Online Mining in Sensor Networks*

Xiuli Ma[1], Dongqing Yang[1], Shiwei Tang[1], Qiong Luo[2], Dehui Zhang[1], Shuangfeng Li[1]

[1](School of Electronics Engineering and Computer Science, National Laboratory on Machine Perception, Peking University, Beijing, China, 100871)
[2](Department of Computer Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong)
Email: {xlma, dqyang, dhzhang, sfli}@db.pku.edu.cn, tsw@pku.edu.cn, luo@cs.ust.hk

Abstract. Online mining in large sensor networks just starts to attract interest. Finding patterns in such an environment is both compelling and challenging. The goal of this position paper is to understand the challenges and to identify the research problems in online mining for sensor networks. As an initial step, we identify the following three problems to work on: (1) sensor data irregularities detection; (2) sensor data clustering; and (3) sensory attribute correlations discovery. We also outline our preliminary proposal of solutions to these problems.

1 Introduction

Recent technology advances have enabled the development of small, battery-powered, wireless sensor nodes [2][6][20]. These tiny sensor nodes, equipped with sensing, computation, and communication capabilities, can be deployed in large numbers in wide geographical areas to monitor, detect and report time-critical events. Consequently, wireless networks consisting of such sensors create exciting opportunities for large-scale, data-intensive measurement and surveillance applications. In many of these applications, it is essential to mine the sensor readings for patterns in real time in order to make intelligent decisions promptly. In this paper, we study the challenges, problems, and possible solutions of online mining for sensor networks.

Research on data mining has been fruitful; however, online mining for sensor networks faces several new challenges. First, sensors have serious resource constraints including battery lifetime, communication bandwidth, CPU capacity and storage [15]. Second, sensor node mobility increases the complexity of sensor data because a sensor may be in a different neighborhood at any point of time [7][19]. Third, sensor data come in time-ordered streams over networks. These challenges make traditional mining techniques inapplicable, because traditionally mining is

* This work is supported by the National Grand Fundamental Research 973 Program of China under Grant No.G1999032705

centralized, computationally expensive, and focused on disk-resident transactional data.

In response to these challenges, we propose to develop mining techniques that are specifically geared for sensor network environments. Our goal is to process as much data as possible in a decentralized fashion while keeping the communication, storage and computation cost low.

As a start point, we propose three operations of online mining in sensor networks: (1) detection of sensor data irregularities, (2) clustering of sensor data, and (3) discovery of sensory attribute correlations. These mining operations are useful for practical applications as well as for network management, because the patterns found can be used for both decision making in applications and system performance tuning. For example, irregularities in sensory data are of interest of monitoring applications. In addition, for this kind of applications, the communication cost can be reduced if only abnormal sensory values, as opposed to all values, need to be transmitted.

The rest of the paper is organized as follows. In Section 2, we illustrate the need for online mining in sensor networks using an example. Our design of online mining in sensor networks is presented in Section 3. We present related work in Section 4 and conclude in Section 5.

2 A Motivating Example

There have been initial applications of sensor networks on wild life habitat surveillance [15], battlefield troop coordination, and traffic monitoring. As a motivating example for online mining in sensor networks, we describe a possible application on wild giant panda monitoring and protection in China. Suppose weather sensors are deployed throughout a panda habitat and wearable sensors attached to the pandas in the habitat. The sensors acquire sensor data on attributes such as temperature, light, sound, humidity, and acceleration. In addition, there is a panda of interest named Huanhuan. The following is a few questions that scientists on site may ask:

1) Is Huanhuan having any abnormal symptoms compared with its past data? What other pandas are having abnormal symptoms and what are these abnormal symptoms?

2) What pandas have a similar physical status to Huanhuan's? What pandas are similar to one other and on what sensory attributes are they similar?

3) What attributes of Huanhuan's are correlated and how are they correlated? What attributes of pandas are correlated with the humidity of their habitat? What symptoms of pandas are correlated to what attributes of the habitat?

Answers to these questions are important for habitat maintenance and panda protection, and these questions all require online mining of sensor data.

3 Design of Online Mining

In this section, we identify the following three problems of online mining in sensor networks and outline our preliminary solutions.

3.1 Detection of Sensor Data Irregularities

The problem of irregularities detection is to find those sensory values that deviate significantly from the norm. This problem is especially important in the sensor network setting because it can be used to identify abnormal or interesting events or faulty sensors.

We break this problem into two smaller problems. One is to detect irregular patterns of multiple sensory attributes and the other to detect irregular sensory data of a single attribute with respect to time or space. The irregular multi-attribute pattern detection problem has the assumption that there are some normal patterns among multiple sensory attributes, which is true in some natural phenomena. Once these normal patterns are broken somewhere, the irregularity is detected and reported. In contrast, the irregular single-attribute sensor data detection problem examines the temporal and spatial characteristics of a sensor node and detects any irregularity in comparison with the node's previous data or the data of the neighbor nodes.

3.1.1 Detection of Irregular Patterns

We propose a new approach named *pattern variation discovery* to solve this problem. Our approach works in the following four steps: i) Selection of a reference frame. This frame consists of the directions along which we want to look for irregularities among multiple sensory attributes. An analyst can explicitly specify the reference frame. It is also possible to discover the reference frame that results in a lot of irregularities. ii) Definition of normal patterns. This definition can be models of multiple sensory attributes or constraints among multiple attributes. iii) Incremental maintenance of the normal patterns. Whenever a sensor gets a new round of readings, the normal patterns are adjusted incrementally. iv) Discovery of irregularity. Whenever a normal pattern is broken at some point along the reference frame, an irregularity appears. That is, the pattern variation happens.

For example, we want to discover the irregular distribution pattern among multiple sensory attributes along time. Then, for each time point, we can put the values of a group of sensory attributes at a series of sensor nodes into a matrix, which represents a distribution status. The problem then becomes to discover the irregular matrix among a set of matrices. An irregular matrix represents that, at the corresponding time point, the distribution pattern of all the sensory attributes on all the nodes are irregular. Because our approach involves a lot of comparisons between matrices, we propose to use the technique of Singular Value Decomposition (SVD) [4]. SVD is a powerful data reduction and approximation technique, which extracts the useful features of a matrix. Using SVD, we can get a vector of singular values out of a matrix. Consequently, matrix comparison becomes vector comparison, which is less computationally expensive and reduces communication cost. Additionally, integrating SVD with the sliding window mechanism, we can handle streaming sensory data.

3.1.2 Detection of Irregular Sensor Data

Detection of irregularities is tightly interrelated to modeling of sensor data. Therefore, we propose to detect irregular single-attribute sensor data with respect to time or space by building models.

For temporal irregularities in sensor data, we build a model of the sensory data as the readings of a node come in. When some reading substantially affects the coefficients of the model, it is identified as an irregularity. With resource constraints of sensor nodes, we may need to approximate the distribution of data instead of maintaining all historical data. In many applications, it suffices to consider the most recent N values in a sliding time window.

For spatial irregularities in sensor data, we build a statistical model of readings of neighboring nodes. If some readings of a node differ from what the model anticipates based on the readings of the neighboring nodes, an irregularity is detected. In order to reduce resource consumption, we may define the neighboring nodes to be those only a single hop away on the network. As a node moves geographically, the parameters of its model is incrementally adjusted. Distributed modeling is also possible.

Finally, there is a tradeoff between model accuracy and resource consumption. On one hand, modeling can reduce resource consumption because only model parameters are stored and transmitted instead of a large amount of sensor data. On the other hand, highly accurate models may result in a large size and frequent updates of data, which increases resource consumption.

3.2 Clustering of Sensor Data

We propose a new approach named *multi-dimensional clustering* of sensor data. This approach works as follows. First, cluster the sensor data along each sensor attribute separately. All resulted clusters form a set of clusters, which we call the Cluster Set. Second, construct a bipartite graph *G* with the Sensor Set (the set of sensor nodes) and the Cluster Set being the two vertex sets. If some sensory attribute value of a sensor *v* belongs to cluster *u*, there is an edge pointing from *v* to *u*. Last, find all of the maximal complete bipartite sub-graphs, i.e., the maximal bipartite cliques of *G*. These cliques identify "which sensor nodes have similar sensory readings on which attributes".

Figure 1 illustrates an example of multi-dimensional clustering. In the bipartite graph, the set of vertices on the top is clusters of sensor data by single-attribute of sensors, e.g., T1, T2, and T3 are clusters by the temperature attribute, L1, L2, L3 by light, and H1, H2, H3 by humidity correspondingly. The set of vertices in the bottom of the figure is the sensor nodes, with the dark ones being representative nodes of a clique. There are three cliques of the sensor nodes in the figure.

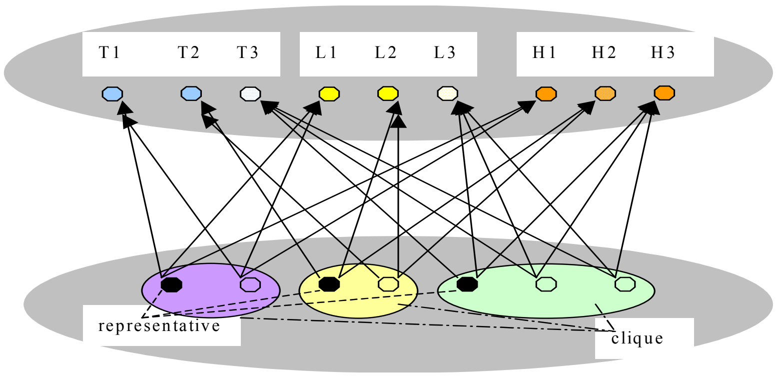

Fig. 1. Multi-dimension clustering

The cliques resulted from multi-dimensional clustering are useful not only for data analysis applications but also for network management and query optimization. Since all sensor nodes in one clique are similar, we can select a representative node for each clique to work on behalf of its clique in order to save power consumption with a reduced accuracy of data. This selection can be based on the residual energy of a node or the distance between the node and the base station. We can also select representative nodes based on a cost function of assigned tasks. In addition, the role of a representative node can be rotated among the nodes in one clique for load balancing.

3.3 Discovery of Sensory Attributes Correlations

Sensory attributes are rarely independent and correlations are common. For example, empirical evidence has shown that temperature and humidity are closely correlated in some natural environment. Therefore, efficiently identifying correlations among multiple sensory attributes is important for data analysis applications. For instance, we can estimate the changes of some attributes from the changes of the correlated attributes.

We treat readings of each sensory attribute as a data stream, i.e., a sequence of data items x_i at the sequence number i. As we are interested in correlation between data changes of various sensory attributes, e.g., correlation between the change of temperature values and that of light values, we replace each sensory data value x_i with its difference from its previous data item, $\Delta_i = x_i - x_{i-1}$. Thus, a time series of sensor data is represented as a sequence of Δ_i.

Let $S_1, \ldots, S_{m-1}, S_m$ be a collection of m sensor data streams, each for one attribute. One way of representing these data streams is to use a matrix A with time points and attributes being row and column indexes. We can then group data by correlated attributes or correlated time points [4][10] in this matrix. Recall that we propose to use the SVD technique (Singular Value Decomposition) for matrix reduction in pattern variation discovery. Here, this technique can also be used to find the best subspace that identifies the strongest linear correlations in the underlying data set [10]. Additionally, SVD tries to identify similarity patterns (rectangular regions) of related values in the *A* matrix, and the similarity of each row with the patterns [4]. It will naturally group similar "attribute-name" into attribute groups with similar behavior. In pattern variation discovery, we use SVD to speed up the comparison among multiple matrices. Here, in correlation discovery, we use SVD to consider the correlation among the rows within one matrix.

Alternatively, we can consider sensory attributes correlations as inter-transaction association rules [13]. For example, a rule or correlation says that "if node A's attribute x goes up at time point 1, B's attribute y will (with an 80% possibility) go up at time point 2 and C's attribute x will go down at time point 3". However, in the context of a large-scale mobile sensor network, the problem is more complex and challenging than traditional market basket analysis.

4 Related Work

There has been much work in the areas of sensor networks, data mining, and data streams, but little work has been done at the intersection of these areas.

Sensor networking protocols have attracted a tremendous amount of research effort [6]. Sensor databases and query processing techniques have been proposed for acquiring and managing sensor data [14][15]. However, existing sensor databases lack support for complex, online mining operations.

There is extensive literature regarding outlier (irregularity) detection [12][21]. However, none of these approaches is directly applicable to a sensor network environment. There is also initial work on modeling sensor data, including a distributed model based on kernel density estimators [16], and a distributed regression framework [9].

Although the clustering problem has been widely studied [5], we have not seen any previous work on multi-dimensional clustering. Existing sensor network clustering methods [3][8][22] mainly concern about the distance among nodes and the network topology, not sensory data. Recently, the clustering problem has also been studied in data streams [1][18].

There is some work on correlated data items [11] with respect to their accesses in order to improve data accessibility in sensor networks. In comparison, we focus on finding out correlations among sensory values. A related problem is identifying correlations among streams [10]. There has also been initial work on online analytical processing and mining for data streams [1][17][18]. However, they seldom consider the unique challenge in sensor networks.

5 Conclusions

We have identified the challenges for online mining in large-scale, mobile sensor network environments. The main concern is to satisfy the mining accuracy requirements while maintaining the resource consumption to a minimum. We identify three research problems to work on: (1) sensor data irregularities detection; (2) sensor data clustering; and (3) sensory attributes correlations discovery. We provide preliminary considerations towards solving these problems. We believe that the patterns discovered can not only enable the applications to gain insight into the sensor data, but also be used to tune the system performance. As future work, we will consider more about energy-awareness, adaptivity, and fault-tolerance of online mining for sensor networks in addition to a further study of our proposed approaches.

References

1. C.C.Aggarwal, J.Han, J.Wang, P.S.Yu. A Framework for Clustering Evolving Data Streams. VLDB2003.
2. I.F.Akyildiz, W.Su, Y.Sankarasubramaniam, and E.Cayirci. Wireless Sensor Networks: A Survey. Computer Networks, Vol.38, No.4, pp.393-422, 2002.
3. S.Bandyopadhyay and E.J.Coyle. An Energy Efficient Hierarchical Clustering Algorithm for Wireless Sensor Networks. IEEE INFOCOM 2003.
4. D.Barbara, W.Dumouchel, C.Faoutsos and P.Haas. The New Jersey Data Reduction Report. IEEE Data Engineering Bulletin, Vol.20, No.4, pp.3-45, 1997.
5. V.Estivill-Castro. Why So Many Clustering Algorithms----A Position Paper. SIGKDD Explorations. Vol.4, Iss.1, 2002.
6. D.Estrin, R.Govindan, J.Heidemann, and S.Kumar. Next Century Challenges: Scalable Coordination in Sensor Networks. MobiCOM 1999
7. M.J.Franklin. Challenges in Ubiquitous Data Management. Informatics 2001.
8. S.Ghiasi, A.Srivastava, X.Yang, and M.Sarrafzadeh. Optimal Energy Aware Clustering in Sensor Networks. Sensors 2002, 2, pp.258-269.
9. C.Guestrin, P.Bodik, R.Thibaux, M.Paskin and S.Madden. Distributed Regression: an Efficient Framework for Modeling Sensor Network Data. IPSN 2004.
10. S.Guha, D.Gunopulos, and N.Koudas. Correlating Synchronous and Asynchronous Data Streams. KDD 2003, pp. 529-534.
11. T.Hara, N.Murakami, and S.Nishio. Replica Allocation for Correlated Data Items in Ad Hoc Sensor Networks. SIGMOD Record, Vol.33, No.1, pp.38-43, March 2004.
12. E.M.Knorr and R.T.Ng. Algorithms for Mining Distance-based Outliers in Large Datasets. VLDB 1998.
13. H.Lu, L.Feng, J.Han. Beyond Intra-transaction Association Analysis: Mining Multi-dimensional Inter-transaction Association Rules. ACM Transaction on Information system, Vol.18, No.4, pp.423-454, 2000.
14. S.Madden, M.J.Franklin, J.M.Hellerstein, and W.Hong. TAG: A Tiny Aggregation Service for Ad-hoc Sensor Networks. In Symposium on OSDI, 2002.

15. S. Madden, M.J.Franklin, J.M.Hellerstein, and W.Hong. The Design of an Acquisitional Query Processor for Sensor Networks. SIGMOD 2003.
16. T.Palpanas, D.Papadopoulos, V.Kalogeraki, and D.Gunopulos. Distributed Deviation Detection in Sensor Networks. SIGMOD Record, Vol.32, No.4, Dec.2003
17. T.Palpanas, M.Vlachos, E.Keogh, D. Gunopulos, and W.Truppel. Online Amnesic Approximation of Streaming Time Series. ICDE 2004.
18. N.H.Park and W.S.Lee. Statistical Grid-based Clustering over Data Streams. SIGMOD Record, Vol.33, No.1, March 2004.
19. F.Perich, A.Joshi, T.Finin, and Y.Yesha. On Data Management in Pervasive Computing Environments. IEEE Transactions on Knowledge and Data Engineering. Vol. 16, No. 5, May 2004.
20. G.J.Pottie and W.J.Kaiser. Wireless Integrated Network Sensors. Communications of the ACM. Vol.43, No.5, pp.51-58, May 2000.
21. S.Sarawagi, R.Agrawal, and N.Megiddo. Discovery-driven Exploration of OLAP Data Cubes. EDBT 1998.
22. O.Younis and S.Fahmy. Distributed Clustering in Ad-hoc Sensor Networks: A Hybrid, Energy-efficient Approach. IEEE INFOCOM 2004.