A Communication-Efficient Distributed Clustering Algorithm for Sensor Networks

Amirhosein Taherkordi
Department of Informatics
University of Oslo
amirhost@ifi.uio.no

Reza Mohammadi
IT Eng. Dept.
Faculty of Engineering
Tarbiat Modares University
r.mohammadi@modares.ac.ir

Frank Eliassen
Department of Informatics
University of Oslo
frank@ifi.uio.no

Abstract

Sensor networks usually generate continuous stream of data over time. Clustering sensor data as a core task of mining sensor data plays an essential role in analytical applications of sensor networks. Although several algorithms have been proposed to address the problem of distributed clustering, in the domain of sensor networks these algorithms face major new challenges such as limited communication bandwidth and constraints in power supply, and storage resources. Moreover, previous studies about clustering in sensor networks have mostly focused on clustering sensor nodes and designing better network topology for the purpose of energy conservation rather than clustering sensor data for future analytical purposes. In this paper a communication efficient distributed algorithm is proposed for clustering sensory data. This approach addresses the limited bandwidth issue through summarized transmissions. Furthermore communication efficiency of the algorithm contributes to reduced power consumption. Time efficiency of the algorithm is evaluated through simulation experiments and the results are presented.

1. Introduction

Recent advances in wireless communications and miniaturization of hardware components have enabled the development of low-cost, low-power, multifunctional and intelligent sensor nodes. These devices are small in size and communicate in short distances over an RF (radio frequency) channel. These tiny nodes, which consist of components for sensing, processing, and communicating data, realize the objectives of sensor networks.

Wireless sensor networks promise significant improvements over traditional sensors. A Wireless Sensor Network (WSN) is composed of a large number of integrated sensor nodes that are densely deployed either inside the phenomenon or very close to it, and collaborate through a wireless network in collecting environmental information or reacting to specific events [1, 2]. Some typical applications of sensor networks include medical monitoring, natural event monitoring, object tracking, monitoring product quality, and combat field reconnaissance [3].

The majority of sensor networks applications fall into the querying class of applications in which it is required to continuously collect and integrate data for future analysis and mining. Special characteristics of WSNs lead us to new research challenges in mining data sensed in sensor network. Sensors have serious resource constraints including power supply, communication bandwidth, processor capacity and storage. Moreover, data which is sensed in WSNs is generated and streamed continuously. These challenges make traditional mining techniques inapplicable in sensor networks [4], as the traditional mining process is usually centralized, computationally expensive, and focuses on disk stored data.

Clustering is a basic task in data mining. It is usually used as a preparatory step in data mining for future data analysis. In this paper, a distributed version of K-Means clustering algorithm is proposed for the purpose of clustering sensors' data in a wireless sensor network.

In the following a brief overview of cluster analysis in data mining and its applicability in sensor networks is given in section 2. Section 3 presents the details and analysis of the proposed algorithm. Section 4 provides some experimental and simulations and finally section 5 concludes the paper.

2. Data Clustering in Wireless Sensor Networks

In this section, we describe our problem in detail. In order to make the problem statement more clear, we first review some concepts related to data clustering in data mining and physical structure of a cluster-based sensor network.

2.1. Cluster Analysis in Data Mining

Data Clustering is defined as the process of organizing a set of objects into groups or clusters so that objects within a cluster have the most similarity to one another and the most dissimilarity to objects in other clusters. This process is commonly seen as a tool for discovering structure in data. The dissimilarity or distance between objects is often measured by some distance function such as Euclidean distance, Manhattan distance, etc [5, 6].

There exist many algorithms for data clustering developed in recent decades. Many of these algorithms are designed to deal with data which is stored in a traditional database management system. But in a distributed environment we intend to analyze data which is distributed across multiple sites. Because of constraints such as limited bandwidth, privacy, data analysis in a distributed environment faces many challenges [7]. There are some algorithms proposed for the problem of cluster analysis in distributed environments. In [7], a number of distributed clustering algorithms are well discussed and grouped based on rounds of message passing between data sites and a central site.

A wireless sensor network is a distributed environment consisting of a large number of low-power sensors. So, data clustering algorithms for sensor networks should address challenges mentioned above as well as sensor constraints in communication and computation.

2.2. Cluster-Based Sensor Networks

Due to energy constraints in sensor networks, managing sensors' energy in a wisely manner is a key requirement to achieve prolonged network lifetime. Among many studies on considering sensors' energy, clustering sensors into groups has proven to be a more efficient and adaptive approach [8]. Cluster-based mechanism [9] is adopted for node communication and routing. In a clustered sensor network, sensors communicate data only to cluster-heads and then the cluster-heads communicate the aggregated data to a processing center or a base station usually called sink. The base station can be a specialized device or just one of the sensors [10]. The essential operation in sensor node clustering is to select a set of cluster heads among the nodes in the network, and cluster the rest of the nodes with these heads. Cluster-heads are selected according to some negotiated rules. Usually, more powerful nodes in the topology play the role of cluster heads and other nodes are responsible for sensing data and forwarding them to cluster nodes.

Communication is usually the main source of energy consumption in sensors [9]. The amount of energy reduction of a sensor during a message passing is highly dependent on the distance between the source and the destination of the message. Since in a clustered sensor network, sensors only communicate data to cluster-heads over smaller distances, the total energy consumption in the network will be much lower than the situation in which every sensor communicates directly to the base station.

There are already a lot of works related to clustering in sensor networks [11]. Several heuristic algorithms have been proposed to choose cluster-heads. Moreover, some approaches attempt to estimate the optimal number of cluster-heads in a sensor network [8].

The main goal of sensor network clustering is to balance energy consumption over the whole network. LEACH, ASCENT, SPAN, GAF, ACE and HEED all try to preserve and balance the energy dissipation of the network using cluster-based architectures, thus prolonging the network lifetime [11].

2.3. Problem Definition

Suppose that N sensor nodes are dispersed uniformly in a field to detect D attributes. Moreover, assume that these sensors are clustered into groups resulting in a clustered sensor network with determined cluster-heads. Our goal is to develop an algorithm that clusters the data generated by sensors in a communication-efficient way. In other words, the algorithm will determine data clusters in a cluster-based sensor network. Each data cluster contains sensor nodes which are similar in data attributes they detect. The problem is illustrated in figure 1.

Previous studies about clustering in sensor networks have just mostly focused on clustering sensor nodes and designing better network topology for the purpose of energy conservation. There have been few attempts to address the problem of clustering sensory data. Notable exceptions are the ones proposed in [4, 12] for clustering in a multi-dimensional sensor dataset. But these approaches do not consider the physical limitations of the sensor networks. In this paper, our data clustering approach is aimed at considering the physical aspects as well.

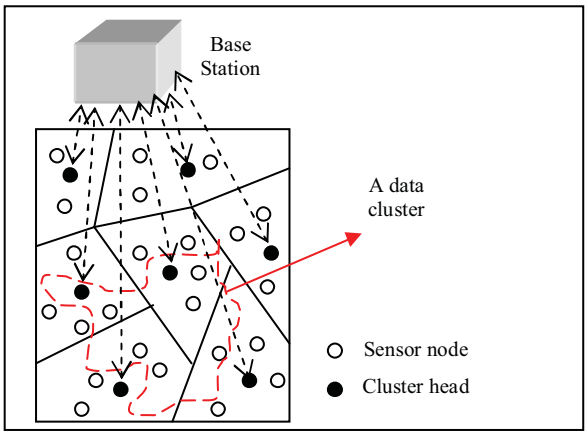

Figure 1. Data clustering on a clustered sensor network

3. The Algorithm

in the previous sections, communication efficiency is a critical issue that must be considered in the algorithm. The cost of transmitting a bit is higher than a computation [10]. In order to achieve the goal of decreasing the communication cost, a common and good idea is to communicate only summarized and sufficient information during the algorithm runtime. In addition, the algorithm should try to perform computations locally in clusters as much as possible. In other words, we aim at applying the computation power of cluster-heads to reduce the overall runtime of the algorithm while controlling the communication overhead. Based on the schema proposed in [6] for parallelizing a family of center-based clustering algorithm, we have used a distributed version of the K-Means algorithm. The reason why we choose this algorithm is due to its scalability, simplicity in implementation and linear computational complexity. Moreover, this is an exact technique as it does not use any approximation[6]. The quality of the clustering produced by the distributed approach is the same as that of the clustering from centralized data while keeping the communication and computation costs low.

Before entering the algorithm procedure, we make the following assumptions:

1) The network is homogenous i.e. all sensors are similar in the type of data items they detect and have the same amount of initial energy.
2) The cluster-heads are powerful enough and have the extra burden of performing required computations and long range transmissions to the base station.
3) Each sensor node can communicate directly only with other nodes in its cluster.

We use the sum of the mean-square error (MSE) of each data point to its center as popular criterion function to control the number of iterations in the K-Means algorithm.

The steps of the algorithm are as follows:

1) Each sensor transmits its data to its cluster-head;
2) The base station adopts the K initial center locations (M_1, M_2,…, M_k) arbitrarily;
3) Repeat:
   a. The base station transmits current center locations to cluster-heads.
   b. Each cluster-head assigns its sensor nodes to the closest center;
   c. Each cluster-head sends back the following data items to the base station:
      i. For each center M_i (i:1..k), the count and vector sum of its local sensory data points assigned to it.
      ii. The sum of the squared distance from each local point to its center.
   d. The base station updates the cluster means;
4) Until the criterion function converges (no change);

3.1. Communication Cost Analysis

In the first step of the algorithm, all sensors send their data to cluster-heads, therefore in next steps each cluster-head has a local copy of data gathered by its sensors. We also assume that each cluster-head as a powerful sensor is able to detect data and has its own data. In order to evaluating the communication cost, we calculate the total number of data transmissions which occurs during message passing between cluster-heads and the base station. For this purpose, we consider the main loop in the algorithm (step 3). The communication cost based on transmissions in this step is:

$$T = T_1 + T_2,$$

where T_1 and T_2 are the total transmissions which occur in steps 3.a and 3.c respectively. In step 3.a, the cost of transmitting center locations to cluster-heads is:

$$T_1 = KDH,$$

where K is the number of center locations, H is the number of cluster-heads and D is the dimensionality of the data.

In step 3.c, cluster-heads send back some statistics to the base station. The base station uses the statistics to update center locations. The transmission cost of sending these statistics is:

$$T_2 = H(KD + K + 1),$$

Consequently, T is given by:

$$T = KDH + H(KD + K + 1) = H(2KD + K + 1),$$

and hence the total number of transmissions which occurs in step 3 is:

$$T_{total} = T.t = Ht(2KD + K + 1), \qquad (1)$$

where t is the number of iterations of the main loop and the communication cost per cluster per iteration will be:

$$T_{cluster/iteration} = 2KD + K + 1 \qquad (2)$$

4. Simulations and Experimental Results

We evaluate the efficiency of the algorithm via simulations. For this purpose, a program is written with Java programming language. We simulate a network with several sensor nodes and some cluster-heads, which are connected to a base station. Each sensor is assigned to one cluster-head. The sensor detects two attributes and sends its data to the cluster-head (D=2). The sensors' data is randomly generated. The number of sensor nodes and desired data clusters (parameter K in the algorithm) are taken as input parameters of the program. The program automatically estimates the optimal number of cluster heads based on the method proposed in [8]. In their method, the optimal number of cluster-heads is determined based on minimizing the total energy consumption in the network. This optimal number is proportional to the square root of N. For large numbers of N, we can take $H \cong \sqrt{N}$ as a rule of thumb. Once the optimal number of cluster-heads is estimated and cluster-heads are determined, the sensor nodes are assigned to cluster-heads uniformly. After configuring the network, the algorithm is run.

For different values of N and K, we run the program several times to get the average number of transmissions over the network during each test.

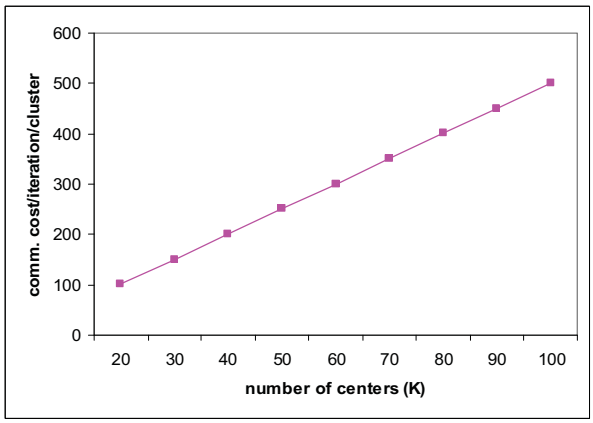

Figure 2. Number of transmissions

Figure 2 depicts changes of the communication cost per cluster per iteration for different values of K ranging from 20 to 100. According to (2), the communication cost is independent of the number of sensors (N) and increases linearly by increasing the number of centers. Note that in real life usage, the dimensionality of the sensor data is usually low (e.g. in our simulation D is 2), so the parameter K has the most effect on T in (2).

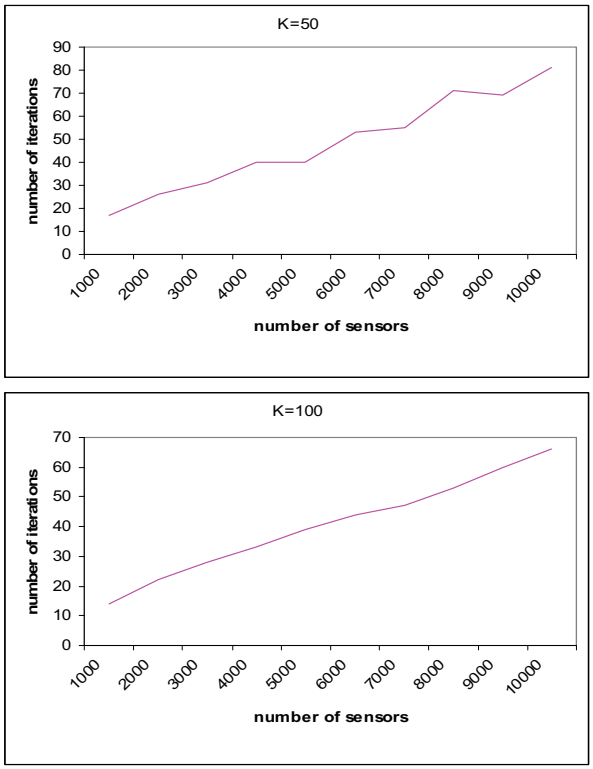

Figure 3. Number of iterations

As (1) shows, the total number of transmissions for each cluster directly depends on t, the number of iterations of the algorithm. Our simulations results shows that the most dominant parameter which affects t is N. Figure 3 illustrates number of iterations for K=50 and K=100 center locations when changing the number of sensors.

5. Conclusions and Future Work

In this paper, we have addressed the problem of data clustering in sensor networks. We proposed a distributed version of the K-Means algorithm to work on clustered network of sensors. The algorithm does not use any approximations, so the quality of the clustering is the same as centralized clustering. To

reduce communication overhead the only summarized and sufficient information is transmitted in each iteration of the algorithm. Although the algorithm keeps the clustering quality, due to iterations, it requires multiple rounds of message passing between cluster-heads and the base station. As simulation result shows, this may have serious effect on communication efficiency when the number of sensors is relatively high.

As future work, we plan to investigate distributed data clustering methods to improve the communication efficiency as well as developing methods to address other main challenges in sensor networks such as handling data streams and constraints on computing resources.

6. References

[1] F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A Survey on Sensor Networks," IEEE Communications Magazine, pp. 102, 2002.

[2] D. Culler, D. Estrin, and M. Srivastava, "Overview of Sensor Networks," IEEE Computer magazine, pp. 41-49, 2004.

[3] M. Ilyas, and I. Mahgoub, HANDBOOK OF SENSOR NETWORKS: COMPACT WIRELESS AND WIRED SENSING SYSTEMS. CRC Press, 2004.

[4] X. Ma, D. Yang, S. Tang, Q. Luo, D. Zhang, et al., "Online Mining in Sensor Networks," Lecture Notes in Computer Science, pp. 544-550, 2005.

[5] J. Han, and M. Kamber, DATA MINING – CONCEPTS AND TECHNIQUES. Morgan Kaufmann Publishers, 2004.

[6] G. Forman, and B. Zhang, "Distributed data clustering can be efficient and exact," SIGKDD Explorations, vol. 2(2), pp. 34-38, 2000.

[7] J.C.d. Silva, C. Giannella, P. Bhargava, H. Kargupta and M. Klusch, "Distributed data mining and agents," Engineering Applications of Artificial Intelligence, vol. 8, pp. 791-807, 2005.

[8] H. Kim, S. W. Kim, S. Lee, and B. Son, "Estimation of the Optimal Number of Cluster-Heads in Sensor Network," LNAI, pp. 87-94, 2005.

[9] S. Ghiasi, A. Srivastava, X. Yang, and M. Sarrafzadeh, "Optimal Energy Aware Clustering in Sensor Networks," Sensors, vol. 2, pp. 258-269, 2002.

[10] S. Bandyopadhyay, and E.J. Coyle, "An Energy Efficient Hierarchical Clustering Algorithm for Wireless Sensor Networks," IEEE INFOCOM, 2003.

[11] S. Lee, J. Yoo, and T. Chung. "Distance-based Energy Efficient Clustering for Wireless Sensor Networks," in Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks, 2004.

[12] G. Cheng, and A. Zell, "CL2: A Multi-dimensional Clustering Approach in Sensor Networks," Lecture Notes in Computer Science, vol. 3289, pp. 234-245, 2004.