Measuring the quality of information in clustering protocols for sensor networks

Pubali Banerjee
Department of Computer Science
Texas A and M University
College Station, TX, USA
email: pubali@cs.tamu.edu

ABSTRACT

The performance of a clustering protocol for sensor network is often measured by the energy spent by the network. But there are other metrics that are of significance, namely the latency of the network and the quality of the overall information sensed by the network. In this paper, we have show three ways to measure the quality of information obtained in a sensor network. Our first method an information loss metric that we have formulated is a statistical method to estimate the quality of information. Our second metric, the entropy ratio metric based on Shannon's information theory gives us a ratio that indicates the quality of information obtained. Our third method is based on variance. This deterministic method essentially calculates the mean square error between the true and the estimated signal over all the nodes. This simple method provides an effective way to measure the amount of information lost.

1. INTRODUCTION

Sensor networks are designed with energy optimization in mind. A large number of nodes are deployed in remote location to obtain aggregated information. The information gathered at every node is aggregated and sent to the base station. The conventional way to measure the performance of clustering protocols is to measure the amount of energy used in the whole process [1][2][3]. Some researchers have used the average number of hops required for data to travel to the base station as a measure of the performance of the protocol. This is in fact a measure of the delay or latency in the protocol. A method to formulate energy spent and delay in clustering protocols for sensor networks is described in [4]. In this paper, we have formulated three metrics to measure the quality of information obtained at the base station. We somehow want to measure the information loss between the initial information obtained at the individual nodes and the aggregated information obtained at the base station.

In the last five years many researchers have tried to come up with clustering protocols for sensor networks. One of the most frequently used distributed clustering protocols for sensor networks is LEACH [1] which we will discuss in a later section. Another proposed cluster protocol is PEGASIS [2] which is a chain based clustering protocol. Yet another approach is the one used in the TEEN protocol[5] where the nodes sense data continuously but they only send data to the cluster head when the sensed value is greater than a certain preset threshold value. A modification of TEEN is suggested in [6]. In this protocol called APTEEN, a hybrid approach that combines proactive and reactive networks are used. Another more recent modification of LEACH is suggested in [3]. According to this variation of LEACH, a node has a higher chance of becoming a cluster head if it is closer to the base station.

In all these protocols the focus is on minimizing the energy spent. Although energy spent is the key concern in sensor nets there are other metrics of performance measure that are of importance as shown in [4].

To the best of our knowledge, past work on the quality of the information collected by clustering protocols for sensor networks has been limited. In [7] the authors try to formalize the data aggregation efficieny in the context of two protocols, a chain based and a clustering based protocol. The protocols are then varied to achieve better data aggregation efficiency. The authors conclude that the larger the size of the data to be aggregated and the smaller the entropy of the data, the greater is the efficiency of the data aggregation process. However, they do not measure the quality of the aggregate data. We on the other hand, want to measure or formalize a method to measure the quality of the data received at the base station and how much it varies from the sensed data. In another paper [8], the authors formulate the amount of data likely to be lost in the whole sensing process in order to evaluate the quality of the aggregated data. They list four possible sources of data loss namely sensor wear and tear, hardware limitations, radio attenuation and network congestions. They then formulate a model that estimates the data loss.

In this paper, we focus on the quality of information received at the base station and the amount of inaccuracy in that data due to the aggregation process and the different ways of measuring that. We use the LEACH protocol as a test bed for our metrics.

2. LEACH

2.1 The LEACH model

LEACH[1] uses a simplified network model. The sensor nodes are uniformly spread over a rectangular area. The base station is assumed to be very far away from this square area. The nodes organize themselves into local clusters. The data is locally sent to the cluster heads. The cluster heads do the data aggregation and then transmit the aggregated data to the base station. A cluster head

WICON 2007, October 22-24, Austin, USA

DOI 10.4108/pwsn.2007.2292

does more work than a non cluster head node. When a cluster head dies, a chunk of the nodes lose communication, i.e. they effectively die. To reduce this, within a cluster, cluster heads are rotated, and all nodes get a chance to be a cluster head equal number of times. This ensures that the energy of all the nodes are balanced. The algorithm works in rounds, one round is defined as setting up the clusters, getting data from all the nodes once, fusing all that data in the cluster heads and sending that data to the base station.

2.2 The LEACH Parameters

In LEACH, there are some parameters whose values determine the performance of the algorithm. One such parameter is the number of clusters in a round. The total energy consumed in a round depends on the number of clusters. If the number of clusters is more, the coverage area of each cluster is small, thereby the energy requires for communication between the cluster head and the members is less. On the other hand, a fewer clusters would mean less overhead to set the clusters up and fewer nodes will have to serve as cluster heads which is an energy expensive event. Thus there exists an optimal number of clusters for which the energy consumption will be minimal. In LEACH, an expression for the optimal number of clusters, k, is derived.

The analysis of LEACH in[1] does not include the MAC layer time scheduling policy nor does it include the number of routing hops in the networks layer. But the optimal number of clusters will depend on these factors as the energy consumed depends on these. Therefore k will change with the MAC protocol scheduling times and the number of routing hops. Again, in LEACH, the cluster heads set up a schedule for the non cluster heads to transmit. The scheme used is TDMA. The value of k will also affect the delay or latency of the algorithm. It will also affect the quality of the data aggregated. Thus there exists an optimal value of k for which the energy consumed per round, the delay incurred and the quality of data aggregated will all be optimal.

In this paper, we develop three different ways to measure the quality of information gathered by a sensor network. We also include the formal derivation of the amount of energy spent and delay [4] for LEACH. In the simulation part, we show how the different ways of measuring the amount of quality of information can be used to obtain the optimal clustering configuration of the protocol.

3. MEASURING THE LOSS OF INFORMATION

In this paper, we provide three ways to measure the loss of information in a sensor network. Our first method is a statistical method that estimates the difference in information content of the true signal and the sensed signal obtained after aggregation at the base station. In the second method we have developed a metric based on Shannon's information theory which is a ratio of the entropy of the sensed signal to that of the signal obtained at the base station. The third method is based on mean square error and variance and estimates the deterministic error between the true signal and the signal obtained at the base station.

3.1 The loss of information metric

3.1.1 Description

The cluster heads fuse data from the member nodes and then transmit the fused data to the base station.If there are fewer cluster heads, more energy is saved. But energy is saved at the expense of information. The base station only knows the aggregate information fused at the cluster heads. Thus, if there are more cluster heads then the base station has more information about the local regions and vice versa. Therefore, the information loss is an important parameter that is a measure of the efficiency of the algorithm [7]. The amount of information loss is dependent on the number of clusters in the network. Next, we formally determine an expression for the total loss of information.

3.1.2 Formulation

Let $f(x, y)$ be the true signal over the coverage area R. Individual sensors,located at random points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, sense $f(x, y)$ with noise, and send

$$S_j = f(x_j, y_j) + \epsilon_j,\ j = 1, \ldots, N \tag{1}$$

to the cluster head where ϵ_j is a white noise process. A cluster head then processes the noisy signal from its nodes and sends the aggregated information Y_i to the base station. Assuming there are k cluster heads operating at a given time, the information received by the base-station is $Y_1, \ldots, Y_k$. We define the *loss of information due to data aggregation and noise* as

$$\Delta_k = \sum_{i=1}^{k} E \int_{A_i} (Y_i - f(x, y))^2 \, dxdy \tag{2}$$

where A_i is the region covered by cluster head i and E stands for the (conditional) expectation (given the location of the sensors (x_i, y_i) and the clusters $A_i, \ldots, A_n$). In addition to minimizing the energy and the total delay, a sensor network protocol must also guard against excessive loss of information, Δ_k.

Suppose that the data aggregation at the cluster head is done using the average of the noisy signals received from the neighborhood sensors. Then,

$$Y_i = \text{avg}\{S_j\} = \text{avg}\{f(x_j, y_j)\} + \text{avg}\{\epsilon_j\} \tag{3}$$

where the average is taken over all sensors (x_j, y_j) that belong to cluster head i.

Note that the sensors are assumed to be uniformly distributed over the region of interest R. If we assume that the number of sensors belonging to a cluster head is large, then by the law of large numbers (LLN),

$$\text{avg}\{f(x_i, y_i)\} \approx \frac{1}{|A_i|} \int_{A_i} f(x_i, y_i) dxdy = c_i \quad \text{say,}$$

where $|A_i| \equiv$ = the area of A_i. Hence, the total loss of information can be approximated by

$$\Delta_k \approx \frac{\sigma^2 k M^2}{N} + \sum_{i=1}^{k} \left(\int_{A_i} c_i - f(x, y) \right)^2 dxdy. \tag{4}$$

Now note that

$$\int_{A_i} (c_i - f(x_i, y_i))^2 dxdy = |A_i| \text{Var}(f(U_i, V_i),$$

where (U_i, V_i) is a random variable that is uniformly distributed over the region A_i. Hence, using the propagation of error formula

[9], we can further approximate Var$(f(U_i, V_i))$ and hence, Δ_k by

$$\begin{aligned}\Delta_k &= \frac{\sigma^2 kM^2}{N} + \sum_{i=1}^{k} |A_i| \Big[\{D_1 f(a_i, b_i)\}^2 \text{Var}(U_i) \\ &\quad + \{D_2 f(a_i, b_i)\}^2 \text{Var}(V_i)\} \\ &\quad + 2\big((D_1 f(a_i, b_i))(D_2 f(a_i, b_i))\big) \text{Cov}(U_i, V_i) \Big], \end{aligned} \tag{5}$$

where $D_j f(x, y)$ is the jth partial derivative of $f(x, y)$ and where$(a_i, b_i) = (EU_i, EV_i)$ = the center of gravity of A_i. For a circular region A_i of radius r,

$$\text{Var}(U_i) = \text{Var}(V_i) = \frac{r^2}{\pi} \quad and \quad \text{Cov}(U_i, V_i) = 0.$$

Hence for the circular coverage region A_i, we have

$$\Delta_k = \frac{\sigma^2 kM^2}{N} + \sum_{i=1}^{k} \pi r^2 \cdot \frac{r^2}{\pi} \Big[(D_1 f(a_i, b_i))^2 + (D_2 f(a_i, b_i))^2 \Big]. \tag{6}$$

As in LEACH, supposing the coverage regions A_i have the same area, the radius r is obtained as

$$k\pi^2 = M^2, \quad \text{i.e., } r = \sqrt{\frac{M^2}{k\pi}}, \tag{7}$$

which may be substituted in (16).

Next we consider the important special cases where the signal $f(x, y)$ has a linear trend and we derive the specific formula for Δ_k . In this case, $D_j f(x, y)$ is a constant for $j = 1, 2$, and from (16) and (17), Δ_k is given by

$$\begin{aligned}\Delta_k &\approx \frac{\sigma^2 kM^2}{N} + k(\pi r^4) \left[\frac{d_1^2}{\pi} + \frac{d_2^2}{\pi} \right] \\ &= \frac{M^4}{\pi^2 k}(d_1^2 + d_2^2) + \frac{\sigma^2 kM^2}{N}. \end{aligned} \tag{8}$$

From (18), it follows that Δ_k decreases as the number of nodes increases. This is intuitively clear as a large number of cluster heads provides a finer approximation to the true signal and therefore reduces the loss of precision. Although an exact expression like (17) is not available in general, a similar conclusion holds.

3.2 The entropy loss ratio metric

The cluster heads fuse data from the member nodes and then transmit the fused data to the base station.If there are fewer cluster heads, more energy is saved. But energy is saved at the expense of information. The base station only knows the aggregate information fused at the cluster heads. Thus, if there are more cluster heads, then the base station has more information about the local regions and vice versa. Therefore, the information loss is an important parameter that is a measure of the efficiency of the algorithm. The amount of information loss is dependent on the number of clusters in the network. Next, we formally determine an expression for the total loss of information.

In Shannon's Information Theory, one of the ways of measuring the information content of a signal is by using the entropy function. In this paper, we choose to use the expected entropy as a measure of the information content of the data.

In LEACH, each node senses and then transmits l bits of data. The cluster head then fuses the $\frac{N}{k}$ l bits of data into m bits of data where m is less than $\frac{N}{k} l$.

At each ordinary node, entropy of l bit signal sensed is given by

$$E_l = -\sum_{n=1}^{l} p_n log(p_n) \tag{9}$$

where p_n is the probability with which the nth bit will take 0 or 1 value.

Therefore, entropy of all the l bit signals sensed at the N nodes can be denoted by NE_l.

At a cluster head, aggregated data is m bits in length. The entropy of this data is given by

$$E_m = -\sum_{n=1}^{m} p_n log(p_n) \tag{10}$$

We define information loss ratio as the ratio of the information loss for data aggregation and the total information content of the sensed data. This can be formulated by the following expression

$$\Delta k = \frac{NE_l - kE_m}{NE_l} \tag{11}$$

A smaller value of Δk corresponds to a better quality of the aggregated data and vice versa.

3.3 The variance of estimation

In many scenarios the sensor nodes are used to monitor a variable in the deployment area. The variable monitored can be modeled using a function of the location of the sensors. Let this function be represented by $f(x, y)$ where (x_i, y_i) is the location of the ith sensor in the deployment area. Let there be i sensor nodes in the MxM deployment area.The sensors use the clustering protocol to report their readings to the CHs. The CHs perform data aggregation using a data aggregation function and send values to the base station for the clusters. In a given cluster, the value that the CH sends to the base station is the value of the estimated signal for all the members. Let the estimated signal be denoted by $f'(x, y)$. We define the variance of estimation, v as follows:
$v = \sum_i (f(x, y) - f'(x, y))^2$ for all i.
A larger value of this metric indicates a poor quality of signal estimation and vice versa.

4. SIMULATION RESULTS

We use a 100m $\times$ 100m deployment area to simulate the sensor network environment.The values of the different parameters used are the following: $E_{fs} = 10 * 10^{-12} Joules$; $Emp = 0.0013 * 10^{-12} Joules$; M=100;N=100; $E_{elec} = 50 * 10^{-9} Joules$; $E_{da} = 5 * 10^{-9} Joules$; $T_{nchtx} = 0.5$; $T_{chtx} = 1$;$T_{chrx} = 1$;d=100;l=1; Δ d=0.05 sec; $d_{fs} = 0.5sec$;$d_1 = d_2 = 1$.

We obtain the different values of the energy consumed per round, the delay and the information loss and the entropy ratio for different configurations of the clusters.The metrics are plotted against the number of clusters. The number of clusters are plotted along the x axes and the metrics are plotted along the y axes.

In Figure 1 the energy consumed in each round of LEACH is plotted against the number of clusters from a physical layer standpoint.

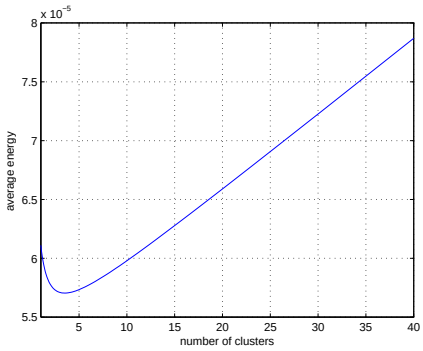

Figure 1: Average Energy per round in LEACH for 100 nodes

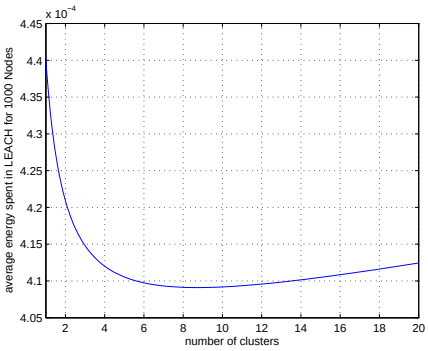

Figure 2: Average Energy per round for 1000 nodes

In Figure 3 we plot energy x delay X information loss against the number of clusters. In Figure 4 we plot the energy x entropy ratio against the number of clusters. We see that the optimal number of clusters is different in each case when different performance measures are used.

In Figure 5 we plot histogram of the variance of the sensed signal from the true signal observed over 200 rounds for the LEACH protocol. Our true signal is of the form

$$f(x,y) = \Big(\sin[2\pi x/100]\Big)\Big(\sin[2\pi y/100]\Big). \qquad (12)$$

5. CONCLUSIONS AND FURTHER WORK

In this paper, we have show three ways to measure the quality of information obtained in a sensor network. The information loss metric that we have formulated is a statistical method to estimate the quality of information. Our second metric, the entropy ratio metric based on Shannon's information theory gives as a number that indicates the quality of information obtained. Our third method, is very simple and is based variance. This deterministic method essentially calculates the mean square error between the true and the estimated signal over all the nodes. This simple method provides an effective way to measure the amount of information lost. In future, we plan to compare these three methods and apply them in different scenarios to see which metric is suitable to be used in what scenario.

6. REFERENCES

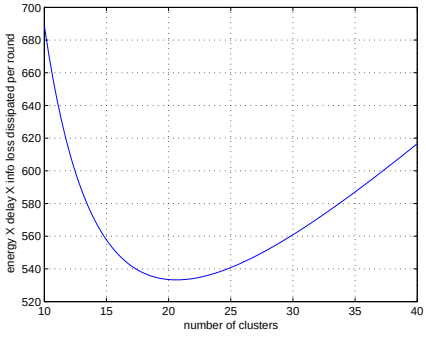

Figure 3: Average Energy x Delay x info loss per round in LEACH for 100 nodes

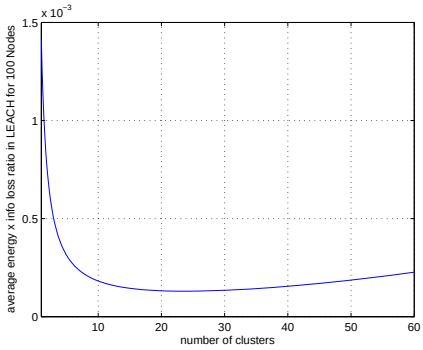

Figure 4: Average Energy x entropy ratio per round in LEACH for 100 nodes

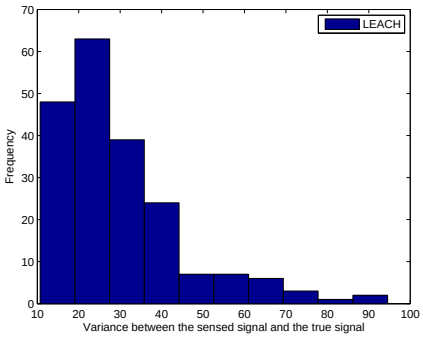

Figure 5: Variance of sensed and true signal in LEACH for 100 nodes

[1] W.Heinzelman and H.B. Chandrakasan, *"An application-specific protocol architecture for wireless microsensor networks," IEEE Transactions on Wireless Communications* . Supplement to IEEE 802.11 Standard-1999 Edition, 2002.

[2] Stephanie Lindsey and Cauligi S. Raghavendra, *"PEGASIS: Power-Efficient Gathering in Sensor Information Systems"* . Aerospace Conference Proceedings, 2002. IEEE, 2002.

[3] Mao Ye; Chengfa Li; Guihai Chen; Wu, J., *"EECS: an energy efficient clustering scheme in wireless sensor networks "*. Performance, Computing, and Communications Conference, 2005. IPCCC 2005. 24th IEEE International, 2005.

[4] P.Banerjee and D. Jacobson, *"Optimal configuration of clustering protocols for sensor networks"* . Proceedings of PDCS 2006,Dallas, TX, Nov 2006, 2006.

[5] A.Manjeshwar and D.P. Agrawal, *"TEEN: A Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks"* . Ist International Workshop on Parrallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, 2001.

[6] A.Manjeshwar and D.P. Agrawal, *"APTEEN: A Hybrid Protocol for Efficient Routing and Comprehensive Information Retrieval in Wireless Sensor Networks"* . Proceedings of the Parrallel and Distributed Processing Symposium, 2002.

[7] T.Pham; E.Kim and M. Moh, *"On Data Aggregation Quality and Energy Efficiency of Wireless Sensor Network Protocols-Extended Summary"*. "Proceedings of the First International Conference on Braodband Networks (BROADNETS'04)", 2004.

[8] Tolstikov, A.; Biswas, J.; Chen-Khong Tham , *"Data loss regulation to ensure information quality in sensor networks"* . Proceedings of Intelligent Sensors, Sensor Networks and Information Processing Conference, 2005.

[9] Stephen B. Vardeman , *"Statistics for Engineering Problem Solving"*. published by IEEE as ISBN 0-7803-1118-3, 1994.