Comparative Study of Association Rule Mining for Sensor Data

Manisha Rajpoot
Department of IT and MCA
Rungta College of Engineering and Technology, Bhilai (CG)-India

Lokesh Kumar Sharma
Department of IT and MCA
Rungta College of Engineering and Technology, Bhilai (CG)-India

ABSTRACT
Knowledge discovery from sensor data is an emerging research area due to many applications of crucial importance to our society. Wireless Sensor Networks produce large scale of data in the form of streams. Association Rule Mining in the sensor data provides useful information for different applications. In this study we analyze the framework of association rule mining for sensor data. Three data mining techniques PLT, SP-Tree and FP-Growth to mine the sensor data are considered in this study. These techniques are experimented with various support values and number of messages. The comparative performance analyses are reported in this paper.

General Terms: Knowledge Discovery from Sensor Data (Sensor-KDD), Wireless Sensor Network, Data Mining.

Keywords: Sensor Data Mining, Association Rule Mining, Pattern Discovery, SP-Tree, FP-Growth.

1. INTRODUCTION
Wide-area sensor infrastructures, remote sensors, and wireless sensor networks yield massive volumes of disparate, dynamic, and geographically distributed data. As sensors are becoming ubiquitous, a set of broad requirements is beginning to emerge across high-priority applications including disaster preparedness and management, adaptability to climate change, national or homeland security, and the management of critical infrastructures. Therefore data mining community interacted to develop the data mining techniques to extract hidden or known information from sensor data.

A sensor network is composed of a large number of sensor nodes, which are densely deployed either inside the phenomenon or very close to it. The position of sensor nodes need not be engineered or pre-determined. This allows random deployment in inaccessible terrains or disaster relief operations [1]. On the other hand, this also means that sensor network protocols and algorithms must possess self-organizing capabilities. Another unique feature of sensor networks is the cooperative effort of sensor nodes. Sensor nodes are fitted with an on-board processor. Instead of sending the raw data to the nodes responsible for the fusion, sensor nodes use their processing abilities to locally carry out simple computations and transmit only the required and partially processed data.

Time is a critical issue in sensor network and introduces the possibility of temporal relations between sensors. These relations are important in that they can help in predicting the sources of future events. Several techniques can be used to extract these temporal relations, among these techniques; data mining has recently received a great deal of attention. However, the stream nature of senor data along with the limited resources of wireless networks, bring new challenges to the data mining techniques that should be addressed Among these challenges are the type of the knowledge to be extracted from the networks and the way to extract the required data to mine the defined knowledge. Recently traditional data mining algorithms are extended for sensor data. In this study we consider three association rule mining techniques namely PLT, SP-Tree and FP-Growth and study the performances of these techniques.

The remainder of the paper is organized as follows: In Section 2 presents brief reviews of related work on sensor data mining. In Section 3 association rule mining techniques are reported. The experimental results and performance analysis are reported in Section 4 and our study is concluded in Section 5.

2. RELATED WORK
Loo et al. [8] have studied the problem of mining the associations that exist between sensor values in a stream of data reported from a wireless sensor network. They proposed a data model that stores the data and presents those in a way that makes it possible to adapt the lossy counting algorithm [9] that makes an online one-pass analysis of the data. In this data model, sensors are assumed to take values from a finite discrete number of values, whereas a quantization method is applied for the continuous values. The time is divided into equal-sized intervals, and a snapshot from the sensor reading is taken whenever there is an update on a sensor reading. These snapshots formulate the contexts of the database. Although taking snapshots at state changes will reduce the redundancy in the data, these snapshots occur randomly; thus, each context is associated with a weight value that indicates for how many intervals this reading is valid (that is, for how long these readings will kept unchanged). The support of the pattern is defined by the total length of non overlapping intervals in which the pattern is valid.

Mining spatial temporal event patterns is another attempt to link the problem of mining sensor data to the association rules' mining problem that was proposed by Roemer [9]. Roemer's approach takes into consideration the distributed nature of wireless sensor networks and proposes an in-network data mining technique to discover frequent patterns of events with certain spatial and temporal properties. In this approach, each sensor should be aware of the events that are within a certain distance from itself (this distance may be a Euclidean distance or a number of hops). The sensor then collects these events and applies a mining algorithm to discover the pattern that satisfies the given parameters. The mining parameters include a minimum support S, a minimum confidence C, a maximum scope, and a maximum history. Each node in the network collects the events from its neighbors within the maximum scope and keeps a history of their events for duration of the

maximum history. After that, each node applies a mining algorithm to discover the frequent patterns (those that have frequency exceeding the given minimum support).

Halatchev and Gruenwald [11] proposed an association rule mining framework to stand the missed readings that result from the loss and corruption of messages while they are routed from sensor nodes to the processing points. Sensor readings are streaming in nature; hence, applying an association mining algorithm such as Apriori [2] directly to the stream of data is not possible in the first place. This situation led the authors to propose the Data Stream Association Rule Mining (DSARM) framework that adapts the "Apriori" algorithm to make it applicable to the data stream received from sensor nodes. There are several modifications that have been made for the Apriori scheme to be adapted for sensor streams. First, rules are generated between pairs of sensors instead of generating all of the possible rules. Second, the association between pairs of sensors is evaluated with respect to a particular state of the sensors, and this modification will lead to rules of the form s1 $\Rightarrow$s2/st which means that s1 determines s2 with respect to state st. Finally, the sliding window technique is implemented to generate the association between sensors within the given window size. To the best of our knowledge, few studies have proposed addressing the problem of extracting data from wireless sensor networks for mining patterns regarding the sensor nodes themselves. All the attempts have focused on extracting patterns regarding the phenomenon monitored by the sensor nodes, in which the mining techniques are applied to the sensed data received from the sensor nodes and accumulated at a central database. In our work, we will propose a solution to extract the behavioral data required for mining patterns regarding the behavior of the sensor nodes in the network (that is, the data used in the mining process is metadata, describing the nodes activities, and it differs from the sensed data). A primary assumption of the proposed data extraction mechanism is to have a flash memory device attached to each sensor to store the metadata about the sensor's behavior that will be used during the extraction process. Several researchers have studied the cost of attaching a storage devise to each sensor. In [10], Mathur et al. have showed that current flash memories offer a low-cost high-capacity energy-efficient storage solution, especially when compared with the transmission of the data.

3. FRAMEWORK OF ASSOCIATION RULE MINING IN SENSOR DATA

Notations for Sensor association rules can be derived based upon the definition of association rules proposed in the domain of transactional databases. It can be represented as follow:

Let S= (s_1, s_2. . . s_n) be a set of sensors in a particular sensor network. Let assume that the time is divided into equal-sized slots (t_1, t_2 . . . t_n) such that $t_{i+1} - t_i = \lambda$ for all $1 < i < n$, where λ is the size of each time slot, and $T = t_n - t_1$ represents the historical period of the behavioral data defined during the data extraction process. Also, P = (s_1, s_2 . . . s_k) $\subseteq$ S as a pattern of sensors is referred.

Definition 3.1. Let Rank(s) be the function that maps each sensor node s (where s ϵ S) to a unique integer number so that the lexicographic order is preserved.

Definition 3.2. Let pos(s) be the function that maps each sensor node s, where s 2 S, in the lexicographic tree to an integer number that represents its position among its siblings relative to the parent node (that is, the lexicographical distance between the node's identifier and its parent identifier).

Definition 3.3. Given a set S of sensors, Path(S) is defined to be the list of all possible paths from the root node to any other sensor node in S's PLT.

Let us consider the example portrayed in Fig. 4. If we omit the value of the root node, then the elements of Path(S) can be listed as follows: Path(S)={ [1], [2], [3], [4], [1,1],[1, 1, 1],[1, 1, 2], [1, 1, 1, 1],[1, 2],[1, 2, 1], [1, 3], [2, 1],[2, 1, 1], [2, 2], [3, 1]}.

Definition 3.4. V(P), the position vector of the pattern P, is defined by the vector [pos(s1), pos(s2), . . . , pos(sk)], where P is a subset of the set S, and {s1, s2, . . . , sk} is the path in the lexicographic tree that maps the elements in P.

Definition 3.5. Given a database DS of epochs. A PLT structure of DS is defined to be a set of partitions, each represented by a tabular structure of the epochs' position vectors in such a way that all position vectors sharing the same sensor as the last element in the vector will appear in the same partition.

4. ASSOCIATION RULE MINING IN SENSOR DATA

Positional Lexicographic Tree (PLT), is able to partition and compressed the data and provides an easy access mechanism for manipulating the data PLT use temporal relations between sensors, these relations are able to generate the set of correlated sensors which can be used later to estimate the value of another sensor, to predict the future sources of events, or to identify faulty nodes. The distributed extraction tries to maximize the network lifetime through optimizing number of exchanged messages sensor pattern tree (SP-tree) for mining association rules for Wireless Sensor Networks data. The important features of SP-tree are (i) it can be constructed with one scan over the sensor epochs, which is highly crucial while the streams of sensor data flow; (ii) it is a frequency-descending tree structure, which enables an efficient FP-growth-based mining technique.

This new data structure is denoted by FP-tree (Frequent- Pattern tree) and is created as follows. The reason to store transactions in the FP-tree in support descending order is that in this way, it is hoped that the FP-tree representation of the database is kept as small as possible since the more frequently occurring items are arranged closer to the root of the FP-tree and thus are more likely to be shared.

The PLT structure to outperform the mining process using the FP-tree. These issues include the following:

1. The partitioning mechanism used in PLT makes it easy to locate the conditional vectors of a particular pattern instead of following the nodes' link as in the FP-tree.
2. PLT partitions are independent. We do not need the entire structure to be in the main memory, as opposed to the FP-tree that requires the whole structure to be in the main memory.
3. The comparison values of the position vectors that are used to locate and insert vectors accelerate the process of accessing the PLT structure, whereas the FP-tree does not include such values.

4. PLT requires smaller variable sizes for storing the positional values, since it uses the lexicographic distance to represent sensors, as opposed to the FP-tree that uses the actual sensors' identifiers.

5. EXPERIMENT AND RESULT ANALYSIS

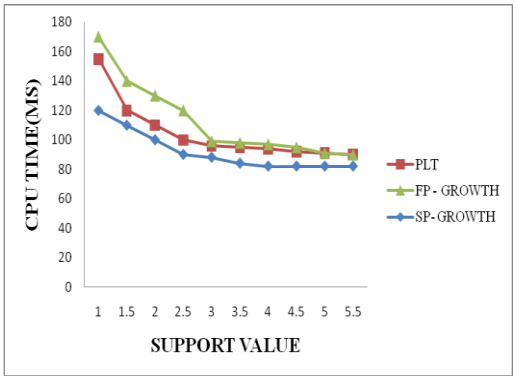

Fig. 1 Support values versus CPU Time for S100 using PLT, FP- Growth and SP-Tree.

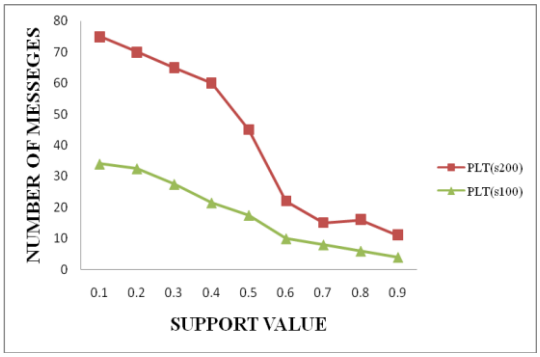

Fig. 2 Support value versus number of messages for s200 and s100 using PLT.

In this experiment we have analyzed for PLT values s100 and s200. From fig. 1 it is shown that at PLT s200, the number of messages is decreased considerably with increase in support values. At PLT s100, the number of messages slightly decreased with increase in support values.

The comparison result for PLT and FP-growth from fig. 2 is that in case of PLT, the CPU time is decreased considerably with increase in support values. For SP-Tree the CPU time initially is very high and then decreases dramatically. There is no change in CPU time in both the cases after support value .4.

In case of PLT, the CPU time is decreased with increase in support values. For SP-Tree the CPU time decreases slightly then there is no change after support value 4. But PLT consumes more CPU time than SP-Tree

6. CONCLUSION

In this work we have analyzed association rule mining PLT, SP-Tree and FP-growth for sensor data. We analyzed PLT for s100 and s200 with support value versus number of messages. At PLT s200, the number of messages is decreased considerably with increase in support values. The CPU time consumed by PLT is more that of SP-Tree and less than FP-growth. Over all FP-growth consumes initially high CPU time in low support values and SP-Tree consumes considerably less CPU time than PLT.

7. REFERENCES

[1] A. Boukerche and S. Samarah, "A Performance Evaluation of Distributed Framework for Mining Wireless Sensor Networks", ANSS'07, pp. 239-246, 2007.

[2] A. Boukerche and S. Samarah, "A Novel Algorithm for Mining Association Rules in Wireless Ad Hoc Sensor Networks", IEEE Transactions On Parallel And Distributed Systems, Vol. 19, (7), pp. 865-877, July 2008.

[3] S. K. Tanbeer, C. F. Ahmed, B. S. Jeon, Y. K. Lee, "Efficient Mining of Association Rules from Wireless Sensor Networks", Proc. of ICACT, Feb. 15-18, pp. 719-724, 2009.

[4] A. Boukerche and S. Samarah, "An Efficient Data Extraction Mechanism for Mining Association Rules from Wireless Sensor Networks", Proc. of ICC 2007.

[5] B. Goethals, M. J. Zaki, "Frequent Item set Mining Implementations," FIMI'04 Brighton, UK, 2004.

[6] D. Culler, D. Estrin and M. B. Srivastava, "Overview of Sensor Networks," Computer, vol. 37 (8), pp. 41-49, August 2004.

[7] S. S. Iyengar and R. R. Brooks, "Distributed Sensor Networks", CRC press, 2004.

[8] K. K. Loo, I. Tong, B. Kao and D. Chenung, "Online Algorithms for Mining Inter-Stream Associations from Large Sensor Networks", Proc. of PAKDD '05 Springer LNCS 3518, pp. 291-302, May 2005.

[9] K. Roemer, "Distributed Mining of Spatio-Temporal Event Patterns in Sensor Networks," Proc. of EAWMS '06, June 2006.

[10] G. Mathur, P. Desnoyers, D. Ganesan, and P. Shenoy, "Ultra Low Power Data Storage for Sensor Networks," Proc. Fifth IEEE/ACM Conf. Information Processing in Sensor Networks (IPSN '06), Apr. 2006.

[11] M. Halatchev and L. Gruenwald, "Estimating Missing Values in Related Sensor Data Streams," Proc. 11th Int'l Conf. Management of Data (COMAD '05), pp. 83-94, Jan. 2005.

[12] A. A. Salah, E. Pauwels, R. Tavenard and T.Gevers, "T-Patterns Revisited: Mining for Temporal Patterns in Sensor Data", Sensors 2010, vol (10), pp. 7496-7513; doi: 10.3390/s100807496.

[13] V. Tseng and K. Lin "Energy efficient strategies for object tracking in sensor networks: A data mining approach", Journal of System Software, vol (80), pp. 1678–1698, 2007.

[14] V. Tseng and E. Lu, "Energy-efficient real-time object tracking in multi-level sensor networks by mining and predicting movement patterns", Journal of System Software, vol (82), pp. 697–706, 2009.