Implementation of the X-means Clustering Algorithm for Wireless Sensor Networks

Abdelrahman Radwan, Nazhatul Hafizah Kamarudin, Mahmud Iwan Solihin, Hungyang Leong, Mohamed Rizon
Engineering Department, UCSI University,
Jalan Puncak Menara Gading, Taman Connaught,
Kuala Lumpur, 56000, Malaysia

Abstract

K-means clustering algorithms in wireless sensor networks (WSNs) are potential solutions that prolong the network lifetime. However, there are limitations hampering these algorithms, such as size of clusters and the number of clusters. This paper proposes implementing X-means algorithm as a new clustering technique that overcomes K-means limitations; clusters constructed using tentative CHs and tentative area of centroids in an initial phase, After that, if a cluster meets splitting criteria, new centroids selected and new clusters constructed.

Keywords: K-means, X-means, Clustering, Wireless, Sensors, Networks.

1. Introduction

Machine to machine communication (M2M) and Internet of Things (IoT) are the next wireless communication technologies, these technologies require low power communication networks[1]; Wireless Sensor network (WSN) is a potential candidate that provides an infrastructure for these technologies. WSN is a group of wireless nodes that equipped with sensors and actuators; the network designed as an application specific and perform dedicated tasks. Wireless nodes operate on batteries as a source of energy and transmit unit drains nodes batteries[2]; therefore, routing data considered a costly and complex task. In literature clustering algorithms have shown promising network life enhancements; classical algorithms such as LEACH[3], HEED[4], and EECS[5] divide their networks into random clusters, and each cluster has a cluster head (CH), nodes collect data from surrounding then rely them to CHs; CHs have to forward the received data to a central unit called Sink using a long range communication called backhauling. The role of CH is frequently rotated to distribute the load of backhaul communication among nodes and re-clustering initiated every round of operation. Another clustering Algorithms are K-means algorithms such as Refs. 6-10; these algorithms estimate the number of clusters and their initial cluster centroids, then average Euclidean distance to centroid determined and nodes assigned to minimum distance centroid, the process repeated until cluster centroid fixed and cluster members are static.

K-means algorithms overcome classical algorithms problems such as unbalanced intra-cluster consumption caused by random centroid location and non-optimal number of CHs; but still these algorithm can form clusters that requires energy resource beyond the CH capabilities such as high number of cluster members that exceed CHs channel time slots; therefore, this article propose to form initial clusters using K-means clustering then X-mean algorithm used to evaluate formed clusters based on splitting criteria and to finalize list of clusters.

The rest of the article is as following, in section two the system model that includes Energy model and distance crossover model, then in section three we propose X-mean algorithm and describe its two operation phases with mathematical equations in details, after that in section four the simulation results and discussion on how the proposed algorithm prolong WSN lifetime is presented, finally in section we conclude our work with future recommendations to further enhance X-means algorithm.

2. System Model

The system model obtained from Ref.3, and it includes two parts; a consumption model for transmit, receive and data processing as in Eq. (1)(2)(3), and a distance model that estimates the crossover distance as in Eq. (4).

$$E_{tx}(b,d) = E_{elec}(b) + E_{tx-amp}(b,d) \quad (1)$$

$$E_{Tx} = \begin{cases} b * E_{elec} + b * \varepsilon_{fs} * d^2, \ d < d_o \\ b * E_{elec} + b * \varepsilon_{mp} * d^4, \ d \geq d_o \end{cases} \quad (2)$$

$$E_{Rx} = b * E_{elec} \quad (3)$$

$$d_o = \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}} \quad (4)$$

For part one, consumption model Eq. (1) represents the transmission consumption and it consists of E_{elec} that is the energy required to aggregate and process a single bit of data and $E_{tx\text{-}amp}$ that is the amplification energy required to transmit data over (d) distance. Eq. (1) expanded to Eq.(2) that describes two types of the amplification energies, where ε_{fs} describes the amplification energy required to transmit data using the freespace model, if the distance between source and destination is less than the crossover distance (d_0); otherwise, ε_{mp} describes the amplification energy that is required to transmit data using the multipath fading model. the crossover distance between the freespace model and the multipath model has a threshold (d_o), it determined by Eq. (4) and described in Ref. 7. The total transmission energy (E_{Tx}) is proportional to the distance between a source node and a destination node, where it corresponds to d^2 when d is less than d_0 and corresponds to d^4 when d is greater than or equal d_0. Receiving consumption (E_{Rx}) represented by Eq. (3) and includes the energy required to process a single bit.

3. Proposed Method

K-means algorithm depend on a predefined value of K to form clusters, K cluster centroids selected randomly, but even though K-means algorithm are deterministic because initial K centers have to be provided; a bad choice of initial centroids leads to poor clustering structures and performance. Therefore, we propose to implement X-means as a solution for K-means problems. The concept of the X-means clustering can be summarized into two main phases; at the first phase random k-means applied to select initial centroids, then nodes assigned to these initial centroids and new centroids position determined by Eq. (5), the process repeated recursively until a final copy of centroids is determined, the final copy of centroids called parent centroids. Then in phase two, every parent gives birth to two children $C_{n1,2}$, the initial location of the children determined as in Eq. (6); it is worth noticing that the children initial positions are the parent (x,y) position, but adjusted by the Euclidean distance (d_n); here the Euclidean distance is the distance between the parent centroid and the furthest node in the cluster as in Eq. (7).

$$P_n = \alpha_{n-1}(x,y) + \frac{1}{S_n}\sum_{i=1}^{S}\beta_i(x,y) \quad \begin{cases} n \in 1,2,3..k \\ i \in 1,2,3..S_n \end{cases} \quad (5)$$

$$C_{n1,2} = P_n \pm d_n \quad (6)$$

$$d_n = max\left(\left\{\sqrt{(x_n - x_i)^2 + (y_n - y_i)^2}\right\}\right), \ i \in S_n, n \in P_n \quad (7)$$

P_n: Parent (n) centroid.
$\alpha_{n-1}(x,y)$: x and y position of previous (n-1) centroid.
S_n: Set of nodes assigned to P_n.
$\beta_i(x,y)$: the x and y position of node (i).
d_n: the maximum distance between parent centroid and the furthest member node (i).

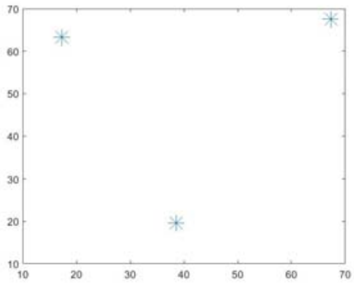
The new positions are the new children $C_{n1,2}$ and nodes assigned to these children according to received signal strength or distance. Then X-means adjust children centroid positions. The mechanism of X-means fairly designed to minimize backhaul distance by adjusting cluster centroids according to their distance to sink as in Eq.(8). Nodes grouped in clusters based on their average distance from children centroids and their distance to sink as in Eq. (9).

$$CC_{n1,2} = \emptyset(x,y) + \frac{1}{S_n}\sum_{i=1}^{S}\beta_i(x,y) \quad \begin{cases} n \in 1,2,3..k \\ i \in 1,2,3..S_n \end{cases} \quad (8)$$

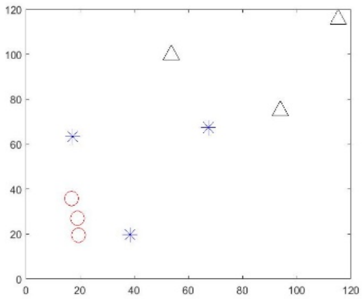
$$AvgD = min((d_c + d_{sink})/2) \quad (9)$$

$CC_{n1,2}$: position of parent (n) Children Centroid.
$\emptyset(x,y)$: position of Sink.
AvgD: average distance to centroid and sink.
d_c: A node Euclidean distance to child centroid.
d_{sink}: A node Euclidean distance to sink.

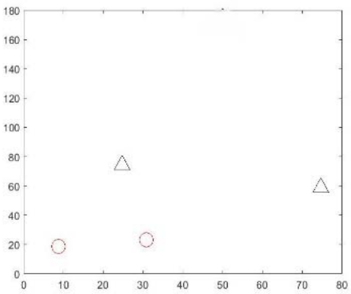
X-means algorithm has to execute recursively until there are no changes at centroids position, the algorithm has three outcomes, first children form their own cluster and parents collapsed, second outcome some of the children and parent diminished, and third outcome children collapsed and parent repositioned to best locations. Fig. 1. shows the new generated children centroids, the first generation is parents centroid generated by k-means as in Fig. 1-A, and the number of generated parents are three, then the implementation of X-means generated children as in Fig. 1-B (marked with triangle and circles), the total of children and parents becomes nine centroids, finally the recursive run of X-means diminishes some of the parents, children, and repositions the remaining children as in Fig. 1-C, the final centroids become static clusters of the network.
Nodes that have minimum distances to centroids selected to become CHs, then if a CH energy dropped below the energy threshold, it steps down and becomes a cluster member and the next closest node to the cluster centroid



A) Parent centroids generated after running K-means



B) Prents give birth to new children centroids



C) Remaining centroids after running x-means

Fig. 1. X-means two phase of centroid generation

become a CH. Once a node energy dropped below the threshold the node determines a new threshold, the new threshold is a step down from the original threshold by (σ) as in Eq. (10).

$$E_{th} = E_{th} - (\sigma * E_{th}) \quad (10)$$

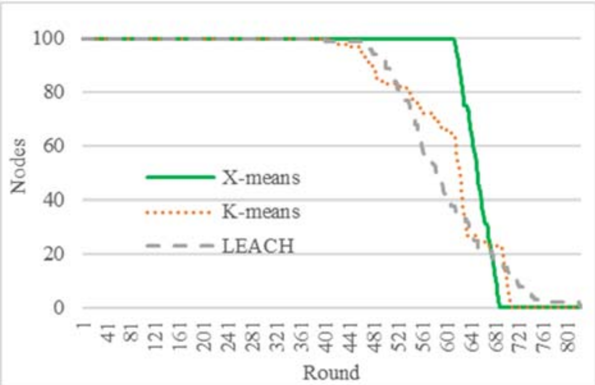
The propose algorithms simulated in MATLAB and the simulation parameters listed in table (1). There are two simulation scenarios considered; first a random deployment of 100 nodes in an area 100 x100, second a random deployment of 200 nodes in an area of 200 x 200. The result is benchmarked against prominent algorithm such as LEACH [3], and a centralized K-means algorithm[7]. X-means implemented to prolong the network lifetime and the deployed sensor nodes always have data to transmit; therefore, the network lifetime measured in round and a single round represents a single packet received from a CH at the sink. At simulated scenarios the sink positioned outside the monitoring field at (50,175) and (100,275) for first and second scenarios respectively. X-means implemented on the sink and its aware of nodes position, thus; clusters formed by sink, but rotating of CH role is a node decision.

Table 1: X-means simulation parameters

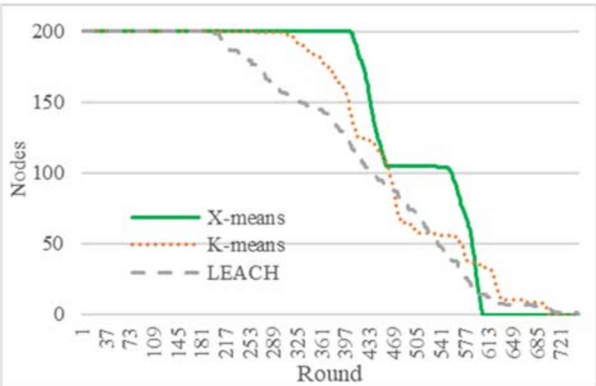
Parameter	Description	Value
d_0	Crossover distance	87m
ε_{fs}	Freespace model amplification energy	10pJ/bit/m²
ε_{mp}	Multipath amplification energy	0.0013pJ/bit/m4
E_{elec}	Single bit processing energy	Tx or Rx = 50nJ/bit Aggregation = 5nJ/bit
E_{init}	Initial Energy	0.5J
K	Initial centroids	3
E_{th}	Energy threshold	0.2J
σ	E_{th} step down	5%
Packet Size	Data size	4.2Kb/packet
Control Packet	Packet Headers	0.2Kb/packet

4. Results and Discussion

The main concept of X-means is to select the best centroid positions, these positions must minimize nodes consumption per round by reducing intra-cluster communication consumption and reducing CHs backhaul consumption. In X-means number of centroids are not determined, but only initial centroids are set. The number of initial centroids has to be set to half of the maximum allowed centroids, for example, in scenario one, the maximum allowed CHs is 5% of the total networks nodes[3], so the number of CHs is 3 after rounding up. After executing X-means the number of parents generated were 3, total children were 6, and final number of centroids were 4, the results of this enhancements shown in Fig. 2-A, the enhancement of centroids number and positions reflect as enhancement of First Node Death (FND), where using X-means algorithm extend FND by 53% over LEACH and 51% K-means. Similarly for the second scenario, the maximum number of CHs is 5%; therefore the maximum number of CHs is 10, and initial centroids/parents required by X-means is 5 centroids and generated children are 10, but the final number of centroids remained 5 and only position of centroids adjusted. The result of simulating the second scenario also shows that x-means prolonged the FND and enhanced the performance over LEACH and K-means. It is worth noticing in X-means network that in both scenarios, after FND occurred a sharp drop in the number of alive nodes occurred too; which indicates that the X-means mechanism in selecting centroid positions and grouping nodes based on their distance from the sink has nearly balanced nodes consumption. In the second scenario, the sharp drop of number of alive nodes stopped at 105 nodes, this because the nodes at the far end of the network depleted more energy during backhaul communication with the sink than those at closer distances to the sink.



A) 100 node networks



B) 200 node networks.

Fig. 2. X-means performance simulation and benchmarking in multiple scenarios.

In our simulation, we also monitored the total network consumption in three periods of network lifetime; the first period represents the duration from the network startup time until 1% of nodes death, then the second period is the energy spent from 1% of nodes death until 50% of nodes death, after that the third period from 50% of nodes death until 100%. X-means algorithm has shown less consumption than LEACH and K-means at the first and the second periods because of centroid optimization, but X-means and K-means consumed higher energy at the end of network lifetime. this is because of static clustering where most of the nodes at close distance from centroids have died and the remaining nodes has high intra-cluster distance, which unlike LEACH that depends on random clustering.

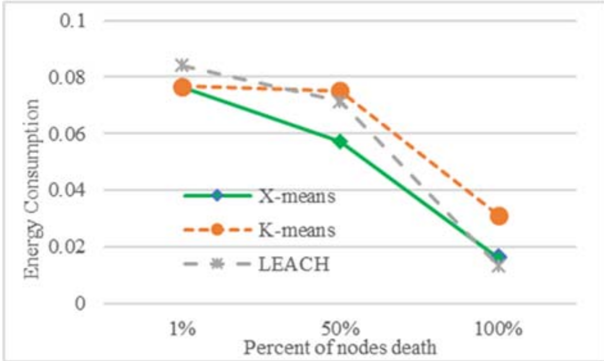


Fig. 3. The average network energy consumption during network lifetime.

5. Conclusion

In this article we proposed X-means as a solution for K-means shortcomings. The algorithm depends on determined value of parents to generate two children for each parent, then children evaluated to generate a final copy of centroids that is less than the maximum number of network centroids and these centroids have to enhance network performance. The simulation scenarios were limited to 100 node deployments and 200 node deployments, thus; the performance is not checked for large scale deployment such as 1000 nodes.

In future work, the design of X-means need to be adjusted to include more than 2 children per parent, in order to increase the chance of finding minimum cost centroids.

Acknowledgements

The author would like to thank UCSI university Trust Fund and Centre of Excellence for Research, Value Innovation and Entrepreneurship (CERVIE) for support this work.

References

1. Li X, Li D, Wan J, Vasilakos AV, Lai CF, Wang S. *A review of industrial wireless networks in the context of industry 4.0. Wireless networks.* 2017 Jan 1;23(1):23-41.
2. Yick J, Mukherjee B, Ghosal D. *Wireless sensor network survey.* Computer networks. 2008 Aug 22;52(12):2292-330.
3. Heinzelman WR, Chandrakasan A, Balakrishnan H. *Energy-efficient communication protocol for wireless*

microsensor networks. In Proceedings of the 33rd annual Hawaii international conference on system sciences 2000 Jan 7 (pp. 10-pp). IEEE.

4. Younis O, Fahmy S. *HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks.* IEEE Transactions on mobile computing. 2004 Oct 1(4):366-79.
5. Ye M, Li C, Chen G, Wu J. *EECS: an energy efficient clustering scheme in wireless sensor networks.* InPCCC 2005. 24th IEEE International Performance, Computing, and Communications Conference, 2005. 2005 Apr 7 (pp. 535-540). IEEE.
6. Periyasamy S, Khara S, Thangavelu S. *Balanced cluster head selection based on modified k-means in a distributed wireless sensor network*. International Journal of Distributed Sensor Networks. 2016 Mar 1;2016:2.
7. Park GY, Kim H, Jeong HW, Youn HY. *A novel cluster head selection method based on K-means algorithm for energy efficient wireless sensor network.* In2013 27th International conference on advanced information networking and applications workshops 2013 Mar 25 (pp. 910-915). IEEE.
8. Sasikumar P, Khara S. *K-means clustering in wireless sensor networks.* In2012 Fourth international conference on computational intelligence and communication networks 2012 Nov 3 (pp. 140-144). IEEE.
9. Razzaq, M., Ningombam, D.D. and Shin, S., 2018, January. *Energy efficient K-means clustering-based routing protocol for WSN using optimal packet size.* In 2018 International Conference on Information Networking (ICOIN) (pp. 632-635). IEEE.
10. Khan A, Tamim I, Ahmed E, Awal MA. *Multiple parameter based clustering (mpc): Prospective analysis for effective clustering in wireless sensor network (wsn) using k-means algorithm.* Wireless Sensor Network. 2012 Jan 1;4(1):18-24.