Energy-efficient Clustering with Fast Data Compression in Sensor Networks

Xiaoling Wu, Mohammed A. U. Khan, Jinsung Cho, Sungyoung Lee[*] and Young-Koo Lee
Department of Computer Engineering, Kyung Hee University, Korea
{xiaoling,khan,sylee}@oslab.khu.ac.kr, {chojs,yklee}@khu.ac.kr

Abstract

Wireless sensor networks (WSNs) are used for various applications, such as habitat monitoring, automation, agriculture, and security. In order to send information from very high number of sensor nodes to the base station, it is necessary and economical to group sensors into clusters to prolong network lifetime. Due to the resource limitation of sensor nodes, the collected information from sensor nodes in the cluster has to be compressed quickly and precisely for transmission. In this paper, we propose a VQ-LBG based approach for cluster formation in WSN. The most distinguishing feature of the proposed method is that both energy efficient cluster formation and fast data compression can be guaranteed. Experiment shows its great improvement over other related methods.

1. Introduction

Wireless sensor networks (WSNs) have received a strong interest in the recent past due to their various applications, such as habitat monitoring, automation, agriculture, and security. The network topology control for large number of randomly placed sensors was studied in the recent years emphasizing the limited battery power. Generally, there are three methods that can be considered as possible networking protocols: direct communication, multi-hop routing, and clustering. In order to send information from a very high number of sensor nodes to the base station, it is necessary and economical to group sensors into clusters to prevent redundant information transmission and prolong network lifetime. Given the parameters for variation of energy consumption in the nodes, there are some main problems: How many sensors should be connected to each cluster head (CH), how many clusters are needed, how to select CH, and where each CH should be positioned. Another typical challenge is that due to the resource limitation of sensor nodes (CPU, memory, bandwidth, and energy), the collected information from sensor nodes in the cluster has to be compressed quickly and precisely for transmission.

The research community is actively looking into these challenges. [1] proposes the LEACH protocol, which is a hierarchical self-organized cluster-based approach for monitoring applications. The data collection area is randomly divided into several clusters. Based on time division multiple access (TDMA), the sensor nodes transmit data to the cluster heads, which aggregate and transmit the data to the base station. A new set of cluster heads are chosen after specific time intervals. A node can be re-elected only when all the remaining candidates have been elected. The work in [2] shows that a 2-tier architecture is more energy efficient when hierarchical clusters are deployed at specific locations. [3] describes a multi-level hierarchical clustering algorithm, where the parameters for minimum energy consumption are obtained using stochastic geometry. HEED [4] selects cluster heads through *O(1)* time iteration according to some metric and adopts the multi-hop communication to further reduce the energy consumption. PEGASIS [5] improves the performance of LEACH and prolongs the network lifetime greatly with a chain topology. But the delay is significant although the energy is saved. In [6], particle swarm optimization (PSO) is used to find the optimal cluster head positions for sensor network deployment. There are some other related works [7~9] which efficiently use energy through clustering. However, none of the above clustering approaches can guarantee both energy efficient sensors clustering and fast data compression. For example, in their proposed cluster based topology, sensing values inside one cluster may not provide the closest correlations, thus the optimal data compression can not be guaranteed. To the best of our knowledge, this is the first paper that suggests a sensor cluster formation method with fast data compression while minimizing energy consumption in sensor networks.

[*] Corresponding author

The proposed approach is based on VQ-LBG design algorithm. It is a lossy data compression method based on the principle of block coding. It provides fast data compression process with minimum average distortion at CHs and uniform cluster formation. This would help in balancing the system load on each CH since all the clusters are balanced, and at the same time, the communication energy consumption will be significantly reduced due to the efficient data compression.

The rest of the paper is organized as follows. Section 2 describes the WSN model we used. The proposed cluster formation algorithm based on VQ-LBG is introduced in section 3. Section 4 presents the experiment results and section 5 concludes the paper.

2. WSN model

In this section we describe our model of a wireless sensor network with nodes homogeneous in their initial amount of energy. We particularly present the energy model and how the optimal number of clusters can be computed. We assume that all nodes are distributed randomly over the sensor field.

Previous work have studied either by simulation [1] or analytically [10], [11] the optimal probability of a node being elected as a cluster head as a function of spatial density when nodes are uniformly distributed over the sensor field. This clustering is optimal in the sense that energy consumption is well distributed over all sensors and the total energy consumption is minimum. Such optimal clustering highly depends on the energy model we use. For the purpose of this study we use similar energy model and analysis as proposed in [1].

According to the radio energy dissipation model, in order to achieve an acceptable Signal-to-Noise Ratio (SNR) in transmitting an l bit message over a distance d, the energy expended by the radio is given by:

$$E_T(l,d) = \begin{cases} lE_{elec} + l\varepsilon_{fs}d^2 & if\ d \le d_0 \\ lE_{elec} + l\varepsilon_{mp}d^4 & if\ d > d_0 \end{cases} \quad (1)$$

where E_{elec} is the energy dissipated per bit to run the transmitter or the receiver circuit, ε_{fs} and ε_{mp} are amplifier constants, and d is the distance between the sender and the receiver. By equating the two expressions at d=d_0, we have $d_0 = \sqrt{\varepsilon_{fs} / \varepsilon_{mp}}$.

To receive l bit message, the radio expends:

$$E_R(l) = lE_{elec} \quad (2)$$

Assume an area $A = L \times L$ square meters over which n nodes are uniformly distributed. For simplicity, assume the sink is located in the center of the field, and that the distance of any node to the sink or its cluster head is $\le d_0$ .Thus, the energy dissipated in the cluster head node during a round is:

$$E_{CH}(l) = (\frac{n}{n_c} - 1)lE_{elec} + \frac{n}{n_c}lE_{DA} + lE_{elec} + l\varepsilon_{fs}d_{toBS}^2 \quad (3)$$

where n_c is the number of clusters, E_{DA} is the processing (data aggregation) cost of a bit per report to the sink, and d_{toBS} is the average distance between the cluster head and the sink. The energy used in a non-cluster head node is equal to:

$$E_{nonCH}(l) = lE_{elec} + l\varepsilon_{fs}d_{toCH}^2 \quad (4)$$

where d_{toCH} is the average distance between a cluster member and its cluster head. The expected squared distance from the nodes to the CH is given by:

$$E[d_{toCH}^2] = \frac{L^2}{2\pi n_c} \quad (5)$$

The energy dissipated in a cluster per round is given by:

$$E_{cluster} \approx E_{CH} + \frac{n}{n_c}E_{nonCH} \quad (6)$$

The total energy dissipated in the network is equal to:

$$E_{total} = l(2nE_{elec} + nE_{DA} + \varepsilon_{fs}(n_c d_{toBS}^2 + nd_{toCH}^2)) \quad (7)$$

By differentiating E_{total} with respect to n_c and equating to zero, the optimal number of constructed clusters can be found [12]:

$$n_{c-opt} \approx \sqrt{\frac{n}{2\pi}}\frac{L}{d_{toBS}} = \sqrt{\frac{n}{2\pi}}\frac{2}{0.765} \quad (8)$$

3. Cluster formation based on VQ-LBG

3.1. VQ design problem

We propose a data compression guaranteed clustering algorithm based on vector quantization (VQ).

Our approach is designed according to the following two observations: (1) In sensor networks, the historical information exhibits similar patterns over time, (2) different measurements are intrinsically correlated, and (3) coordinates of sensor nodes can be regarded as a feature pattern. VQ is a lossy data compression method based on the principle of block coding. It is a fixed-to-fixed length algorithm. LBG is a VQ design algorithm based on a training sequence. The use of a training sequence bypasses the need for multi-dimensional integration. A VQ that is designed using this algorithm are referred to as an LBG-VQ [13].

The VQ design problem can be stated as follows. Given a vector source with its statistical properties known, given a distortion measure, and given the number of codevectors, find the set of codebook and a cluster formation (partition) which results in the smallest average distortion.

We assume that there is a training sequence consisting of M source vectors:

$$\tau = \{X_1, X_2, ..., X_M\}.$$

This training sequence can be obtained from some large database. M is assumed to be sufficiently large so that all the statistical properties of the event source are captured by the training sequence. We assume that the source vectors are k-dimensional, e.g.,

$$X_m = \{x_{m,1}, x_{m,2}, ..., x_{M,k}\},\ m= 1, 2,..., M$$

Let N be the number of codevectors and let $C = \{c_1, c_2, ..., c_N\}$, represents the codebook. Each codevector is k-dimensional, e.g., $c_i = \{c_{i,1}, c_{i,2}, ..., c_{i,k}\}$, i=1, 2,…, N. Let S_i be the encoding region associated with codevector c_i and let $P = \{S_1, S_2, ..., S_N\}$ denote the partition of the space. If the source vector X_m is in the encoding region S_i, then its approximation $Q(X_m) = c_i$, if $X_m \in S_i$. Assuming a squared-error distortion measure, the average distortion is given by:

$$D_{ave} = \frac{1}{Mk}\sum_{m=1}^{M}\|X_m - Q(X_m)\|^2$$

where $\|e\|^2 = e_1^2 + e_2^2 + ... + e_k^2$. The design problem can be briefly stated as follows: Given τ and N, find C and P such that D_{ave} is minimized.

3.2. Optimality criteria

If C and P are a solution to the above minimization problem, it must satisfy the following two criteria.

- Nearest Neighbor Condition:

$$S_i = \{X : \|X - c_i\|^2 \le \|X - c_{n'}\|^2\} \\ \forall n' = 1,2,...,N\}$$

  This condition means that the encoding region S_i should include all vectors that are closer to c_i than any of the other codevectors. For those vectors lying on the boundary, any tie-breaking procedure will do.

- Centroid Condition: $c_i = \dfrac{\sum_{X_m \in S_n} X_m}{\sum_{X_m \in S_n} 1}$,

  i=1,2,…,N

  It says that the codevector c_i should be average of all those training vectors that are in encoding region S_i. We should ensure that at least one training vector belongs to each encoding region to guarantee that the denominator in the above equation is never 0.

3.3. LBG design algorithm

The LBG-VQ design algorithm is an iterative algorithm which alternatively solves the above two optimality criteria [13]. The algorithm requires an initial codebook $C^{(0)}$. This initial codebook is obtained by the splitting method. In this method, an initial codevector is set as the average of the entire training sequence. This codevector is then split into two. The algorithm runs with these two vectors as the initial codebook. The two codevectors are splitted into four and the process is repeated until the desired number of codevectors is obtained. The algorithm is summarized below.

1) Given τ. Fixed $\varepsilon > 0$ to be a very small value.

2) Let $N = 1$ and $c_1^* = \frac{1}{M}\sum_{m=1}^{M} X_m$, and calculate

$$D_{ave}^* = \frac{1}{Mk}\sum_{m=1}^{M}\|X_m - c_1^*\|^2.$$

3) Splitting: For *i=1,2,...,N*, set $c_i^{(0)} = (1+\varepsilon)c_i^*$, $c_{N+i}^{(0)} = (1-\varepsilon)c_i^*$ and set N=2N.

4) Iteration: Let $D_{ave}^{(0)} = D_{ave}^*$. Set the iteration index $i = 0$.

a) For m= 1, 2,…, M, find the minimum value of $\left\| X_m - c_n^{(i)} \right\|^2$ over all n=1, 2,…, N. Let n^* be the index which achieves the minimum. Set

$$Q(X_m) = c_{n^*}^{(i)}$$

b) For n=1, 2,…,N, update the codevector

$$c_n^{(i+1)} = \frac{\sum_{Q(X_m)=c_n^{(i)}} X_m}{\sum_{Q(X_m)=c_n^{(i)}} 1}$$

c) Set i=i+1.

d) Calculate $D_{ave}^{(i)} = \frac{1}{Mk} \sum_{m=1}^{M} \left\| X_m - Q(X_m) \right\|^2$

e) If $(D_{ave}^{(i-1)} - D_{ave}^{(i)}) / D_{ave}^{(i-1)} > \varepsilon$, go back to Step *a)*.

f) Set $D_{ave}^{*} = D_{ave}^{(i)}$. For n=1, 2,…,N, set $c_n^{*} = c_n^{(i)}$ as the final codevectors.

5) Repeat Steps 3) and 4) until the desired number of codevectors is obtained.

3.4. Proposed Algorithm

Considering both finding the initial cluster heads position for proper cluster formation and fast data compression, we propose our algorithm applied in sensor networks for both cluster formation and data compression as follows:

Step1: Create the codebook from training data set of the sensors.

Step2: Transmit the codebook to the base station.

Step3: Base station determines the CHs which has shortest distance to codebook.

Step4: Let the sensor collect data and fill the local buffer.

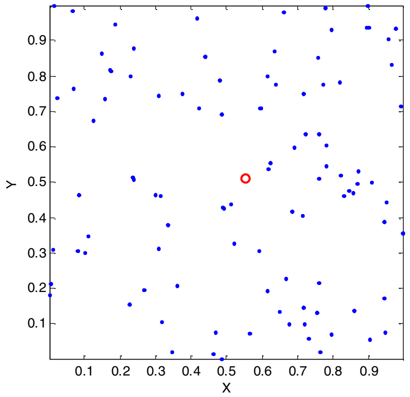
Step5: Compute the codebook update locally at CH and send to BS.

Step6: Sensor node with highest energy inside a fixed cluster becomes new CH.

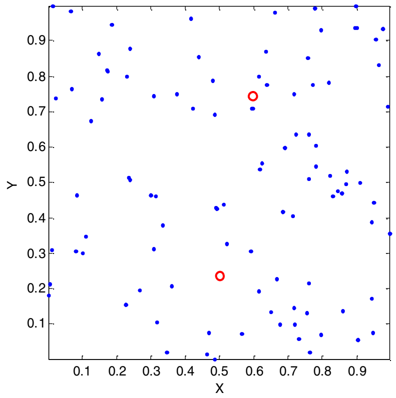
Step7: Go to step 4 and repeat until the last node dies.

4. Performance Evaluations

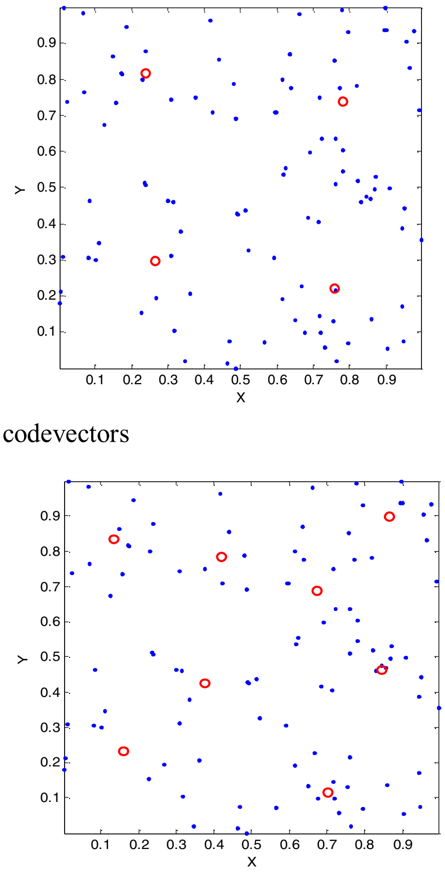
We simulate a wireless sensor network in a 100m×100m field. We set E_{elec} as E_{elec}=50nJ/bit, E_{DA} as E_{DA}=50$nJ/bi/report$ and the amplifier constant is taken as ε_{fs}=10$pJ/bit/m^2$, ε_{mp}= 0.0013$pJ/bit/m^2$, thus d_0 can be computed as about 87m. We set total sensor nodes number n=100, and the optimal number of constructed clusters can be approximately 8 according to Eq (8). So the number of codevectors N is 8. We take $\tau = \{X_1, X_2\}$ with X_1 as the location of sensor nodes, and X_2 as temperature and humidity sensing values. We regard the source vectors as 2-dimensional. So the sample data set is a 100×4 matrix. Without losing generality, we regard the sensing values to be the coordinates divided by 100. Thus when we simulate, we standardize the coordinates range as [0, 1] instead of [0, 100m].



(a) 1 codevector



(b) 2 codevectors



(c) 4 codevectors

(d) 8 codevectors

Figure 1. Snapshots of the VQ-LBG algorithm based codebook creation

Figure 1 (a)-(d) show four snapshots of the VQ-LBG algorithm based codebook creation. The red circles represent codebook. Figure 2 is the final cluster formation in which the nodes with same color are bunched into one cluster. The sensor node in one cluster which has the shortest Euclidean distance to its codevector is selected as CH. Figure 3 is an example of randomly generated 8 cluster heads marked as red stars which form uneven cluster distribution used in LEACH. We can see that compared with VQ-LBG based cluster formation result, randomly generated cluster heads cannot always guarantee uniform cluster formation.

Table 1 shows the loop count and distortion verses number of centers during the algorithm execution process. It is obvious that the distortion is decreased significantly after the final codebook with 8 centers is determined.

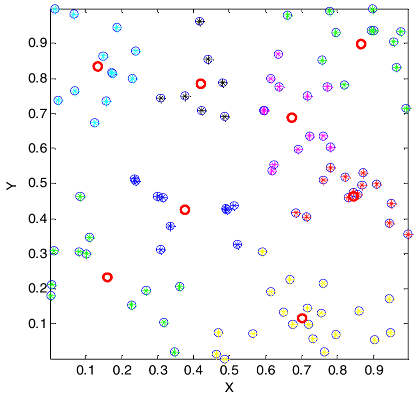


Figure 2. Final cluster formation in which nodes with same color are bunched into one cluster

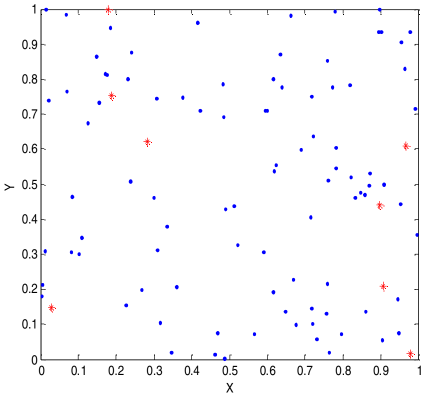


Figure 3. An example of randomly generated 8 CHs marked as red stars which form uneven cluster distribution

Table 1. Loop count and distortion vs No. of centers during the execution

No. of centers	Loop count	distortion
2	9	10.834
4	6	4.20282
8	8	1.83539

Figure 4 exhibits the distribution of the number of clusters in randomly selected 50 rounds in both proposed algorithm and LEACH. The number of clusters varies widely in each run in LEACH; on the other hand, the cluster number varies narrowly at the optimal range in proposed algorithm. Although the clusters are fixed in our algorithm and only the cluster head nodes are rotated, once the clusters are formed, there is no set-up overhead at the beginning of each round. Depending on the cost of forming adaptive clusters, our approach where the clusters are formed once and fixed and the cluster head position rotates among the nodes in the cluster is more energy efficient than LEACH. Besides that, our method at the same time guarantees fast data compression which is also an important issue in WSNs due to the scare resources of sensor node.

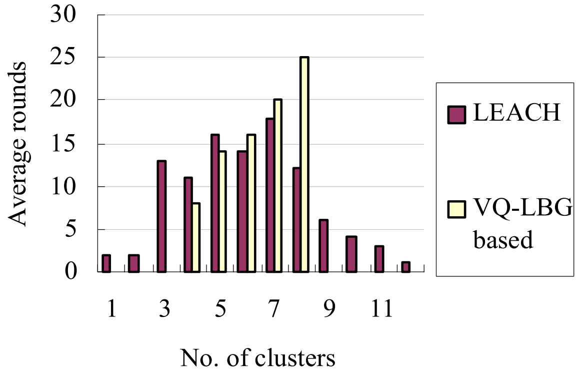


Figure 4. The number of clusters in each round in both proposed algorithm and LEACH

5. Conclusion and future work

In this paper, we proposed a VQ-LBG algorithm based clustering approach for sensor networks which provides fast data compression process with minimum average distortion at CHs. The resulted uniform cluster distribution also balanced the system load on each CH since all the clusters were balanced, and at the same time, the communication energy consumption was significantly reduced due to the efficient data compression.

We used fixed clusters and rotate cluster head nodes within the cluster. However, it may increase non-cluster head node energy dissipation and increasing inter-cluster interference. In the future work, we will consider dynamic cluster head election mechanism based on the above research work. And heterogeneous sensor network environment and Gaussian sensor distribution will also be further studied. Since in large scale sensor networks multi-hop communication is a mainstream technique for energy saving, we will also try to remove the assumption of single-hop and design an energy efficient protocol for both intra-cluster and inter-cluster data transmission in the future work.

6. Acknowledgement

This research was partially supported by the Driving Force Project for the Next Generation of Gyeonggi Provincial Government in Republic of Korea.

7. References

[1] Wendi B. Heinzelman, Anantha P. Chandrakasan, and Hari Balakrishnan, "An Application-Specific Protocol Architecture for Wireless Microsensor Networks", IEEE Transactions on Wireless Communications, Vol. 1, No. 4, 2002, pp. 660 – 670.

[2] J. Pan, L. Cai, Y. T. Hou, Y. Shi, and S. X. Shen, "Optimal base-station locations in two-tiered wireless sensor networks", IEEE Transactions on Mobile Computing (TMC), 4(5), 2005, pp. 458–473.

[3] S. Bandyopadhyay and E. J. Coyle, "An energy efficient hierarchicalclustering algorithm for wireless sensor networks", In Proceedings of the IEEE Conference on Computer Communications. INFOCOM, 2003.

[4] O. Younis, et. al., "HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks", IEEE Transactions on Mobile Computing, 3(4), 2004, pp. 660-669.

[5] S. Lindsey, et. al., "PEGASIS: Power-Efficient Gathering in Sensor Information Systems", IEEE Aerospace Conference Proceedings, Vol. 3, 9-16, 2002, pp. 1125-1130.

[6] Xiaoling Wu, Shu Lei, Yang Jie, Xu Hui, Jinsung Cho and Sungyoung Lee, "Swarm Based Sensor Deployment Optimization in Ad hoc Sensor Networks", ICESS' 05/ LNCS (SCIE), Springer, 2005, pp. 533-541.

[7] S. Bandyopadhyay, et. Al, "An Energy-Efficient Hierarchical Clustering Algorithm for Wireless Sensor Networks", IEEE INFOCOM, 2003.

[8] H. Chan, et. al., "ACE: An Emergent Algorithm for Highly Uniform Cluster Formation", the First European Workshop on Sensor Networks (EWSN), 2004.

[9] J. Kamimura, et. al., "Energy-Efficient Clustering Method for Data Gathering in Sensor Networks", 1st Annual International Conference on Broadband Networks, 2004.

[10] S. Bandyopadhyay and E. J. Coyle, "An energy efficient hierarchical clustering algorithm for wireless sensor networks", Proceedings of INFOCOM, 2003.

[11] S.Bandyopadhyay, E.J.Coyle, "Minimizing communication costs in hierarchically-clustered networks of wireless sensors", Computer Networks, vol. 44, No. 1, 2004, pp. 1–16.

[12] Georgios Smaragdkis, Ibrahim Matta and Azer Bestavros, "SEP: A Stable Election Protocol for clustered heterogeneous wireless sensor networks", Proc. of the Int'l Workshop on SANPA, 2004.

[13] http://www.data-compression.com/vq.shtml