addition the number of distance calculations increases exponentially with the increase of the dimensionality of the data.

A. K-Means clustering Algorithm

K-means algorithm is based mainly on the Euclidian distances and cluster head selection depends on residual energies of nodes. So here the central node collects the information about the node id, position and residual energy of all nodes and stores this information in a list in the central node. After getting

Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-02

this information from all nodes it starts performing the clustering algorithm (k-mean) [11].

Algorithm:

1. If we want to cluster the nodes into 'k' clusters, take 'k' number of centroids initially at random places
2. Calculate the Euclidian distance from each node to all centroids and assign it to centroid nearest to it. By this 'k' initial clusters are formed .Suppose there are n nodes are given such that each one of them belongs to Rd. The problem of finding the minimum variance clustering of this nodes into k clusters is that of finding the k centroids { mj }k j = 1 in Rd such that,

$$(1/n)*(\sum min_j d^2(X_i, m_j)), for\ i-1\ to\ n$$

for i = 1 to n, where d(X_i, m_j) denotes the Euclidean distance between Xi and mj. The points {j}k i = 1 are known as cluster centroids or as cluster means.

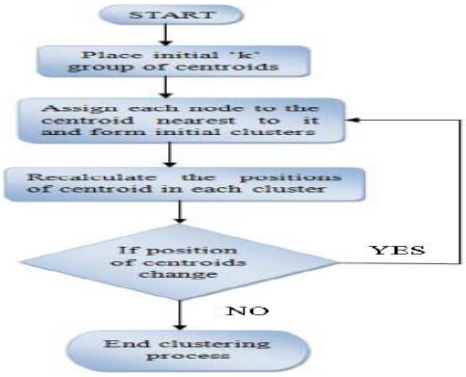

Fig4. Flow chart showing sequence of k-means algorithm

3. Recalculate the positions of centroids in each cluster and check for the change in position from the previous one
4. If there is change in position of any centroid then go to STEP 2, else the clusters are finalized and the clustering process ends

By this the clustering of nodes into 'k' number of clusters is done [11] and the cluster heads in each cluster are to be chosen as shown in Fig4.

IV. Experiments and results

In our work we carry out the simulation with 150 sensors deployed randomly assuming that they are deployed in clusters with inter-clusters communication will happen only through cluster heads of the respective clusters.

a. Maximum and Minimum Range of Node

We see in the four clustered network that change in the parameter combination results in an alternation in the range (maximum to minimum) of nodes per cluster. Our experiments show that by increasing or decreasing we can alter node range in any specific cluster.

Table1. Average nodes per cluster for experiment

Number of clusters	Average nodes/clusters
2	75
3	50
4	37.3
5	30

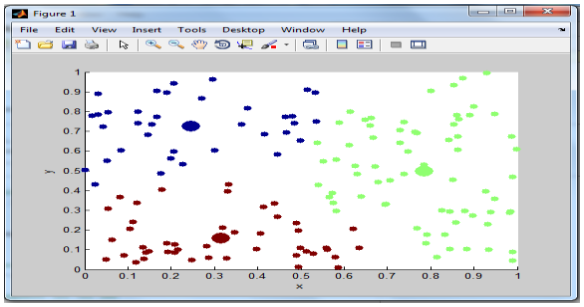

Fig5 Three clusters each having 50 sensors

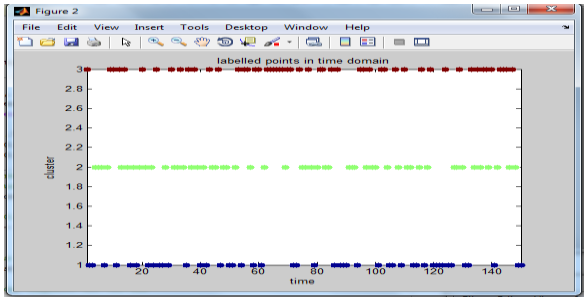

Fig6 Labeled points in time domain during clustering (3 clusters)

b) Centroids with Low Power Requirement

If the clusters that we choose have centroids with high energy that will results in a less energy efficient. From our experiment we see that the increase in the number of parameter in five clustered network results in centroids with low power requirement for the same cluster.

Table2. Parameters used in simulation study

Parameter	value
Node energy	1J-3J
Latency	Depends on the distance
Number of nodes	150

Table3. List of intra cluster distance for multiple parameters for different cluster size

For K=2	K=3	K=4	K=5
29.3617	24.7769	21.4021	18.8705
28.7003	22.1308	19.5844	15.2605
27.7292	21.6504	18.9276	13.962
26.8674	21.6016	18.77	13.4374
26.5038	21.6016	18.6667	13.2889

Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET-2014

ISBN: 978-1-63248-028-6 doi: 10.15224/ 978-1-63248-028-6-01-02

V. Conclusion and Future work

A brief introduction about WSNs, routing techniques, clustering is presented. There are so many routing techniques developed to solve the energy consumption problem and for enhancing the lifetime of WSNs. some of them are discussed above such as LEACH, HEED, TEEN. But mostly every technique has certain limitations like various clustering techniques do not suitable for large scale networks. But K-means clustering algorithm has biggest advantage of clustering large data sets and its performance increases as number of clusters increases and the performance of K- mean algorithm is better than Hierarchical Clustering Algorithm. some of the further work that is also done for further research work is:

a) Different WSN is made up for different purposes. So the effect of parameters will vary from network to network. In out experiment we tried to work with those parameters which are common for almost all the sensors. But finding the right combination of parameters and their optimum value for specific WSN is a great challenge.

b) For our experiment we have worked with uniform distributed data type. Further experiment can be done by choosing exponential, Negative exponential distribution.

REFERENCES

[1] Ameer Ahmed Abbasi a, Mohamed Younis, "A survey on clustering algorithms for wireless sensor networks", Computer Communications 30 Elsevier (2007) 2826–2841.

[2] Malay K. Pakhira, "A Modified k-means Algorithm to Avoid Empty Clusters", International Journal of Recent Trends in Engineering, Vol 1, No. 1, May 2009.

[3] Xiang Mina,b, ShiWei-rena, JiangChang-jianga, ZhangYing, "Energy efficient clustering algorithm for maximizing lifetime of wireless sensor networks", Int. J. Electron. Commun. (AEÜ) , Elsevier (2010).

[4] Chi-Tsun Cheng, Member, IEEE, Chi K. Tse, Fellow, IEEE, and Francis C. M. Lau, Senior Member, IEEE," A Clustering Algorithm for Wireless Sensor Networks Based on Social Insect Colonies", IEEE SENSORS JOURNAL, VOL. 11, NO. 3, MARCH 2011.

[5] Jalil Jabari Lotf, Seyed Hossein Hosseini Nazhad Ghazani, "Clustering of Wireless Sensor Networks Using Hybrid Algorithm", Australian Journal of Basic and Applied Sciences, 5(8): 1483-1489, 2011.

[6] K.Ramesh and Dr. K.Somasundaram, "A Comparative Study of Clusterhead Selection Algorithms in Wireless Sensor Networks", International Journal of Computer Science & Engineering Survey Vol.2, No.4, November 2011.

[7] K.Ramesh and Dr. K.Somasundaram, "A Comparative Study of Clusterhead Selection Algorithms in Wireless Sensor Networks", International Journal of Computer Science & Engineering Survey Vol.2, No.4, November 2011

[8] Luke K. Wang, Chien-Chang Wu,"A Practical Target Tracking Technique in Sensor Network Using Clustering Algorithm", November 2012.

[9] Asif Khan, Israfil Tamim, Emdad Ahmed, "Muhammad Abdul Awal, Multiple Parameter Based Clustering (MPC): Prospective Analysis for Effective Clustering in Wireless Sensor Network (WSN) Using K-Means Algorithm", January 2012.

[10] Arash Ghorbannia Delavar, Abootorab Alirezaie Article: KGAWSN: An Effective Way to Reduce Energy Consumption in Wireless Sensor Networks by K-means and Genetic Algorithms. International Journal of Computer Applications, June 2012.

[11] P.Sasikumar, Sibaram Khara, "k-MEANS Clustering in Wireless Sensor Networks", Fourth International Conference on Computational Intelligence and Communication Networks,2013.

[12] Atefeh Heydariyan, Amir Abbas Baradaran, "IMKREC: Improved k-means Algorithm Method for Reducing Energy Consumption in WSN", Journal of Basic and Applied Scientific Research, Basic. Appl. Sci. Res., 2013.

[13] Chunfei Zhang, Zhiyi Fang, "An Improved K-means Clustering Algorithm", Journal of Information & Computational Science, 2013.

[14] Amit Bhattacharjee, Balagopal Bhallamudi and Zahid Maqbool, "Energy-Efficient Hierarchical Cluster Based Routing Algorithm In Wsn: A Survey", International Journal of Engineering Research & Technology, May – 2013.

[15] Qing Yan Xie, Yizong Cheng," K-Centers Min-Max Clustering Algorithm over Heterogeneous Wireless Sensor Networks", IEEE 2013.

I,Navjot Kaur Jassi Pursuring M.tech from Guru Nanak Dev University,Amritsar,Punjab,India