Improving K-Means Algorithm by Grid-Density Clustering for Distributed WSN Data Stream

Yassmeen Alghamdi[1], Manal Abdullah[2]
Faculty of Computing and Information Technology, Department of Computer Science
King Abdul-Aziz University, KAU
Jeddah, Saudi Arabia

***Abstract*—At recent years, Wireless Sensor Networks (WSNs) had a widespread range of applications in many fields related to military surveillance, monitoring health, observing habitat and so on. WSNs contain individual nodes that interact with the environment by sensing and processing physical parameters. Sometimes, sensor nodes generate a big amount of sequential tuple-oriented and small data that is called Data Streams. Data streams usually are huge data that arrive online, flowing rapidly in a very high speed, unlimited and can't be controlled orderly during arrival. Due to WSN limitations, some challenges are faced and need to be solved. Extending network lifetime and reducing energy consumption are main challenges that could be solved by Data Mining techniques. Clustering is a common data mining technique that effectively organizes WSNs structure. It has proven its efficiency on network performance by extending network lifetime and saving energy of sensor nodes. This paper develops a grid-density clustering algorithm that enhances clustering in WSNs by combining grid and density techniques. The algorithm helps to face limitations found in WSNs that carry data streams. Grid-density algorithm is proposed based on the well-Known K-Means clustering algorithm to enhance it. By using Matlab, the grid-density clustering algorithm is compared with K-Means algorithm. The simulation results prove that the grid-density algorithm outperforms K-Means by 15% in network lifetime and by 13% in energy consumption.**

Keywords*—WSNs; data mining; clustering; data stream; grid density

I. INTRODUCTION

In recent years, a widespread use of WSNs are found in various applications. A WSN is a specific kind of ad-hoc networks that is able to sense and process information. They can be used in many areas such as environmental, industrial, military, and agriculture fields. WSNs consist of built-in, independent and tiny equipment's called sensor nodes. Sensor nodes contain four main components: energy source, processing unit, sensing [1] unit and transducer. Sensor nodes are mainly used in processing data and report parameters continuously. The reports are transferred by sensor nodes and collected by special controllers called Base Stations (BSs). A WSN has many resource constraints including high computational power and limited energy source. WSNs depend on their nodes that consumes battery energy. Unfortunately, the WSNs nature makes it difficult to recharge sensor nodes batteries. Therefore, energy efficiency is an important design objective in WSNs [2] and their algorithms should be precisely designed based on energy saving.

In some WSN applications, data that WSNs process usually contain a large amount of datasets that flow rapidly in a very high speed and arrive online. Data are considered to be unlimited and the arriving order of elements being processed is out of control. Such data are called Data Streams [1, 3].

The widespread deployment of WSNs and the need for aggregating data streams requires an efficient organization of network topology to reach load balancing and extension of network lifetime. This is performed by using mining techniques. Clustering is a data mining technique that is considered to be an efficient tool in WSNs to solve the problem of network lifetime, energy consumption, data aggregation, load balancing, scalability [4], delay and delivering data stream packets. It organizes WSNs into a connected hierarchy. In general, two categories of network structure are found in WSNs, flat and clustered (i.e. hierarchical) [5]. At any case, clustering plays an important part in network organization, and also affects network performance. To reach many advantages, clustering is preferred in mining WSNs data. In a clustered WSN, the network is divided into groups called clusters, each cluster has a leader elected from sensor nodes called Cluster Head (CH). Data streams are aggregated from nodes by their CH inside a cluster. Then it is transmitted from CHs to the BS. Transmitting streaming data in the wireless environment by a multi-hop communication to reach the BS resumes nodes energy leading to shorten the lifetime of a network.

As mentioned previously, a WSN suffers from power consumption during data stream transmission. Sensor nodes should be energy efficient. Energy efficiency affects the entire WSN lifetime. Therefore, to gain WSN's operational long-lasting, consuming energy is considered during designing WSNs algorithms. Furthermore, since sensor nodes are in difficult-to-reach locations, replacing batteries is unpractical. A WSN can achieve energy saving from clustering algorithms. However, to achieve better energy conservation, data stream mining [6] must be formed in a distributed manner, due to their resource constraints.

Clustering algorithms are designed to obtain load-distribution between CHs, high connectivity, saving energy and fault tolerance. In WSNs, using resources and reducing energy is provided in clustering technique by decreasing number of nodes that transmit data streams through long distance transmission. Clustered WSN algorithms running streaming data are usually partitioned in two main steps, cluster formation step and data transmission step [7]. But specifically, the cluster-based operation of clustered WSN algorithms consist of

rounds. Rounds involve cluster creation, CH-Election, and data transmission.

Grid-based clustered WSNs, are type of networks [4] where a sensed area is partitioned into a number of equally sized small cells called grids. Grid-based clustering scheme has proven to have a fast processing time compared to other types of clustering algorithm schemes due to computational operations are performed on grid cells instead of the whole dataset stream.

This paper develops a distributed clustering algorithm for WSNs based on the well-known K-Means to enhance it. The algorithm is based on combining a grid technique and a density technique. Besides clustering advantages, density technique can find arbitrary shaped clusters with noise, while grid technique is used to avoid clustering quality problems by discarding the boundary nodes of grids. This combination of techniques decreases algorithm computational time, reduce energy consumption and thus increasing network lifetime resulting desirable simulation results. To reach this paper aims, the algorithm must converge the limited dataset streams as fast as possible, to ensure that a processor can take on next set of streams. The paper provides an evaluation of our grid-density clustering algorithm to prove its efficiency by comparing its final results with K-Means results. The remaining of this paper is ordered as follows. Literature review on clustering algorithms is provided in section II. An overview on K-Means clustering algorithm is given in section III. The proposed grid-density clustering algorithm is explained in section IV. Simulation analysis and results are discussed in section V. Concluding is given in section VI.

II. Literature Review

Some clustering algorithms are found in WSNs process traditional sensed data. Based on network structure, algorithms found in WSNs can be divided into two classes: algorithms for either flat networks or hierarchical networks. In a flat network structure, all nodes have the same tasks and perform exact functionalities. Data transmission is done in a hop-by-hop manner using flooding. Some clustering algorithms in flat WSNs include Energy-Aware Routing (EAR), Gradient-Based Routing (GBR), Sequential Assignment Routing (SAR) and many more. These clustering protocols are efficient in networks with a small-scale. However, flat WSN algorithms are unfavorable in networks with large-scale due to resource limitation, but nodes such networks preform more data processing [5]. In a hierarchical structure, nodes have different functions and are organized into groups according to specific requirements or metrics. Generally, each cluster has a specific CH and other sensor nodes. CHs have the highest energy inside clusters to perform processing and transferring data, while other nodes with low energy perform sensing [5]. Some clustering algorithms in a hierarchical WSN topology include Low-energy Adaptive Clustering Hierarchy (LEACH), Energy-Efficient Uneven Clustering (EEUC) algorithm, Algorithm for Cluster Establishment (ACE). Clustering technique is an important scheme in hierarchical WSNs due to many advantages, such as data aggregation [5], scalability, less load, low energy consumption and more robustness.

Other clustering algorithms stream data stream in other environments rather than WSNs. In 2006, Feng Cao proposed the DenStream algorithm for clustering dynamic data stream [8]. It is an effective method that can discover clusters with arbitrary format in data streams, but it is insensitive to noise [9]. Heng Zhu Wei proposed a density and space clustering algorithm called CluStream [8]. CluStream is a clustering data stream algorithm based on K-Means that is inefficient to get clusters of arbitrary formats and cannot process outliers. Further, they have to predetermine a parameter K (i.e. number of clusters) [10].

K-Means is used in an offline phase of some algorithms such as Clustream. It is a divide and conquer schemes that partition data streams into segments and discover clusters in data streams. K-Means has a number of limitations. Firstly, it doesn't reveal clusters with arbitrary formats and usually identifying spherical clusters. Secondly, it is unable to discover outliers and noise. Thirdly, K-Means requires multiple scans of data, making it impractical for huge data stream [8, 10]. STREAM and CluStream are data stream clustering algorithms that are extensions of K-Means [11].

LOCALSEARCH, STREAM, DenStream and CluStream are clustering algorithms involving data streams. They disregard grid border problems. Data streams come with a large number in chronological order, and makes original grids no longer adapt to new data mapping, so a large number of data is likely to fall on grids borders. But if the data is simply discarded, it affects the clustering quality. If grids are updated in time, cost is greatly increased and the clustering efficiency is affected greatly [8].

D-Stream is a real-time density-grid stream data clustering algorithm where nodes are assigned to grids and grids are gathered to form clusters based on their density. D-Stream clustering quality depends on the lowest grid structure level. This may reduce the clusters accuracy despite the technique processing time speed [11]. D-Stream assigns input data into grids by using an online component. It also computes density of each grid and performs clusters based on their density by using an offline component [10]. MR-Stream is an algorithm that can cluster data streams at various resolutions. It divides a given data space to cells and a data structure tree that keeps the space dividing. MR-Stream increases the clustering performance by determining the exact time to generate clusters [11]. FlockStream is a density clustering algorithm that is based on the concept of bio-inspired model. It uses the flocking model, where independent micro-cluster agents form clusters together. FlockStream combines online and offline components where agents form clusters once required. It can get clustering results without performing offline clustering. DenStream, MR-Stream, D-Stream and FlockStream are density-based clustering algorithms carrying data streams. They can affectively reveal clusters with arbitrary shapes and handle noise, but their quality decrease when using clusters with variant densities. LOCALSEARCH algorithm [8] uses dividing and conquering to partition data streams into segments, and discovers clustering of data streams in finite space, by using the K-Means algorithm.

A framework to dynamically cluster multiple evolving data streams called Clustering on Demand (COD) was proposed [12]. It produces a summary hierarchy of data statistics in the online phase, whereas clustering is performed in the offline phase [12]. It summarizes data streams using the Discrete Fourier Transform (DFT) technique. Then it applies a K-Means algorithm to cluster the summarized data streams [12]. An Online Divisive-Agglomerative Clustering (ODAC) algorithm was also proposed to incrementally construct a tree-like hierarchy of clusters using a top-down strategy. The previous techniques assume that all data streams are gathered at a centralized site before they are processed [12]. Many density-based clustering algorithms for multi density datasets are not suitable for data stream environments. First, they need two-pass of data and this condition is impossible for data streams where they arrive continuously and need a single scan to be performed. GMDBSCAN and ISDBSCAN use a two-pass data. Second, some multi-density clustering algorithms require using the whole data. Third, other algorithms have a high execution time which makes them unsuitable when applying data streams. DSCLU [11] is considered to be a density clustering algorithm run streaming data in multi-density environments.

Another class of clustering algorithm is when applying data streams on WSNs. It is divided into two subsections [13]: algorithms based on Fuzzy clustering in WSNs and algorithms based on multimedia streaming data streams found in Multimedia Wireless Sensor Networks (MWSNs). Our proposed density-grid clustering algorithm is similar to research study scope of algorithms belongs under this class. Fuzzy C-Means or Fuzzy Clustering Means (FCM), is a widely used data stream mining clustering algorithm in WSNs. Most clustering algorithms are descendant from FCM when solving data stream problems in WSNs. FCM requires prior information about how many needed clusters C to divide the data space. The clusters number C is unknown previously.

An algorithm based on FCM, a distributed WSN data stream clustering algorithm called SUBFCM (Subtractive Fuzzy Cluster Means) is proposed to decrease nodes energy consumption and extend network lifetime in WSNs involving data streams. The SUBFCM focuses on the clustering problem on data streams. Simulations show that the energy efficient algorithm SUBFCM can obtain clustering with less energy than the FCM. SUBFCM reduces the overall data transmission needed without affecting vital information in data streams [13]. SUBFCM is a result of blending a subtractive clustering algorithm with the FCM.

For algorithms based on multimedia streaming data, in WMSNs, multimedia clustering protocols use the quality of service (QoS) parameters [14]. The requirements of QoS differ based on to the multimedia applications type. QoS has several metrics such as jitter, delay, bandwidth, reliability [15] and packet loss [14]. A lot of multimedia applications are time critical, they require to be managed with a limited time. Sensors in multimedia are able to grab image, audio, video, and so on. Then send \ multimedia content by sensors [15]. FoVs is a wireless multimedia sensor network clustering algorithm proposed based on Overlapped Field of View (FoV) areas. FoVs prolongs network lifetime and saves energy [16].

III. K-Means Algorithm

The simplest algorithm that solve a well-known clustering problem is called the K-Means clustering algorithm. K-Means has an efficient CH selection method to maximize energy efficiency of a WSN. K-Means is based on finding a CH that minimizes the sum of Euclidean distances between CH and nodes [17, 18]. It reduces communication overhead, energy consumption and extends network lifetime. It is used to partition a sensed area into K clusters. The procedure follows a simple way to classify a given dataset through a certain number of clusters fixed a priori [18]. In K-means, there is a distance threshold called R for calculating distance between CH and BS. If their distance is less than R, they use a single-hop transmission, otherwise, they use a multiple-hops transmission [7]. There is also an energy threshold called E for all CHs. If CH energy is less than E, then CH broadcasts a quit message to all nodes inside the cluster. Hence, other nodes which have higher residual energy are elected to become CHs [7]. Nodes near boundary region in K-Means are affected since, the degree of belongingness is described in terms of either zero or one. For this reason, K-Means clustering is called hard clustering. Edge nodes may have the same degree of belongingness to more than one clusters. In K-Means, there is an optimal cluster formation. Nodes are assigned to a cluster based on the degree of belongingness when network area deployment. Degree of belongingness needs to be computed in each round for every node inside a cluster. Obviously, the major limitation of K-Means clustering algorithm is predetermining parameter K [6]. The traditional K-Mean algorithm in figure 1.

K − Means Algorithm

Input: K, n

1- Set (k) as number of clusters.

2- Set (n) as total number of nodes.

3- Initialize value (k) to estimate each cluster centroid.

4- Assign each (n) to its cluster whose centroid has nearest distance.

5- Centroid recalculation (k) after assigning each node to its cluster.

6- Repeat (3) and (4) until

6.1 no node changes its cluster assignment.

6.2 or until (k) no longer moves.

Fig. 1. K-Means Algorithm Pseudo Code.

IV. Grid-Density based Clustering Algorithm

Clustering WSN algorithms could be considered under specific schemes as shown in figure 2, such as, hierarchical scheme, grid scheme, heuristic scheme, weighted scheme, PSO-Based scheme. The developed grid-density algorithm is a grid scheme. Figure 2 summarizes clustering schemes in WSNs with an example of each clustering scheme. The grid-density clustering algorithm is a clustering algorithm that forms clusters based on density of each grid in a gridded WSN. Grid-density algorithm is proposed based on K-Means to enhance it.

It solves the same problems that K-Means clustering algorithms solve. But grid-density clustering algorithm forms clusters in a different manner, where cluster formation is not based on the Euclidian distance calculation. Cluster formation is based on finding the density of each grid. Additionally, it doesn't require a predetermination of clusters number as required in K-Means.

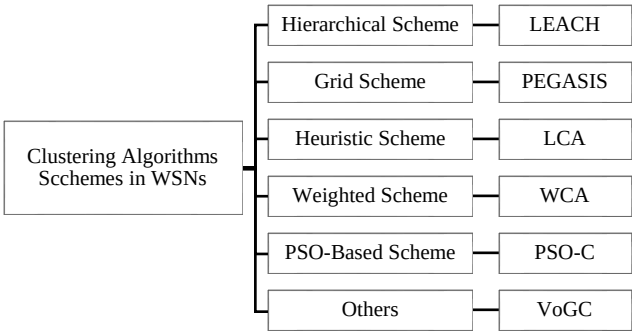
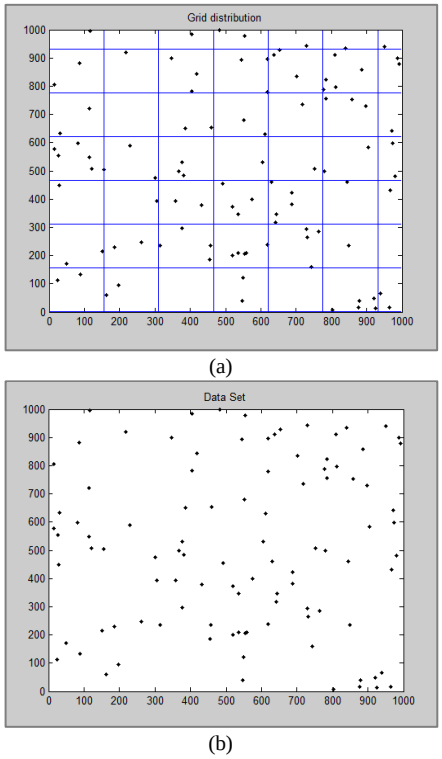


Fig. 2. Clustering Algorithms Schemes in WSNs.

To form network clusters, the developed scheme is done by dividing a sensor network area into equal size of grids. The area is divided by a value called grid size g where $\boldsymbol{g} \in \boldsymbol{X}$, and $\boldsymbol{g} \in \boldsymbol{Y}$. Grids then are classified based on their densities by using a specific value called threshold $\boldsymbol{\sigma}$. By using the suggested algorithm and both values $\boldsymbol{g}$ and $\boldsymbol{\sigma}$, grids close to each other are combined after finding their density to form arbitrary shaped clusters. Empty grids are used as delimiters to reduce the algorithm execution time. Cluster formation process in the grid-density algorithm is based on $\boldsymbol{g}$ and $\boldsymbol{\sigma}$ to find number of clusters $\boldsymbol{C}$. After forming clusters, the grid-density selects a CH for each cluster based on nearest distance to the BS. Figure 3 (a) shows a sensed area that is already divided into grids assuming that $\boldsymbol{g = 155}$ by using the grid-density algorithm. In contrast, figure 3 (b) presents the same sensed area in K-Means that is not gridded.



(a)

(b)

Fig. 3. (a) Gridded WSN in Grid-Density Clustering Algorithm, (b) Sensed Area in K-Means.

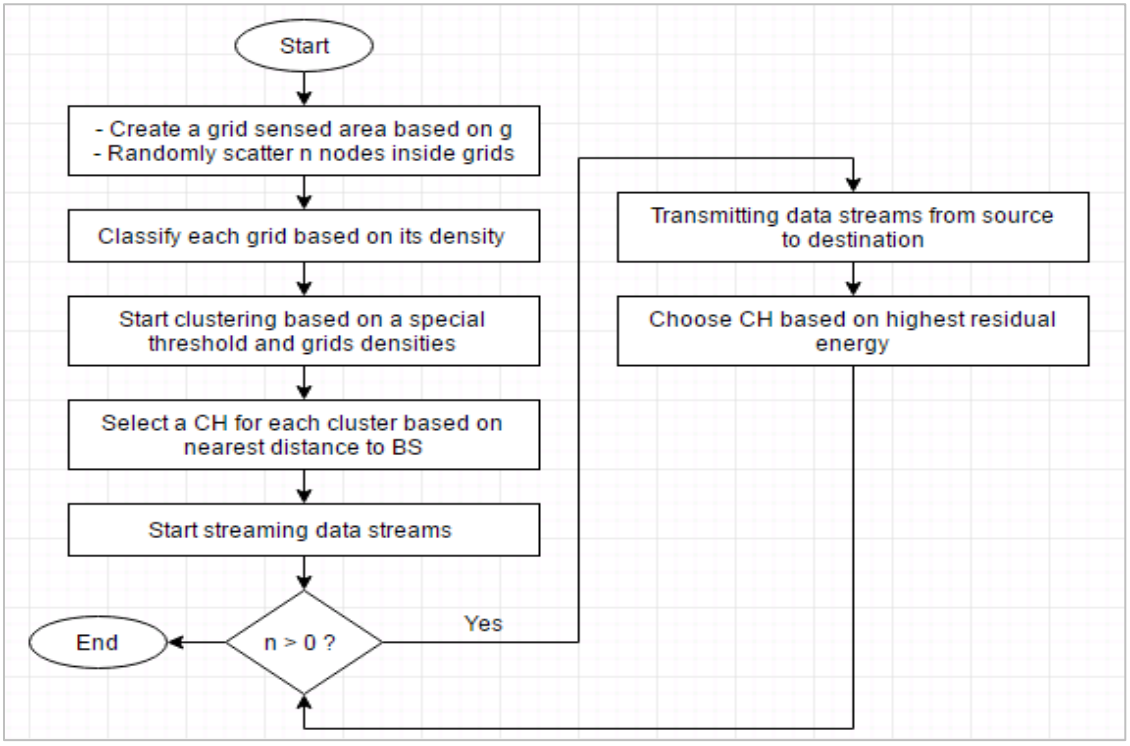
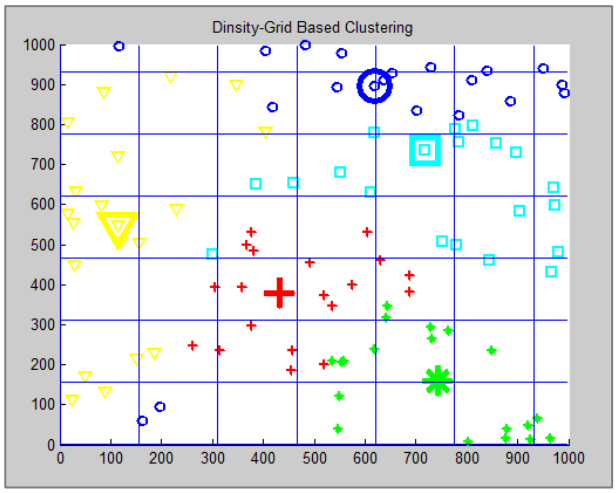


Fig. 4. Grid-Density Clustering Algorithm Flowchar.

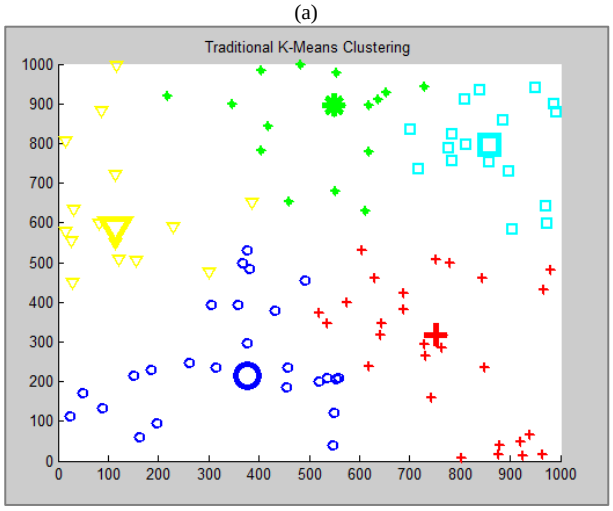
After forming network clusters and choosing CHs initially, the network is ready to stream data. To stream data, the algorithm goes through several rounds until the end of network lifetime. Each round consists of two steps, transmitting data and choosing CHs. First step is responsible for transmitting sensed streaming data from source nodes to the final destination at the BS through CHs. The second step is to rotate the role between CHs based on nodes highest residual energy. The procedure used to obtain the final experimental results for both competitors is by running the grid-density algorithm first to form its clusters, then gain number of clusters $\boldsymbol{C}$. After that, the traditional K-Means is run individually using the predetermined value $\boldsymbol{C}$ gained from the grid-density algorithm. Comparisons between competitors is done based in network lifetime and energy consumption for the entire network. Figure 4 represents grid-density algorithm flowchart.

V. Simulation Analysis and Results

To implement and evaluate the grid-density algorithm, Matlab version R2008b was used. In our research experiments, Matlab is used in a machine with Windows 7 Service Pack 1 with 1TB disk space, 64-bit operating system, Intel ® Core ™ i7 processor and 8GB RAM.



(a)

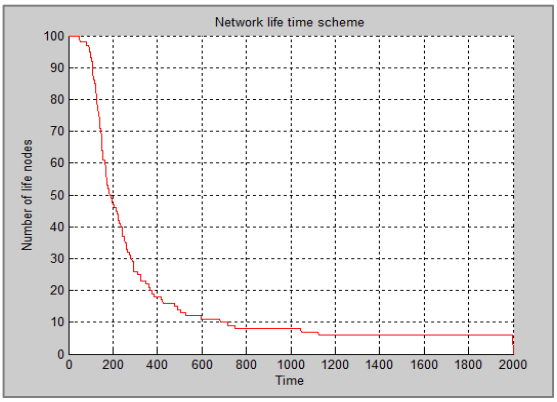


(b)

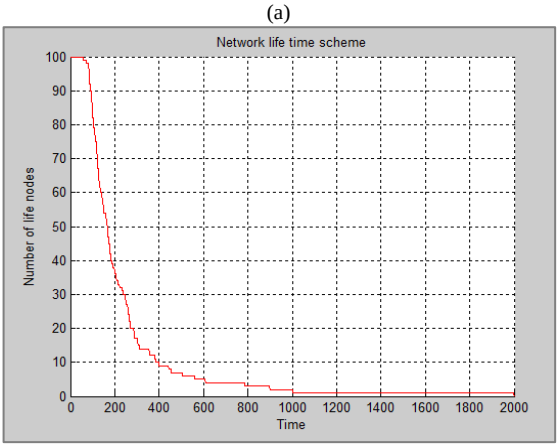
Fig. 5. (a) Cluster Formation in Grid-Density Clustering Algorithm, (b) cLUSTER FORMATIOn in K-Means.

After several simulation experiment, the following experiment has the best results and is chosen to compare between the two competitors' final performance metrics results in terms of network lifetime and energy consumption. Assuming that grid size $\boldsymbol{g} = 155$ and threshold $\boldsymbol{\sigma} = 5$. Both algorithms in their experiment are streaming the same dataset stream packet with size 126 byte/message in a (1000×1000) m^2 sensed area and BS located at the center, with $\boldsymbol{n} = 100$ node scattered randomly each with an initial energy equal to $1 Joule$. At cluster formation process in the grid-density algorithm, clusters number obtain from this experiment is $\boldsymbol{C} = 5$. Figure 5 (a) represents five individual clusters formed in the grid-density algorithm, each with a clear CH. By using number of clusters $\boldsymbol{C} = 5$, gained from the grid-density algorithm, then applying $\boldsymbol{C}$ as an input for K-Means on the given sensed area, figure 5 (b) shows the cluster formation result in K-Means each with a clear CH. It is clear from figure 5 that the sensed area is gridded in the grid-density algorithm while not gridded in K-Means.

In figure 6 (a) and (b), both graphs present network lifetime for grid-density algorithm and K-Means consequently. X-axis presents the running time in seconds (sec.) and Y-axis presents number of live nodes. It is clearly shown that the proposed algorithm extends network lifetime more than K-Means, where K-Means nodes stats to die at 1000 sec before the proposed algorithm nodes. From experimental results, the grid-density algorithm extends network lifetime by about 15% more than K-Means.

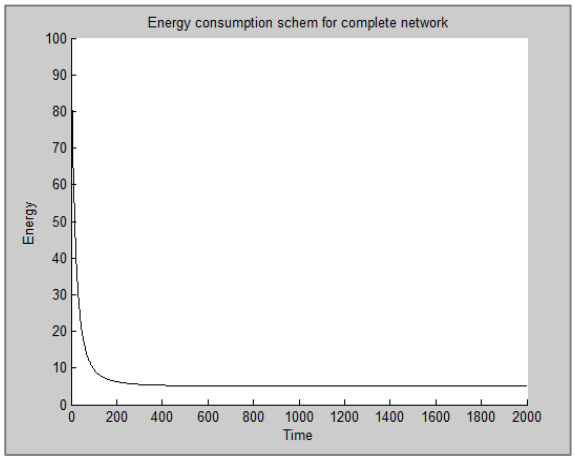


(a)

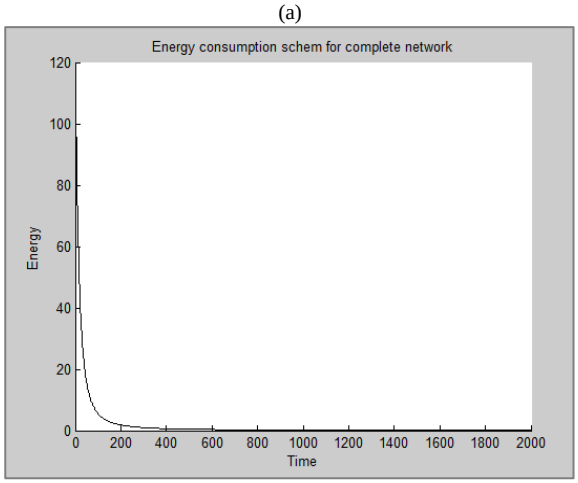


(b)

Fig. 6. (a) Network Lifetime in Grid-Density Clustering Algorithm, (b) Network Lifetime in K-Means.



(a)



(b)

Fig. 7. (a) Energy Consumption in Grid-Density Clustering Algorithm, (b) Energy Consumption in K-Means.

Figure 7 (a) and (b) present graphs of grid-density algorithm and K-Means consequently for their energy consumption. X-axis presents the energy in percentage in Joule while Y-axis presents the running time in Second (sec.). It is found that the proposed algorithm reduces energy consumption by about 13% less than K-Means.

The grid-density algorithm processes small grids, were all operations are performed on grid cells rather than processing the whole sensed area space and exhaustion the network as found in K-Means. Gridding reduces network exhaustion, thus reduces energy consumption that in turn extends network lifetime.

VI. Conclusion

Recent years witnessed a wide use of WSNs in several applications and it has become an interesting research area of study in data mining field. This paper provided an overview on clustering algorithms. It proposed a WSN clustering algorithm involving data streams based on the well-known K-Means. The proposed grid-density clustering algorithm is based on the concept of finding the density of each grid to form clusters in a WSN. The grid-density clustering algorithm results are regarding some performance metrics that are compared with K-Means algorithm results. Simulation results prove that the grid-density algorithm outperforms K-Means by 15% in network lifetime and by 13% in energy consumption performance metrics.

References

[1] A. L. de Aquino, C. M. S. Figueiredo, E. F. Nakamura, L. S. Buriol, A. Loureiro, A. O. Fernandes, et al., "A sampling data stream algorithm for wireless sensor networks," in Communications, 2007. ICC'07. IEEE International Conference on, 2007, pp. 3207-3212.

[2] O. Younis, M. Krunz, and S. Ramasubramanian, "Node clustering in wireless sensor networks: recent developments and deployment challenges," Network, IEEE, vol. 20, pp. 20-25, 2006.

[3] A. L. de Aquino, C. M. Figueiredo, and E. F. Nakamura, "Data Stream Algorithms For Processing of Wireless Sensor Network Application Data."

[4] M. Abdullah, H. N. Eldin, T. Al-Moshadak, R. Alshaik, and I. Al-Anesi, "Density Grid-Based Clustering for Wireless Sensors Networks," Procedia Computer Science, vol. 65, pp. 35-47, 2015.

[5] X. Liu, "A survey on clustering routing protocols in wireless sensor networks," sensors, vol. 12, pp. 11113-11153, 2012.

[6] H. Sabit and A. Al-Anbuky, "Multivariate spatial condition mapping using subtractive fuzzy cluster means," Sensors, vol. 14, pp. 18960-18981, 2014.

[7] S. Zhong, G. Wang, X. Leng, X. Wang, L. Xue, and Y. Gu, "A Low Energy Consumption Clustering Routing Protocol Based on K-Means," Journal of Software Engineering and Applications, vol. 5, p. 1013, 2012.

[8] C. Jia, C. Tan, and A. Yong, "A grid and density-based clustering algorithm for processing data stream," in Genetic and Evolutionary Computing, 2008. WGEC'08. Second International Conference on, 2008, pp. 517-521.

[9] F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-Based Clustering over an Evolving Data Stream with Noise," in SDM, 2006, pp. 328-339.

[10] Y. Chen and L. Tu, "Density-based clustering for real-time stream data," in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, pp. 133-142.

[11] A. Amini, H. Saboohi, and T. Y. Wah, "A multi density-based clustering algorithm for data stream with noise," in Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on, 2013, pp. 1105-1112.

[12] J. Yin and M. M. Gaber, "Clustering distributed time series in sensor networks," in Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, 2008, pp. 678-687.

[13] H. Sabit, A. Al-Anbuky, and H. Gholam-Hosseini, "Distributed WSN data stream mining based on fuzzy clustering," in Ubiquitous, Autonomic and Trusted Computing, 2009. UIC-ATC'09. Symposia and Workshops on, 2009, pp. 395-400.

[14] J. R. Diaz, J. Lloret, J. M. Jimenez, and J. J. Rodrigues, "A QoS-based wireless multimedia sensor cluster protocol," International Journal of Distributed Sensor Networks, vol. 2014, 2014.

[15] M. Abazeed, N. Faisal, S. Zubair, and A. Ali, "Routing protocols for wireless multimedia sensor network: a survey," Journal of Sensors, vol. 2013, 2013.

[16] V. Kumar, S. Jain, and S. Tiwari, "Energy efficient clustering algorithms in wireless sensor networks: A survey," 2011.

[17] D. A. S. Juned M. Khan, "PERFORMANCE COMPARISON OF FCM AND K-MEAN CLUSTERING TECHNIQUE FOR WIRELESS SENSOR NETWORK IN TERMS OF COMMUNICATION OVERHEAD," Global Journal of Advanced Engineering Technologies and Science, pp. 26 - 29, May, 2014.

[18] A. Thangavelu and A. Pathak, "Clustering Techniques to Analyze Communication Overhead in Wireless Sensor Network," Editorial.