Volume 5, Issue 12, December 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Energy Efficient Quality Improving Clustering Technique

Purnima Bholowalia*, Arvind Kumar
CSE, LPU, Phagwara, Punjab,
India

Abstract— Deciding the number of clusters denoted 'k' by the user is pre-requisite for almost every clustering algorithm. Quality of clustering algorithm and cost of routing computations highly depends upon the no. of clusters determined by the user. But it is not an easy task every time for user to chosen best 'k' always because absence of prior knowledge about quantity of datasets for clustering is available. In our work, we have focused on determining no. of clusters automatically by using incremental method for better quality clustering. We evaluate value of 'k' by implementing incremental method in which the incremental value drop dramatically and provide us a angle to define best 'k'.

Keywords— Quality Clustering, Elbow method, Lowering routing computations, k-mediod, Energy.

I. INTRODUCTION

As the market of computers is growing rapidly, unexpectedly increase in the processing power demands the use of wireless sensor networks heavily. Wireless sensor networks composed of vast number of sensor nodes for sensing data. Nodes are deployed densely either in WSN itself or somewhere near to it. These sensor nodes are low in cost but other hand these have limited energy. In remote areas, the deployment of WSNs is increasing to monitor and sense events' data. Sensor nodes sensed data and disseminate sensed data to far away placed base stations for further processing of sensed data. For long life of network, all nodes must consume energy in some synchronous way so that some of nodes remain alive. Many energy-efficient algorithms exist developed by many researchers. As shown in figure 1.1, WSN consists numerous sensor nodes and these nodes linked with another node called sink and further this sink node linked with Internet for passing information about events occurred and sensed by sensor nodes in field to task manager node (user). Satellite associated is also shown in this figure. In this figure, 'A' node sensed data in sensor field but due to short transmission range with 'Sink' node 'A' node cannot pass data directly to 'Sink' node. So to overcome this problem, shown path has developed to transfer sensed data to end-user. Path includes following steps:

- 'A' node sensed data in sensor field.
- 'A' node send data to its neighbour 'B' node
- Then 'B' node pass data to its neighbour 'C' node
- And data transferred by path *A-B-C-D-E-Sink.*
- Now 'Sink' is having sensed data and internet connection.
- Sensed data delivered by 'Sink' to 'end-user' through internet.

Sensed data analyses cooperatively and autonomously so that any redundant data observed can be prune and not to deliver unnecessary data repeatedly by sink to the user. All nodes finds their neighbour nodes autonomously for communication with each other using Single-hop or Multi-hop methods of communication.

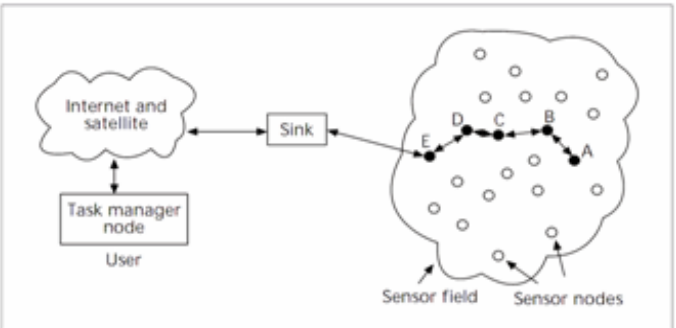

Fig. 1 Sensor nodes in a sensor field

In WSNs micro-sensing technologies are embedded for sensing and detecting events in many environments like battle-field, border-protection, health-related areas, disasters areas they deployed. In battlefield, tanks of enemy are detected and tracked by sensor nodes deployed in wsn. In buildings, personnel are tracked by sensor nodes. To monitor traffic on road, wsn are used. To detect rain and fire in forest, wsn are highly used in efficient manner. To monitor environmental

pollutants, wsn are used. WSN is mainly used these days in tracking sun rays for collecting the solar power, hence used for electricity production. Apart from these usage and advantages, WSN has some limitations also such as limited computational resources, power restrictions, and randomly changed topology, error-prone medium. Energy consumption is one of most important limitation and many researches are on-going to make energy-efficient ways for WSN.

One technique named 'Clustering' is very efficient in reducing traffic among network and base-station. Clustering saves energy by reducing energy consumption and also prolong lifetime of network. Small units called clusters are formed which are very easy to manage as compared to large whole network unit. Clustering improves scalability and bandwidth. It is an important factor upon which efficiency of network depends. Inter-clusters communication conserves bandwidth, reduces redundant data transfer. A part from these advantages, one issue of clustering scheme that cannot be neglect is that is how to determine the optimal number of clusters so that energy consumption can be reduced. The proposed method decides the number of cluster for efficient clustering algorithm PAM. In a network, it is very hard to find small clusters when there are larger clusters present near and far distant places. And it becomes difficult to find best 'k' value when there are many overlapping clusters available. Hence, it is very difficult to find 'k' value. PAM clustering technique is used to reduce the sum of data points' dissimilarities and for this purpose PAM search for possible medoids 'k' among data points and allocate each data point to a nearest medoids. PAM requires pre-determined value of 'k' as input for cluster analysis. Many methods for estimation of 'k' value have been present only focusing on global structure of cluster but no method focus on separation of overlapping clusters. Hence the wrong estimation of 'k' value is possible. To avoid this situation of overlapping between small and large clusters, we propose a method to evaluate no. of clusters based on focusing is it better to merge or separate clusters nearby or overlapping.

II. LITERATURE REVIEW

Said BEN ALL et al. (2010) "HABRP an energy efficient protocol" have proposed an energy efficient routing protocol for heterogeneous wireless sensor network. In this, the author worked on reducing the failure factor of sensor nodes. In this, the author introduced protocol for prolong the time interval before the first nodes' death so that there will be an increase in the lifetime of an heterogeneous WSNs. Also, the protocol elected some high-energy nodes(NCG) as "cluster heads" to collect the information about the cluster members and forward this data to Gateways.

Alain Bertrand Bomgni et al. (2010) "An Energy-Efficient Clique-Based Geocast Algorithm for Dense Sensor Networks" proposes an algorithm which is clique based. This algorithm guarantees the delivery of packets sent by the sink node to all nodes which are deployed in different geocast regions. In this proposal, the suggested algorithm is a hybrid clustering scheme. In this scheme, firstly, the network is partitioned into cliques by using an existing energy-efficient protocol based on clustering. Secondly, the cluster-heads of cliques are divided into sets using an energy-efficient hierarchical clustering turn-by-turn. This approach consumes less energy due to which it comes into the category of energy-efficient clustering algorithm. In this algorithm, the CH is situated at some centre point of the cluster in the network. This approach uses less energy for transmission among cluster heads because of nature of clusters i.e. each cluster is a clique and each sensor is at one hop to the CH.

Geon Yong Park et. al. has worked on Wireless sensor network and proposed an efficient cluster head selection method using K-means algorithm to maximize the energy efficiency of wireless sensor network. Their results have shown that the proposed approach allows better performance than the existing hierarchical routing protocols such as LEACH and HEED in terms of network lifetime."

Amir Sepasi Zamati et. al. proposed an Energy Efficient Protocol with Static Clustering for Wireless Sensor Networks enhancing LEACH. It is a dynamic clustering protocol based on the LEACH. In this, node is chosen as CH temporary first of all, and then it helps in choosing the best CH for network and hence increases lifetime of network.

It has 3 phases:

- ***Setup Phase:*** In this scheme, the desired no. of clusters is set initially, say 'k'. BS sends k-1 messages to sensor nodes of network through different transmission channel. Now, nodes who hear message k=1 will take 1 as their cluster id and so on for all 'k' (k=1,2,3…. k-1). Nodes which do not join any cluster, they will set 'k' as their cluster ID and will inform to BS using CSMA for sending JOIN-REQUEST message. For sending data, nodes uses TDMA based schedule and they can only send data in their respective scheduled time slot. It reduces collision so that energy can be preserved which will enhance the network lifetime in return.
- ***Selection Phase:*** In this phase, all nodes sends their level of energy to temporary CH, then T-CH compares residual energy levels of all nodes and choose the node as CH for that particular cluster which has highest energy level and node as T-CH which has lowest energy level for next round.
- **Steady State Phase:** In this phase, all nodes sends their collected data to CH and CH receives data, aggregates data and computes data so that only the important data must be selected for transmission to BS directly.

III. CLUSTERING

One technique named 'Clustering' is very efficient in reducing traffic among network and base-station. Clustering saves energy by reducing energy consumption and also prolong lifetime of network. Small units called clusters are formed which are very easy to manage as compared to large whole network unit. Clustering improves scalability and bandwidth. It is an important factor upon which efficiency of network depends. Inter-clusters communication conserves bandwidth, reduces redundant data transfer. A part from these advantages, one issue of clustering scheme that cannot be neglect is

that is how to determine the optimal number of clusters so that energy consumption can be reduced. The proposed method decides the number of cluster for efficient clustering algorithm PAM.

- *Sensor Node:* Sensor node is the important component of WSN because of its has multiple roles. It senses data, stores data, routes data and processes data.
- *Clusters:* Clusters are small manageable units which performs tasks such a simplifying the communication.
- *Clusterheads:* Clusterheads are special nodes who act as the leader and also organize cluster activities. It collects data from several sensor nodes and then aggregates those data and also organizes the schedule of a cluster for communication with BS.
- *Base Station:* Base station is a central component which collects data from several nodes distributed at different locations. The deployment of base station is also a critical issue of WSN. It acts as an intermediate between the network and end-user.
- *End User:* End User is a vital component if any network. It can be a computer or a PDA which generates a query to sensor network over the internet in a particular application.

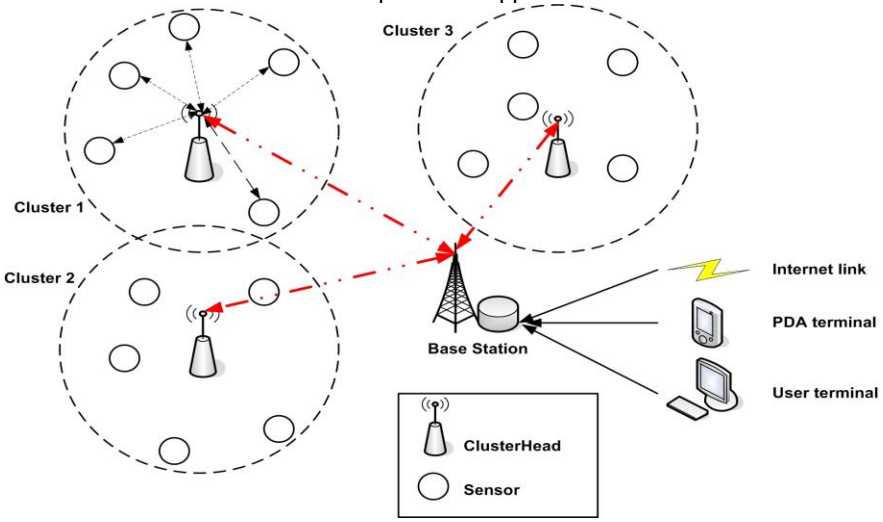

Fig 2. General Sensor Network Architecture

Clustering saves energy by reducing energy consumption and also prolong lifetime of network. Small units called clusters are formed which are very easy to manage as compared to large whole network unit. Clustering improves scalability and bandwidth. It is an important factor upon which efficiency of network depends. A part from these advantages, one issue of clustering scheme that cannot be neglect is that is how to determine the optimal number of clusters so that energy consumption can be reduced.

Two methods of partitioning based clustering are:

- k-means
- k- medoids.

The most common method of k-medoid clustering is PAM. PAM reduces sum of data points' dissimilarities to closets medoids. It is a robust clustering algorithm than K-means. PAM requires pre-determined value of 'k' as input as similar in k-means.

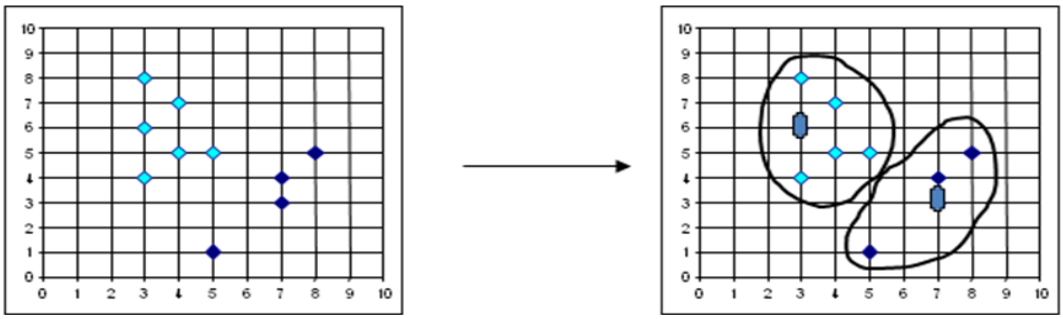

Fig 3. k-means and k-mediod

IV. PAM (K-MEDIODS)

PAM starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total cost and improves quality of clustering. K-Means algorithm has already been proved to be faster than K-Mediod algorithm in the cluster nodes. PAM algorithm is still the most balanced and most adaptive algorithm to most of the WSN scenarios. K-means is very sensitive to outliers, but K-mediod is not. Alike k-means, K-mediods takes most centrally located objects called medoids in a cluster.

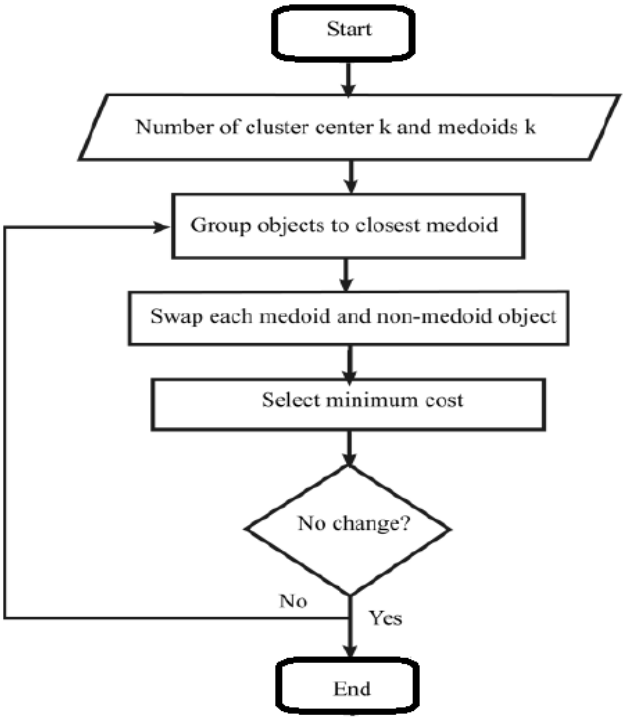

Fig 4. PAM Clustering Flow-chart

In this algorithm, initially 'k' value i.e. no. of clusters is inputted to algorithm. At every iteration step, algorithm determines is there any non-mediods which can be replace current mediod for improving cluster quality. Quality measuring function represent by sum of all non-mediod points within cluster of their mediod.

Algorithm 1: PAM Clustering

[1] Select ***k*** representative objects arbitrarily

[2] For each pair of non-selected object ***h*** and selected object ***i***, calculate the total swapping cost TC_{ih}

$$\textit{Total swapping cost } TC_{ih}=\sum_j C_{jih}$$

[3] For each pair of ***i*** and ***h***,

a. If $TC_{ih} < 0$, ***i*** is replaced by ***h***

b. Then assign each non-selected object to the most similar representative object

[4] repeat steps 2-3 until there is no change

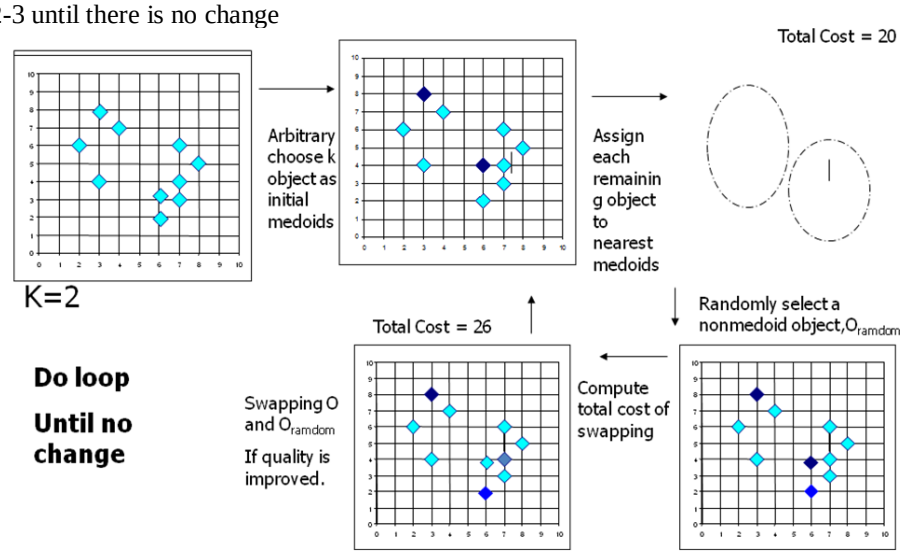

Fig 5. PAM algorithm flow-diagram

In this algorithm, initially 'k' value i.e. no. of clusters is inputted to algorithm. At every iteration step, algorithm determines is there any non-mediods which can be replace current mediod for improving cluster quality.
Quality measuring function represent by sum of all non-mediod points within cluster of their mediod. The K-medoids uses the Euclidean distance on the basis of each node's position coordinates (X & Y) in two way ground.
The K-medoids then computes the cost of swapping

$$\textit{Total swapping cost } TC_{ih} = \sum_j C_{jih}$$

Case 1. If total cost of swapping is less than 0, swapping is possible and '***i***' will be replaced or swapped with '***h***'.

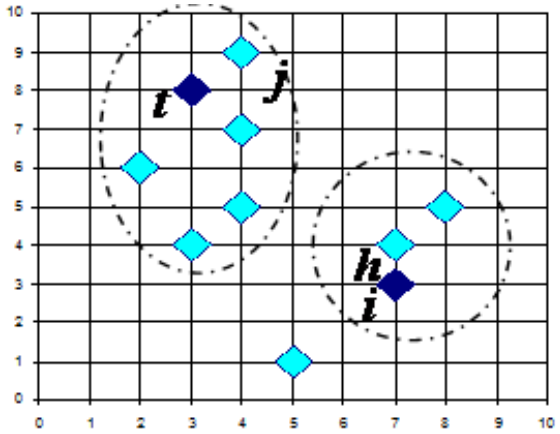

Fig 6. Case I. When Cjih = 0

Case 2. Else new non-selected object will be selected as 'h'and again swapping cost calculated for new '***h***' and ***'i'.***

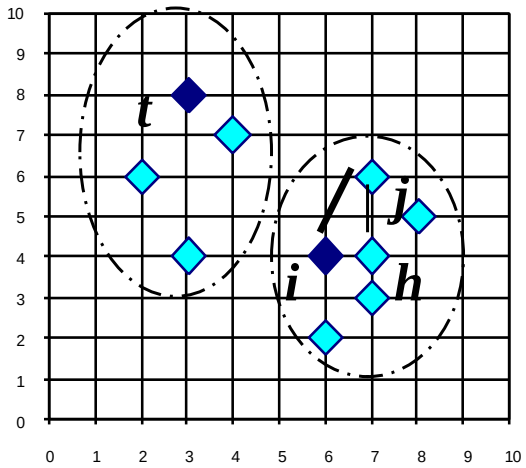

Fig 7. Case II When Cjih = d(j, h) – d(j, i)

V. PROPOSED ALGORITHM

The proposed method decides the number of cluster for efficient clustering algorithm PAM. In a network, it is very hard to find small clusters when there are larger clusters present near and far distant places. And it becomes difficult to find best 'k' value when there are many overlapping clusters available. Hence, it is very difficult to find 'k' value. PAM clustering technique is used to reduce the sum of data points' dissimilarities and for this purpose PAM search for possible medoids 'k' among data points and allocate each data point to a nearest medoids. PAM requires pre-determined value of 'k' as input for cluster analysis. Many methods for estimation of 'k' value have been present only focusing on global structure of cluster but no method focus on separation of overlapping clusters. Hence the wrong estimation of 'k' value is possible. To avoid this situation of overlapping between small and large clusters, we propose a method to evaluate no. of clusters based on focusing is it better to merge or separate clusters nearby or overlapping. The proposed method as been compared with that of existing PAM algorithm through simulation experiments.

Algorithm 1: Elbow Method to determine K

[1] Initialize $k=1$
[2] Start
[3] Increment the value of $k
[4] Measure the cost of the optimal quality solution
[5] If at some point the cost of the solution drops dramatically
[6] That's the true k.
[7] End

Proposed MethodAlgorithm

1. Start PAM()
{
Select ***k*** representative objects arbitrarily
For each pair of non-selected object ***h*** and selected object ***i***, calculate the total swapping cost TC_{ih}
Total swapping cost $TC_{ih}=\sum_j C_{jih}$
For each pair of ***i*** and ***h***,
If $TC_{ih} < 0$, ***i*** is replaced by ***h***

Then assign each non-selected object to the most similar representative object
repeat steps 2-3 until there is no change
Output: k cluster solutions.
}

2. For every k=1,2,3,4,5,6……..k
3. Calculate energies
4. Measure energies of the solution
5. If Nep(k)>Nem(k)
6. {
7. Increment the value of k
}
8. Else
{
If at some point the energies of the solution drops dramatically
That's the true k
}
9. End.

The algorithm reached optimal 'k' value such that the distance between border of clusters is minimized, it helps to reduce transmission cost. However, as the value of 'k' increases, we have reach an optimal 'k' where the merge energy is minimal. Any further increase in the 'k'resulted in higher energy.

VI. RESULTS AND DISCUSSIONS

The proposed method decides the optimal number of cluster for PAM algorithm. It works on studying on stability of clusters whether it is possible to separate them or merge them to find best 'k'. PAM algorithm is used for division of data into best 'k' clusters. It finds small clusters those are overlapping with larger clusters. Our method will analyzes to find optimal 'k' value by comparing energies of clusters at each 'k' value starting from 2, then 3 and so on. In our method, average data points distance and energies among clusters are calculated for every k, starting with k=2 and if it does not satisfy the condition then there will increase in 'k' and new incremented 'k' will apply. Our method analyzes separated clusters on the basis of their border distances and overlapping clusters on the basis of their data points distances. Average data point distance is calculated between all data points in overlapping clusters. Inter-clusters distances are also measured.

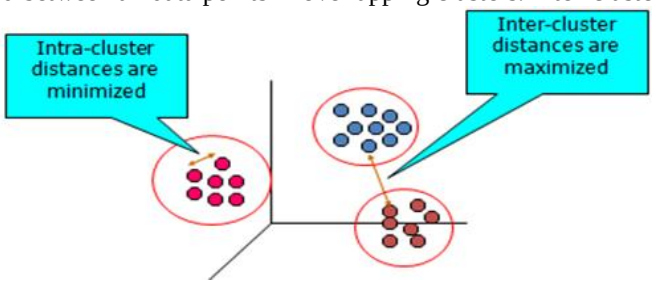

Fig. 8: Inter-cluster and Intra-cluster analyses

The partition energies Nep(k) and merge energies Nem(k) for each 'k' are calculated. Euclidean distances and under Pearson correlation coefficients are used to calculate distances among data points. To find best 'k' value we compare Nep(k) and Nem(k) at all 'k' value (k=1,2,3,4…….). Error rates are analysed while finding true value 'k'. In this paper, we are proposing an effective and efficient 'k' for defining no. of clusters using for k-medoid clustering algorithm(PAM). We use K-mediods because it is better than k-means as in k-mediods cluster must be in center but in k-means the centers could anywhere. k-Medoid is more robust to outliners than k-Means and has high computation complexity but results in more quality clustering. Computer simulation will be performed in the NS2 environment. In our method, we use energy-evaluation functions Nep(k) and Nem(k). DIS_b stands for distance between two clusters' border. Nep(k) stands for the partition energy and Nem(k) stands for the merge energy. The partition energies Nep(k) and merge energies Nem(k) for each 'k' are calculated. Clusters can be separate when DIS_b >0 and Nep(k) – Nem(k) >0. Clusters cannot be separate when DIS_b <=0 and Nep(k) – Nem(k) <=0. Euclidean distances and under Pearson correlation coefficients are used to calculate distances among data points. To find best 'k' value we compare Nep(k) and Nem(k) at all 'k' value (k=1,2,3,4…….). Error rates are analysed while finding true value 'k'.

TABLE I

K	*Nep*	*Nem*
2	0.3288	0.2559
3	0.2942	0.2346
4	0.1296	0.1869
5	0.1844	0.4170
6	0.1877	0.3227

7	0.2037	0.4239
8	0.2037	0.4239
9	0.0901	0.2636
10	0.0901	0.2636

In this paper, we are proposing an effective and efficient 'k' for defining no. of clusters using for k-medoid clustering algorithm(PAM). We use K-mediods because it is better than k-means as in k-mediods cluster must be in center but in k-means the centers could anywhere. k-Medoid is more robust to outliners than k-Means and has high computation complexity but results in more quality clustering. Computer simulation will be performed in the NS2 environment.

Simulations are done in MATLAB to find correct 'k' and give energies for partition clusters and merge clusters. K-mediod algorithm is adopted to evaluate value of 'k' on the basis of which whole dataset is divides into clusters. Our incremental method then processes for correct 'k' value on output solution of k-mediod. MATLAB have many build-in method like statistical tool box, which facilitate the calculation of various types of statistical errors to measure the cluster quality. Our method is compared with build in methods Silhouette index, Davies-Bouldin Index (DB index), GAP statistics. Tables I shows emerge and partition energies for every k value.

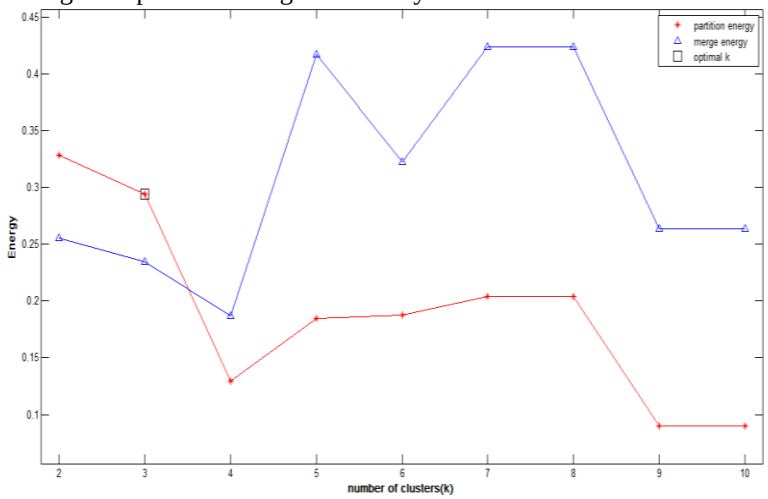

Fig 9. Results showing k=3 as best value for 'k'

\

In figure 9, we show results on dataset 'leukemia'. Firstly, PAM is run and it divides whole dataset into 'k' clusters. For every value of 'k' clustering solution is obtained; partition energies (Nep) and merging energies (Nem) are calculated with obtained clustering solution for every 'k'. After this, our method compares Nep and Nem for every 'k' and then finalize the best value for 'k'. As in figure 3, Nep > Nem at k=3, when value of 'k' increment i.e k=4, then Nem become more than Nep (Nem>Nep), so it means according to our method best value of 'k' is 3. The algorithm reached optimal 'k' value such that the distance between border of clusters is minimized, it helps to reduce transmission cost. However, as the value of 'k' increases, we have reach an optimal 'k' where the merge energy is minimal. Any further increase in the 'k'resulted in higher energy.

VII. CONCLUSIONS

As energy-awareness is very crucial in wsn, we reached at the optimal number of clusters for PAM. However, deciding no. of clusters is an important problem. Many clustering algorithms usually take this input from user. Our proposed method generate no. of clusters that are result better output. It calculate distances between datapoints within overlapping clusters and also distance for separated clusters and give best value of 'k' by focussing on separability degree using PAM algorithm. has been proved to be enough efficient to make the whole process energy efficient and improves the network lifetime. The scheme can be improved to produce more quality clusters with much better cluster number estimation algorithm.

ACKNOWLEDGMENT

I thanks to my respectful guide Asst. Prof. Arvind Kumar who have contributed towards development of the paper.

REFERENCES

[1] Geon Yong Park, Heeseong Kim, Hwi Woon Jeong, and Hee Yong Youn, "A Novel Cluster Head Selection Method based on K-Means Algorithm for Energy Efficient Wireless Sensor Network", 2013 27th International Conference on Advanced Information Networking and Applications Workshops, 20130325.

[2] Heinzelman, W., Chandrakasan, A., and Balakrishnan, H., "Energy-Efficient Communication Protocols for Wireless Microsensor Networks", Proceedings of the 33rd Hawaaian International Conference on Systems Science (HICSS), January 2000.

[3] John M. Shea, Joseph P. Macker, "Automatic Selection of Number of Clusters in Networks using Relative Eigenvalue Quality, IMCC, vol. 1, pp. 131-136, IEEE 2013.

[4] Study on WSN Topology Division and Lifetime XU Jiu-qiang, WANG Hong-chuan,LANG Feng gao,WANG Ping,HOU Zhen-peng, IEEE, 2011

[5] Hierarchical Adaptive Balanced energy efficient Routing Protocol (HABRP) for heterogeneous wireless sensor networks BY Said BEN ALL*, Abdellah EZZATI, Abderrahim BENI HSSANE, Moulay Lahcen HASNAOUI, IEEE, 2010

[6] Raymond Wagner, Shriram Sarvotham, Hyeokho Choi, Richard Baraniuk, "Distributed Multiscale Data Analysis and Processing For Sensor Networks", Rice University Technical Report, February 9, 2005.

[7] Scott Briles, Joseph Arrowood, Dakx Turcotte, Etienne Fiset, "Hardware-In-The-Loop Demonstration of a Radio Frequency Geolocation Algorithm", Proceedings of the Mathworks International Aerospace and Defense Conference, May 24-25, 2005.

[8] Sundeep Pattem, Bhaskar Krishnamachari, and Ramesh Govindan, "The Impact of Spatial Correlation on Routing with Compression in Wireless Sensor Networks," ACM/IEEE International Symposium on Information Processing in Sensor Networks (IPSN), April 26-27, Berkeley, CA 2004.

[9] Dan Pelleg and Andrew Moore. X-means: Extending K-means with efficient estimation of the number of clusters. In Proceedings of the 17th International Conf. on Machine Learning, pages 727–734. Morgan Kaufmann, San Francisco, CA, 2000.

