A Reasonable Exploration of Data Mining Techniques in Wireless Sensor Network

Shubhie Agarwal
M.Tech Scholar
Department of CSE
KIET Group of Institutions, Ghaziabad

Seema Maitrey
Assistant Professor
Department of CSE
KIET Group of Institutions, Ghaziabad

Poonam Rana
Assistant Professor
Department of CSE
KIET Group of Institutions, Ghaziabad

Pankaj Singh Yadav
Assistant Professor
Department of CSE
Raj Kumar Goel Instt. of Technology, Ghaziabad

ABSTRACT

Data mining in WSN is the process of extracting model and pattern that are application oriented with possible accuracy from continuous, rapid flow of data. The whole huge amount of data cannot be stored and processed immediately. That is why the mining algorithm should be fast enough to process high speed arriving data. There are many conventional data mining techniques, but they are not able to handle dynamic amount of data. It is difficult to handle WSN data. There are several challenges that it has to face in WSN. The main aim of wireless sensor networks is to transmit data in such a manner that increased lifetime of the network and energy efficient routing can be done with significant accuracy. Data mining is the process of discovering interesting patterns (or knowledge) from large amounts of data. Knowledge discovery process attains several steps and can be interactive, iterative and user-driven. Data mining techniques of wireless sensor network are different from traditional techniques. Data mining techniques can be frequent pattern mining, sequential pattern mining, clustering and classification. All these techniques can use centralized or distributed approach, even after that the focus is decided that either you can focus on application or performance of wireless sensor network. Data mining techniques that work on sensor network-based application are still facing shortcomings in existing techniques. By seeing these shortcomings and special characteristics of WSNs, there is a need for data mining technique designed for WSNs. In this paper, we are finding the difference between traditional and sensor data processing. Also comparing the different data mining techniques used in wireless sensor network where all these methods have their own processing architecture and method of sensor distribution depending upon the attributes.

Keywords

Centralized Mining, Distributed Mining, Sequential Mining, Pattern Mining, WSN

1. INTRODUCTION

In Wireless sensor network a large number of sensor nodes are deployed in medium (e.g. Air, Water) by wireless communication technology, these nodes are responsible to send measurement data to sink node for further processing. Sensors coordinate to perform sensing of environment over large physical area and enable reliable monitoring and controlling in various applications. Wireless sensor network provide bridges between the virtual world of information technology and real physical world. Data mining in WSN is the process of extracting model and pattern that are application oriented with possible accuracy from continuous, rapid flow of data. The Whole huge amount of data cannot be stored and processed immediately. That is why the mining algorithm should be fast enough to process high speed arriving data. There are many conventional data mining techniques, but they are not able to handle dynamic amount of data. For conventional data mining techniques it is difficult to handle WSN data.

There are several challenges that it has to face in WSN. The sensor nodes are deployed at different such geographical location, where resources cannot be upgraded and changed frequently. Resources like battery, memory, bandwidth, processing power are limited for wireless sensor network due to size and capabilities. Due to variant nature of WSN and dependency on application specific generation of data wireless sensor networks generate high speed data. In several domains data arrives so fast than we are able to mine. Spatiotemporal embedding of sensor nodes may cause many classical data processing techniques perform very poor.

The mining data generated becomes old very fast, due to continuous rapid arrival of data and affect the result generated. Most of data mining technique basically collects the data at some central location and after some time process that data and generates the results accordingly. This process is called offline mining that is quite inadequate to meet the requirement of handling distributed stream data generated by wireless sensor networks. The raw data generated by wireless sensor networks cannot be transformed as per mining requirement due to limited bandwidth. Knowledge base in some structural format transforms the data. A time variant motion is possible in wireless sensor networks. It increases the dynamicity and complexity of algorithm. Being resources constrained wireless sensor networks cannot posses high computation power. New algorithms have been created, and some of the data mining algorithms have been modified to handle the data generated from sensor networks. Number of knowledge discovery methodologies, techniques, and algorithms have been proposed during the last ten years.

The rest of the paper is organized as follows. After the introduction in Section 1, how traditional data mining process is different with data mining process in WSNs and challenges of data mining for WSNs data are discussed in Section 2. It mention the taxonomy of categorizing the existing data mining techniques for WSNs and the comparison of data mining techniques for WSNs in Section 3. Finally, the paper ends with the conclusion and the future research directions in Section 4.

2. RELATED WORK

Traditional data mining is centralized, computationally expensive, and focused on disk-resident transactional data. In comparison with traditional data-sets, the WSNs data flows continuously in systems with varying update rates. Due to huge amount and high storage cost, it is impossible to store the entire WSNs data or to scan through it multiple times. These characteristics of sensor data and the special design issues of sensor networks make traditional data mining techniques challenging. Hence, it is crucial to develop data mining technique that can analyze and process WSNs data in multidimensional, multilevel, single-pass, and online manner. We can observe from Table 1 the differences that exist between traditional data and WSN data.

Table 1:- Difference between traditional and Sensor data processing

	Traditional Data	**WSN Data**
Data Type	Static	Dynamic
Memory	Unlimited	Restricted
Computation Capacity	High	Low
Architecture	Centralized	Distributed
Processing Time	Unlimited	Restricted
Power	No Constraint	Limited
Data Flow	Stationary	Continuous
Data Length	Bounded	Unbounded
Responsiveness	Non-Real Time	Real Time

Traditional growth based mining technique like Apriori[1] and frequent pattern got a chance to perform in wireless sensor network by converting them into frequent and sequential version of pattern mining and provides the association among large WSN data. Clustering adopted the basic functions of K-mean, hierarchal and correlation based methods. Clustering focuses on distance among the data points. Classification adopted the basic functions of decision tree, rule based, nearest neighbor and support vector methods. Classification focuses on the model used. As we know that sensor nodes have limited resources (power, computation, memory, bandwidth) distributed processing should be done in one pass for proper mining locally and then result should be aggregated. We can limit the messages of communication and energy of sensor nodes by transferring data to central server. Centralized processing collects the data from entire network and analyze at central server. Central server have unlimited resources so we do not focus on accurate algorithm in this approach. Either the mining technique can focus on performance issues or application issues. It means either we can focus on resource constrained that are limited to improvise the performance of WSN or we can focus on data precision, accuracy, scalability, robustness of application. Following are several techniques that are used for mining in WSN.

2.1 Frequent Pattern Mining

This technique is used to find the set of variables that co-occur frequently in data set. It tries to find out relation between the variables. The motive is to make CPU and I/O intensive methods, dynamic in nature like WSN. The basic frequent mining technique is association rule based Apriori[]1] technique. Basically iterations are performed over database to find the frequent patterns and possible patterns (candidate pattern) for next iteration. If we suppose the frequent pattern is of length K then possible frequent pattern (candidate pattern) will of length K-1. Now we can generate the association rules by computing Support and Confidence. But it becomes difficult in WSN because of no centralized processing, researches have suggested FP-growth[2] method in which we reduce the database in two parts and eliminate candidate generation. By checking both the databases one by one and comparing them we can generate FP-tree.

Centralized Approach like DSARM (data stream association rule mining)[3] suggests identifying sensors that report the same data for a number of times in sliding window then estimate the missing data from a sensor by using the data reported by its related sensors. A new technique of data estimation called CARM[4,5] (closed item sets based association rule mining) is suggested which can derive the most recent association rules using CFI-stream[6] and maintains an in memory data structure direct update (DIU). We can use online one pass method in which frequent value is generated by transforming the stream data into Interval List (IL)[7,8]. The time is divided into equal-size interval. Rule-learning model finds strong rules from sensor reading and can trigger the sensor as per requirement. Transactions are batched in this approach and sensor streams are collected by some sensor. This sensor is responsible to generate association rule. A newer approach is tree based data structure SP-tree[9] can be used to generate the association rules by obtaining the frequency of event detection sensor data. When the frequency based tree is generated FP-growth[2] mining technique is used to find frequent pattern. Distributed Approach like in-network[10] data mining uses spatial and temporal properties to discover frequent pattern. We can generate several parameter variables and its upper bound can be used as a event for a sensor. Each node collects events of its neighbor in scope. Now mining algorithm is used to find the frequent pattern and converted as association rule. A distributed data extraction methodology can be used to aggregate the data on sensor done. In this method the sink node is responsible to broadcast the parameters to remaining nodes. Each node having its own buffer maintains messages after a time interval each node checks the messages in buffer and if it is satisfying the support value then only messages will be transferred to sink node. Distributed positional lexicographical tree structure (PLT)[11] uses event-detecting sensor and recursive method is applied generate association rule. PLT structure based method is better than FP-growth[2] method in term of CPU and memory usages.

2.2 Sequential Pattern Mining

It is more complex extension of frequent pattern mining. Frequent patterns are discovered in sequential databases that stores records that is sequence of ordered events without focusing on time slot. Centralized approach provides a multidimensional relational sequence mining [12,13] method to find hidden frequent temporal correlation between sensor nodes. Abstraction is done to segment and label the real valued time series, then interval based operator are added to enrich the knowledge, it is more capable than other method to generate human readable pattern. MavHome smart home architecture [14] is episode discovery (ED) based method for mining sequential pattern from time ordered transactions.

Decisions are based on predicted activities. Values that can be predicted include the usage pattern of devices. An anomaly detection method[15] is used to improvise railway maintenance. They extracted the abnormal behavioral data that was based on environmental and structural changes in data. A new algorithm MSAPs (Mining sequential alarm pattern)[16] from GSM system. Here sequential events are identified by defining time interval between adjacent events. The knowledge is extracted to find relevance between two events. Distributed approach like MLOT (multilevel object tracking)[17] uses mining technique on movement log inn sensor network. A multilevel hierarchical structure is adopted and MPG (movement patter generation) algorithm is used to obtain movement pattern. An Object tracking method TMP-mine (Temporal movement patterns)[18] is used to predict the location of next object for saving energy. Using these rules we can do tracking of object in energy efficient way. We can use prediction based tracking method by sequential pattern mining in which inherited patterns of objects movements and sequential patterns are used to predict to which next location moving object will head towards.

3. CLUSTER BASED MINING

In clustering data is categorized in such a subset manner that each subset forms a cluster. This technique is most useful for those application which require scalability to a large no. of sensor nodes. We can do aggregation of data to summarize the overall transmitted data.

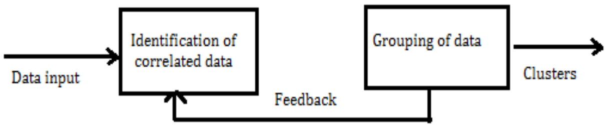

Fig.1: Clustering in wireless sensor network

Researchers have suggested two types of clustering Node and Data Clustering. Node clustering may vary depending on properties like deployment of node, network architecture, operation model. Some of node clustering techniques are dependent of data clustering. Basically in data clustering, they group some sensor nodes and elect one head out of them. This head is used to find data correlation among the nodes. We can use K-mean, data-correlation or hierarchical method for clustering and correlation identification. Centralized approach like EEDC (energy efficient data collection)[19] uses on-demand clustering in which sensor possessing same sensed values are used as one cluster. A sink node is responsible to compare the data in different node with user defined measures. Graph based technique generates the dissimilarity measures as an edge and it is compared with the threshold decided. The cluster is treated as a clique and minimum no. of cliques are to cover all the vertexes. Now the energy-optimized clusters are generated that are even balanced.

Distributed Approach like H-Cluster[20] can take sensor data as input and generate cluster as output . Hilbert-Map method can be used to generate n-dimensional data space. Basically there are two step in first step, connected features are merged with local cluster features in dimensional space then in second step local clusters are connected with global clusters. This method is efficient in managing data losses, power optimization and quality of cluster. Another proposed method DCC (Data-correlation based clustering)[21] is useful in reducing the size of data by finding the probability to become a cluster head. Spatial suppression is done at cluster head and comparison between the representative and sensed value is done and correlation method is used to suppress the size of data. Another technique CAG (Cluster aggregation)[22] suggests generating cluster of nodes sensing similar values within the threshold. Spatial and temporal correlation is used for grouping the data. CAG transmit one reading per group, when the threshold gets changed the membership of a node can be changed as per associated neighbor. Query and attributes of data based method of clustering [23] is motivated for efficient data dissemination in WSN. Here clustering is performed by a map of hierarchy of data attributes to network topology. Base station asks the nodes to become cluster head. Nodes wait for random time and when ready to nominate itself as cluster head, broadcast the message to all the nodes. A node joins the cluster head that is reachable in least no. of hops.

4. CLASSIFICATION

Classification method is useful in assigning the new object to a class of predefined object. Classification can use decision-tree, nearest neighbor or support vector techniques based on classification model used. Decision tree is classifies in the form of tree, nearest neighbor classifies in data set from and support vector partition data belonging to different class format.

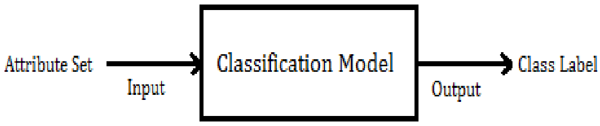

Fig. 2: Classification Map

Centralized approach like DT (Decision tree) based technique[24] uses frequent pattern as a episode and a matrix is generated then pattern classifier generate a labeled training data set using divide and conquer approach. By recursively selecting the attributes set we can partition the training data set in to subset until leaf of tree. Another method NNTC (Nearest neighbor trajectory classification)[25] initially stores every training set by its label then for prediction the distance between these set is computed. The K closet training sets are stored and most frequent labels are checked for prediction. Another method LW (Light weight classification)[26] a one pass algorithm is used for mining on- board mining of data stream in sensor network. Here nearest instances stored in memory are searched when a new element is arrived. Acceptable similarity measure is used to identify the elements and weight increments are done is class label matches otherwise decremented. It the weight becomes zero then element is released from memory.

Distributed approach like FVLD (feature vector of low dimension)[27] uses cluster based approach for detection and classification. Here monitoring area is divided into clusters and cluster head is selected. All nodes send the feature vectors to cluster head and cluster head does the classification using decision fusion and maximum likelihood. Another method SVM (support vector machine)[28] based is incremental technique for learning classification rules. Support vector are less in comparison to all sample values. Due to communication reduction the cluster head becomes energy efficient.

Inferences Drawn out of Literature Search

- Sensor data is of large volume, real time and continuous. The type of data dependent technique is selected on the basis of what kind of attributes and correlation it posses.
- Sometimes the unique characteristic of WSN data is used to determine the method of mining. It can use real time or one pass/scan algorithm to generate expected result.

- Mining techniques are influenced by node properties like connectivity, mobility, role of node, task of node.
- Method of mining is also dependent on source that can be synthetic or real. Synthetic sources are imulation tools and that are validate with real values.
- Most of the method takes homogeneous data. Aggregation of homogeneous and Heterogeneous data can improvise the accuracy of mining.
- Attribute dependency is ignored by most of techniques and focus is towards spatial or temporal correlation. That is why the complexity of mining is increased.
- The mobility parameter is ignored by most of techniques, that can provide random topological changes and variable data can be generated.
- Majority of techniques uses centralized approach in which data is transmitted to a node called sink. The method causes much communication overhead and responses are delayed.

5. COMPARATIVE ANALYSIS

Basically there are four types of data mining methods sequential pattern, frequent pattern, clustering based and classification based method, all these methods have its own processing architecture and method of sensor distribution depending upon the attributes. All these methods cannot be evaluated by some practical modeling method evaluated by simulation tools; they can be checked for their performances on the basis of some analytical methods. After doing the comparative analysis we came to know that most of the techniques are based on sensor attribute that is it can be homogeneous on heterogeneous distribution of sensor nodes and processing can be done at a central location or every node can do computation and processing in distributed manner. Central processing and distributed processing both possess it own drawback and features. The comparison among these methods are displayed in the following table:

Table 2: Comparison among different methods

Name	Mining Method	Processing Architecture	Sensor Attributes	Evaluation Technique
DSARM	Apriori like	Centralized	Homogeneous	Simulation
Distributed data aggregation	Apriori like	Distributed	Homogeneous	Analytical
Lightweight rule learning	Apriori like	Centralized	Heterogeneous	Simulation
CARM	FP-growth	Centralized	Heterogeneous	Simulation
Episode Discovery	Generalized Sequential pattern	Centralized	Homogeneous	Analytical
MPG	Apriori like	Distributed	Homogeneous	Analytical
CPD	PSP	Centralized	Heterogeneous	Analytical
PTSP	Sequential pattern generation	Distributed	Homogeneous	Simulation
CAG	Correlation based Clustering	Distributed	Homogeneous	Simulation
EEDC	Correlation based Clustering	Centralized	Heterogeneous	Analytical
DHCS	Hierarchal Clustering	Distributed	Heterogeneous	Analytical
Prediction framework	Decision Tree	Distributed	Homogeneous	Simulation
NNTC	Nearest Neighbor	Centralized	Homogeneous	Simulation
LW Class	KNN Classifier	Centralized	Heterogeneous	Analytical
FVLD	Max. Likelihood classifier	Distributed	Heterogeneous	Simulation

6. CONCLUSION

Data mining techniques of wireless sensor network are different from traditional techniques. Data mining techniques can be frequent pattern mining, sequential pattern mining, clustering and classification. All these techniques can use centralized or distributed approach, even after that the focus is decided that either you can focus on application or performance of wireless sensor network. It is observed from the analysis of existing data mining work on sensor network-based application there are still shortcomings in existing techniques. By seeing these shortcomings and special characteristics of WSNs, there is a need for data mining technique designed for WSNs. After comparative analysis we can conclude that we are either using centralized or distributed approach and both have their own drawbacks due to communication overhead in centralized approach the responses and performance get reduced as the time interval increases. We can use distributed approach in integration with centralized approach to improvise the throughput and response time of data mining. A hybrid approach of data mining can perform better than the previously available techniques of mining.

7. REFERENCES

[1] R. Agrawal, T. Imieli´nski, and A. Swami, “Mining association rules between sets of items in large databases,” in Proceeding of SIGMOD, pp. 207–216.

[2] J. Han, J. Pei, Y. Yin, and R. Mao, “Mining frequent patterns without candidate generation: a frequent-pattern tree approach,” Data Mining and Knowledge Discovery, vol. 8, no. 1. pp. 53–87, 2004.

[3] M. Halatchev and L. Gruenwald, “Estimating missing values in related sensor data streams,” in Proceedings of the 11th International Conference on Management of Data (COMAD’05), 2005.

[4] J N. Jiang, “Discovering association rules in data streams based on closed pattern mining,” in Proceedings of the SIGMOD Workshop on Innovative Database Research, 2007.

[5] N. Jiang and L. Gruenwald, “Estimating missing data in data streams,” Advances in Databases: Concepts, Systems and Applications,pp. 981–987, 2007.

[6] N. Jiang and L. Gruenwald, “CFI-stream: mining closed frequent item sets in data streams,” in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’06), pp. 592–597, August 2006.

[7] K. Loo, I. Tong, and B. Kao, “Online algorithms for mining inter-stream associations from large sensor networks,” in Advances in Knowledge Discovery and Data Mining, pp. 291–302, 2005.

[8] G. S. Manku and R. Motwani, “Approximate frequency counts over data streams,” in Proceedings of the 28th International Conference on Very Large Data Bases, pp. 346–357, 2002.

[9] S. K. Tanbeer, C. F. Ahmed, B.-S. Jeong, and Y.-K. Lee, “Efficient mining of association rules from Wireless sensor networks,” in Proceedings of the 11th International Conference on Advanced Communication Technology (ICACT ’09), pp. 719–724, February 2009.

[10] K. Romer,“Distributed mining of spatiotemporal event patterns in sensor networks,” in Proceedings of the 1st Euro-American Workshop on Middleware for Sensor Networks (EAWMS ’06), 2006.

[11] 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, no. 7, pp. 865–877, 2008.

[12] F. Esposito, T. M. A. Basile, N. Di Mauro, and S. Ferilli, “A relational approach to sensor network data mining,” Information Retrieval and Mining in Distributed Environments, pp. 163–181, 2010.

[13] F. Esposito, N. Di Mauro, T. M. A. Basile, and S. Ferilli,“Multi-dimensional relational sequence mining,” Fundamenta Informaticae, vol. 89, no. 1, pp. 23–43, 2008.

[14] D. J. Cook, M. Youngblood, E. O. Heierman III et al., “MavHome: an agent-based smart home,” in Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications (PerCom ’03), pp. 521–524, March 2003.

[15] J. Rabatel, S. Bringay, and P. Poncelet, “SO MAD: sensor mining for anomaly detection in railway data,” in Advances in Data Mining. Applications andTheoretical Aspects, pp. 191–205, 2009.

[16] P. H. Wu, W. C. Peng, and M. S. Chen, “Mining sequential alarm patterns in a telecommunication database,” in Databases in Telecommunications II, pp. 37–51, 2001.

[17] V. S. Tseng and E. H.-C. Lu, “Energy-efficient real-time object tracking in multi-level sensor networks by mining and predicting movement patterns,” Journal of Systems and Software, vol.82, no. 4, pp. 697–706, 2009.

[18] V. S. Tseng and K.W. Lin, “Energy efficient strategies for object tracking in sensor networks: a data mining approach,” Journal of Systems and Software, vol. 80, no. 10, pp. 1678–1698, 2007.

[19] C. Liu, K.Wu, and J. Pei, “A dynamic clustering and scheduling approach to energy saving in data collection from wireless sensor networks,” in Proceedings of the 2nd Annual IEEE Communications Society Conference on Sensor and AdHoc Communications and Networks (SECON ’05), pp. 374–385, September 2005.

[20] L. Guo, C. Ai, X. Wang, Z. Cai, and Y. Li, “Real time clustering of sensory data in wireless sensor networks,” in Proceedings of the IEEE 28th International Performance Computing and Communications Conference (IPCCC ’09), pp. 33–40,December 2009.

[21] M. H. Yeo, M. S. Lee, S. J. Lee, and J. S. Yoo, “Data correlation based clustering in sensor networks,” in Proceedings of the International Symposium on Computer Science and its Applications (CSA ’08), pp. 332–337, October 2008.

[22] S. Yoon and C. Shahabi, “The Clustered Aggregation (CAG) technique leveraging spatial and temporal correlations in wireless sensor networks,” ACM Transactions on Sensor Networks, vol. 3, no. 1, Article ID1210672, 2007.

[23] K.Wang, S. A. Ayyash, T. D. C. Little, and P. Basu, “Attribute based clustering for information dissemination in wireless sensor networks,” in Proceedings of the 2nd Annual IEEE Communications Society Conference on Sensor and AdHoc Communications and Networks (SECON ’05), pp. 498–509, Santa Clara, Calif, USA, September 2005.

[24] B. Chikhaoui, S. Wang, and H. Pigot, “A new algorithm based on sequential pattern mining for person identification in ubiquitous environments,” in Proceedings of the 4th International Workshop on Knowledge Discovery form Sensor Data (ACM Sensor KDD ’10), pp. 20–28,Washington, DC, USA, 2010.

[25] K. Sharma, M. Rajpoot, and L. K. Sharma, “Nearest neighbour classification for wireless sensor network data,” International Journal of Computer Trends and Technology, no. 2, 2011.