Poster Abstract: A Clustering Method that Uses Lossy Aggregation of Data

Apoorva Jindal
Department of Electrical Engineering - Systems
University of Southern California
Los Angeles CA-90089
apoorvaj@usc.edu

Konstantinos Psounis
Department of Electrical Engineering - Systems
University of Southern California
Los Angeles CA-90089
kpsounis@usc.edu

ABSTRACT

Wireless sensor networks are characterized by dense deployment of sensor nodes which collectively communicate sensed data to the sink. However, due to the spatial correlation between sensor observations, it is not necessary for every node to transmit its data. We propose a clustering method which exploits the above observation. We do not make any assumption on the nature of data, and hence the algorithm will be valid for a broad range of conditions. The paper shows how to calculate the optimal cluster size. We also discuss the structure of the complete architecture which is still under development.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design

General Terms

Algorithms, Performance

Keywords

Cluster, Wireless Sensor Networks, Lossy Data Aggregation, Spatial Correlation, Energy Distortion Tradeoff

1. INTRODUCTION

We focus on the wireless sensor network applications which require periodic data gathering from all the nodes. Since these sensors are densely deployed and they sense a common phenomenon, it is expected that a high degree of spatial correlation will exist in the sensor network data. Vuran et al [5] have argued that due to the presence of this spatial correlation, it is not necessary for every node to transmit its data. Instead, a smaller number of sensor measurements might be adequate to communicate the event features to the sink within a distortion constraint. The authors in [5] based their protocol on the assumption that all nodes are sensing data from a single source and that the sensed data is jointly gaussian.

We propose a clustering method which exploits spatial correlation by allowing transmission of data to the sink from only one node (called the representative node) in the cluster. This enables other nodes in the cluster to go into the sleep mode. The representative node of a cluster is not fixed; it will keep changing with time to increase the network lifetime. Once formed, the cluster itself is not going to change, only the representative node of the cluster will change. There have been clustering based protocols proposed in the past, e.g. [3] [1] [2], but all of them assume lossless aggregation of data and they calculate the cluster size by minimizing the communication energy cost.

Our main contribution is the introduction of lossy aggregation in clustering based protocols. In this paper, we give an algorithm to calculate the optimal cluster size given a distortion constraint. Unlike other schemes, the proposed algorithm is independent of the nature of data.

SenSys'04, November 3–5, 2004, Baltimore, Maryland, USA.
ACM 1-58113-879-2/04/0011.

2. BASIC IDEA AND ASSUMPTIONS

We assume that the nodes are located on a regular rectangular grid structure. We also assume that the nodes know their locations or their coordinates. The sink is assumed to be located at the origin.

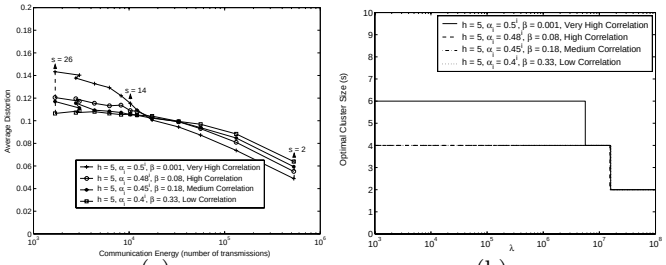
Before proceeding, we should have a model to be able to generate synthetic data at these nodes. Jindal et al [4] proposed a model to generate synthetic data representing a wide variety of correlation structures. To evaluate our algorithm, we use the provided tool to generate synthetic data.

The cluster structure is a square of size s. Only one node (the representative node) in the cluster will transmit its data value to the sink. Each node in the cluster has an equal probability of being selected as the representative node. The sink will assume that the data value at all other nodes in the cluster is the same as the value it receives from the representative node of the cluster. This will lead to an error or distortion of the data. This distortion averaged over all the nodes is an important metric. Normally the network will put a bound on the acceptable average distortion.

As the cluster size s increases, less number of nodes transmit to the sink which will lead to a decrease in energy (less number of transmissions) as well as an increase in average distortion. This suggests a tradeoff between distortion and energy. We display this tradeoff for different correlation structures in Figure 1(a). The grid is assumed to have 128×128 nodes. The cluster structure is a square of side s. s is varied from 2 to 26 and the average distortion and average communication energy costs are measured and plotted. The plot shows that a decrease in energy is accompanied with an increase in distortion until the cluster size becomes so big that the data in the cluster become uncorrelated.

Based on the energy-distortion tradeoff, we give an algorithm to calculate the optimal cluster size in the next section. We assume that the sink is not energy constrained and it has the processing capability to perform complex calculations.

3. HOW TO FIND THE OPTIMAL CLUSTER SIZE



(a) (b)

Figure 1: (a) Energy-Distortion tradeoff. Average distortion vs average communication energy costs for different correlation structures (b) Optimal cluster sizes for different values of λ for different correlation structures. $D_{max} = 0.09$

In the first time slot, all the nodes send their sensed data and their locations to the sink. Thus, the sink has complete information about the network topology and the data values at the nodes. The sink uses this information to get an estimate of the average distortion $\hat{D}$.

Let s be the cluster size, $x_i, 1 \leq i \leq s^2$, be the data value at node i in the cluster and d_{ij} be the distance between the nodes x_i and x_j. Then, if $V_{ij} = \left(\frac{x_i - x_j}{x_i}\right)^2$, the average distortion for the scheme described in Section 2 can be estimated by,

$$\hat{D} = \frac{1}{s^2} \sum_{i=1}^{s^2} \frac{1}{s^2} \sum_{j=1}^{s^2} V_{ij}, \tag{1}$$

For an isotropic stationary random process, the value of V_{ij} will depend only on d_{ij}, so we denote V_{ij} by $V(d_{ij})$. The sink can estimate the value of $V(d)$ by,

$$\hat{V}(d) = \frac{1}{m(d)} \sum_{1}^{m(d)} \left(\frac{x_i - x_j}{x_i}\right)^2, \tag{2}$$

where $m(d)$ is the number of points which are at a distance d from each other, ie. the sum is over all points for which $d_{ij} = d$. The sink can estimate the average distortion for a given cluster size by using Equations (1) and (2).

Since the sink knows the network topology, it can easily calculate the energy consumption. We assume that it takes one unit of energy to transmit one data value over one hop. Let the dimensions of the grid be $N \times N$ and let $N = a \cdot s$, that is a is the ratio of the grid size over the cluster size. Using simple calculations, it can be shown that the average energy to collect data at the sink is $E = a^2(as - 1)$.

The sink can calculate the optimal cluster size by solving the constrained minimization problem,

$$\min_{s}(E + \lambda D) \text{ s.t. } D \leq D_{max},$$

where D_{max} denotes the maximum average distortion that can be tolerated. The value of λ is an engineering decision which will be governed by whether the system pays more attention to distortion or energy. A higher value of λ indicates more attention towards distortion. The value of optimal cluster sizes (s) for different values of λ for $D_{max} = 0.09$ is plotted in Figure 1(b).

4. FUTURE WORK

After determining the optimal cluster size, the next step is to form a cluster. Since the sink knows the complete topology, either a centralized algorithm running at the sink or any of the numerous distributed clustering algorithms [6] can be used to form a cluster.

Now one needs to decide which node in the cluster will be the representative node. According to our scheme, each node in the cluster has an equal probability of being selected as the representative node. This can be achieved by a predetermined static schedule in which every node wakes up for a time slot, transmits its data value and then goes to sleep. It wakes up again after every other node in the cluster got a chance to transmit. Another way is to pick the representative node based on the residual energy. The selected node transmits for a few time slots and then triggers another selection. The static schedule requires time synchronization between all the nodes in the cluster and the residual energy based representative node selection method has the overhead of exchanging control messages. We propose to study both the schemes and get a better understanding of the issues associated with both.

Next, we want to determine how many nodes in the cluster should go to sleep simultaneously. Since, a node which is transmitting uses up more energy than a node which is just listening which in turn uses much more energy than a node which is sleeping, it makes sense to let a large number of nodes goto sleep simultaneously. This will lead to connectivity issues for large cluster sizes. One way to solve this problem is to increase the transmission range of the representative node; another way is to keep more than one node in the cluster awake. This suggests a tradeoff between the power level used for transmission and the number of awake nodes in the cluster (only one node in the cluster will transmits its data, others will be awake to maintain connectivity). We expect, since the power decays as the square of distance, it will be more efficient to keep more than one node awake in the cluster though only one of them will transmit its data.

5. CONCLUSIONS

We propose a clustering method using lossy aggregation. Lossy aggregation introduces distortion in the system. We present an algorithm to determine the optimal cluster size under a distortion constraint. Unlike previous algorithms, our algorithm does not assume any correlation pattern for the data.

6. REFERENCES

[1] S. Bandyopadhyay and E. Coyle. An energy-efficient hierarchical clustering algorithm for wireless sensor networks. In *Proceedings of IEEE Infocom'03*, Apr. 2003.

[2] C. Chiasserini, I. Chlamtac, P. Monti, and A. Nucci. Energy efficient design of wireless ad hoc networks. In *Proceedings of European Wireless*, Feb. 2002.

[3] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan. An application-specific protocol architecture for wireless microsensor networks. *IEEE Transactions on Wireless Communications*, 1:660–670, Oct. 2002.

[4] A. Jindal and K. Psounis. Modelling spatially correlated sensor networks data. In *Technical Report, CENG-2004-10, USC*, 2004.

[5] M. C. Vuran and I. F. Akyildiz. Spatial correlation-based collaborative medium access control in wireless sensor networks. *submitted for publication, 2004.*

[6] O. Younis and S. Fahmy. Distributed clustering in ad-hoc sensor networks: A hybrid, energy-efficient approach. In *Proceedings of IEEE Infocom'04*, Mar. 2004.