Mining Data Generated by Sensor Networks: A Survey

Azhar Mahmood, Ke Shi and Shaheen Khatoon
School of Computer and Applied Technology,
Huazhong University of Science and Technology (HUST), Wuhan, China

Abstract: Sensor Networks (SNs) produces huge amount of data which offer promising prospect for the application of data analysis techniques to extract useful information for end users. Knowledge discovery from sensor data is an emerging research area due to their wide range of potential application of crucial importance to our society. Since sensors nodes are resource constraints in term of memory, communication bandwidth and battery power, therefore data mining techniques required to meet these constraints while applying to SN data. This study presents the classification and evaluation of Data Mining (DM) methods applied on sensor networks data, leading to more efficient techniques for resources optimization. Plenty of research work done on sensor networks but there is none of such study exist that focuses exclusively on data mining methods. The main contribution in this study is classifying and comparing the published data mining approaches in sensor networks and highlights the challenges as well as future trends in applying data mining methods on SNs data.

Keywords: Data mining, sensor network, stream data, association rule mining, clustering, classification, sequential patterns, frequent pattern

INTRODUCTION

In information world data mining is gaining a rapid popularity because of its aim towards extraction previously unknown and potentially valuable information (such as knowledge rules, constraints, regularities) from large amounts of data where the data can be stored in databases, data warehouses, OLAP (On Line Analytical Process) or other repository information (Halkidi, 2000, Han and Kamber, 2001). It involves integration of techniques from multidiscipline such as database technology, statistics, machine learning, neural networks, information retrieval etc. (Fayyad et al., 1996; Hand et al., 2001). Data mining is used to solve wide range of problems in various applications such as product analysis, understanding consumer research marketing, e-commerce, demand and supply analysis, direct marketing, health industry, e-commerce, stocks and real estates, Customer Relationship Management (CRM), telecommunication industry and financial sector investment. Besides these general applications, nowadays data mining is also applied on domain specific applications such as biomedical and bioinformatics (Ling, 2008; Bajcsy et al., 2005), education (Ranjan and Khalil, 2008) software engineering (Khatoon et al., 2011a, b), fraud detection (Thiruvadi and Patel, 2011), distributed (Hui, 2011; Yao et al., 2012). The data set of these applications is different from general ones which call for special mining requirements.

With the advancement in sensors’ technology sensor networks are increasingly finding its applications in many domains such as, habitat monitoring (Rozzyev et al., 2011) object tracking (Chaudhary et al., 2009), environment monitoring (Sabri et al., 2011; Lee and Chen, 2008; Kumar, 2011), military (Arampatzis et al., 2005; Yick et al., 2008). An enormous amount of data has been collected from these applications. Data mining techniques can be applied to this data for solving various problems. The traditional architecture in SNs deployed over an area of interest to sample data and send the results to a data-collection point, i.e., base station (BS) through an onboard radio transceiver. Such technological development has encouraged researchers to visualize the limited capabilities of the individual sensors in a large scale network that can operate unattended (Estrin et al., 1999; Pottie and Kaiser, 2000; Akyildiz et al., 2002; Chong and Kumar, 2003; Wang et al., 2003a). In many applications SNs are deployed on hostile and difficult to access location with limited energy, storage, computational power and bandwidth. Development of algorithms that consider the characteristics of sensor networks such as energy and computation constraints, network dynamics etc. constitutes an active area of current research (Gao et al., 2010; Han and Gao, 2008).
The extraction of useful knowledge from raw sensor data is a difficult task and conventional tools might not be able to handle the massive quantity, the high dimensionality and the distributed nature of the data. For such reasons, in recent years a great interest emerged in the research community in applying data mining techniques to the large volumes of sensor data.

In this study, SNs techniques with respect to sensor data mining process are critically analyzed. As a result a qualitative analysis is presented in tabular form. This work not only provides significant contributions to the SNs research, also exposes how challenging to compare different techniques due to the unavailability of complete data set used to evaluate the study. Since most of the studies have only reported the results without the availability of implemented tools therefore it is difficult to compare the existing techniques quantitatively.

CLASSIFICATION OF DATA MINING TECHNIQUES FOR SENSORS’ DATA

Data mining community has observed that extracting knowledge from SNs data imposes two kinds of challenges. At one hand, network side processing techniques require real-time analysis methodologies and systems to handle dynamic data streams or events; on the other hand centralized processing through high end computing is required for generating off-line predictive insights which facilitates real-time analysis. To address these challenges there are primarily two approaches for mining SNs data are adopted in literature. First, data from SNs is directed to central location where the entire processing take places at central side called centralize data models. Second, instead of sending the raw data to the central site, sensor nodes use their processing abilities to locally carry out mining processing and transmit only the required and partially processed data called local model. This paper attempts to characterize and classify the related approaches whether it applies mining techniques on network side or central side.

Mining techniques at network side: This section discusses the techniques that use the distributed nature of sensors and apply distributed mining algorithms on sensor data. In these techniques most of the work is pushed to the sensors themselves to build local models. The techniques are classified under following types on the basis of major data mining task it used for finding meaningful patterns.

Frequent pattern mining: Loo et al. (2005) proposed a framework for extracting association rules from sensor network. They proposed a model that analyzes the representation of sensor data streams for mining by proposing an Interval-List (IL). The analysis of IL based presentation of stream data showed favorable results using synthetic data set.

Romero (2006) proposed an in-network data mining technique to discover frequent patterns of events with certain spatial and temporal properties. Major issues in this approach are memory consumption of itemset discovery algorithms and the communication overhead of event collection.

Boukerche and Samarah (2007) proposed a distributed data extraction methodology for extracting association rules from SN by attaching a storage device to each node. Major issues in this methodology are increase in cost for node buffer and also delay in crucial messages in case of high support value.

Boukerche and Samarah (2008) proposed sensor association rules in which the event-detecting sensors are the main objects of the rules regardless of their values. A data structure called Positional Lexicographic Tree (PLT) is maintained to store the sensor's event-detecting status. The issue is multi database scans and the additional PLT update during mining limits the efficient use of this approach in handling SN data.

Chong et al. (2008) proposed a model that finds strong rules from sensor readings and use these learnt rules as a triggers to control sensor network operations or supplement sensor operations. To mine the rules A. priori is modify to count the number of transactions that are frequent instead the itemset with in transaction and transaction are processed in a batches b1, b2,..., bk.

Sequential pattern mining: Tseng and Lin (2007) and Tseng and Lu (2009) proposed an object tracking strategy named Multi Level Object Tracking (MLOT) to discover sequential patterns in object tracking sensor networks (OTSNs) by mining the movement log in sensor networks. A multi-level hierarchical structure is adapted by using the clustering mechanism that represents the hierarchical relations among sensor nodes to achieve the goal of keeping track of moving objects in a real-time manner. Through experimental evaluation of various simulated conditions, the proposed method is shown to deliver excellent performance in terms of both energy efficiency and timeliness.

Samarah et al. (2011) proposed Prediction-based Tracking technique by using Sequential Patterns (PTSPs). This technique helps to predict the future location of a moving object with the minimum number of sensor nodes while keeping the other nodes in sleep mode. The simulation results compared with other predictive
techniques shows that PTSPs is much efficient in energy consumption and missing rate recovery mechanism.

Esposito et al. (2011) presented a relational framework to identify the hidden frequent temporal correlations between sensor nodes. The framework exploits the relational language to describe the temporal evolution of a sensor network along with contextual information. By taking into account interval-based temporal data along with contextual information about events it discovers interesting and more human readable patterns.

**Clustering:** Dai et al. (2008) proposed a cluster formation protocol which consumes less energy and achieves longer network lifetime called Algorithm of Cluster head Election by counting (ACE-C). Sensor nodes transmit information only to cluster heads and then heads will communicate the aggregated information to the center. Heinzelman et al. (2000), proposed a popular clustering protocol for SNs called Low Energy Adaptive Clustering Hierarchy (LEACH). It forms clusters based on the received signal strength and uses the Cluster Head (CH) nodes as routers to the base-station. LEACH forms clusters by using a distributed algorithm, where nodes make autonomous decisions without any centralized control. LEACH Centralized (LEACH-C) (Heinzelman et al., 2002), improves LEACH in terms of data delivery by the use of an overlooking base station to evenly distribute cluster head nodes throughout the sensor network. The base station does this by computing the average node energy and allows only nodes over the average node energy to become cluster heads for the current iteration.

Younis and Fahmy (2004), proposed energy-based distributed clustering protocol named Hybrid Energy-Efficient Distributed clustering (HEED) which considers a hybrid of energy and communication cost when selecting Cluster Heads (CHs). The protocol probabilistically selects several nodes as CHs according to their residual energy. In HEED, each node is mapped to exactly one cluster and can directly communicate with its CH.

Taherkordi et al. (2008) proposed a communication efficient distributed protocol for clustering sensory data. A distributed version of k-means clustering algorithm is adopted which reduces communication transmission, time and power consumption of sensor nodes. Major issues are extra memory for cluster head and computation power for summarization of data before transmitting to sink. Guo et al. (2009) proposed H-Cluster a distributed algorithm to cluster sensory data. The results show that H-Cluster algorithm is much efficient in energy conservation and the quality of cluster data in small SN. Beyns et al. (2005) proposed a cluster-based architecture for wireless sensor networks. A prediction model is executed to predict the spatiotemporal correlation of cluster heads. Several centralized clustering techniques, such as k-means and hierarchical clustering (Han and Kamber, 2001), used in this method.

Yoon and Shahabi (2007) presented the Clustered Aggregation (CAG) algorithm that forms clusters of nodes sensing similar values. However, CAG is a lossy clustering algorithm (most sensory readings are never reported) which trades a lower result precision.

**Classification:** McConnell and Skillcorn (2005), presented a classification framework for building and deploying predictors in sensor networks. The proposed framework is useful especially for sensor networks limited power resources, computation, bandwidth and data confidentiality problems. Major issues in this framework are we need an expensive CPU on each sensor node for computation and building local predictive model and also extra memory is required to store local predictive model.

Sharma et al. (2011) proposed a methodology for classifying the sensors data by using a Nearest Neighbor Trajectory Classification (NNTC). When a request to classify a query vector is made the closest training vector(s), according to a distance metric are located. The classes of these training vectors are used to assign a class to the query vector. The nearest-neighbor method predicts the class of a test example.

Malhotra et al. (2008) proposed a distributed classification scheme for classifying the audio signals from wireless sensor network. A distributed cluster-based algorithm for detection and classification of vehicles has been proposed. Two approaches are proposed: first combines extracted features and the second combines' individual decisions. Classification results using decision fusion and a Maximum Likelihood (ML) classifier led to the best results.

Flouri et al. (2006a, b) proposed a distributed and incremental techniques for training Support Vector Machines (SVMs) in sensor networks. Two distributed algorithms for training a SVM are applied to the classification problem in WSN. The challenge for these SVM formulations for this application is their computational complexity.

**Mining techniques at central side:** Plenty of centralized techniques are proposed in literature which is classified under major data mining method use to discover meaningful patterns.
**Frequent pattern mining:** Le-Gruenwald (2005), proposed a data model to store and updates the data generated from sensor networks. They adopt a centralized extraction methodology where all the data is collected at sink for further analysis. Data Stream Association Rule Mining (DSARM) framework is proposed to identify the missed readings that result from the loss and corruption of messages.

Jiang (2007) and Jiang and Gruenwald (2007) proposed a data estimation technique, called CARM (Closed Itemsets based Association Rule Mining) which can derive the most recent association rules between sensors based on the current closed itemsets in the sliding window. The technique is based on the Closed Frequent Itemsets mining algorithm in data streams called CF-stream (Jiang and Gruenwald, 2006). It maintains an in-memory data structure, called Direct Update (DUI) to store closed itemsets. The results show the algorithm achieve time and space efficiency. However, the algorithms estimates the missing data according to the frequent patterns which are pre-computed based on the existing data. If the pattern containing the missing data does not appear in the frequent patterns, the missing data cannot be estimated.

Chi et al. (2004) proposed an algorithm called Moment to mine closed frequent itemsets over a data stream sliding window. Moment stores much more information other than the current closed frequent itemsets which consumes much memory, especially when the support threshold is low. Furthermore, the exploration and node type checking are time consuming.

Giannella et al. (2003) proposed FP-tree-based algorithm called FP - Stream to mine frequent itemsets at multiple time granularities over data streams. A novel tilted-time window is proposed to calculate the frequent patterns for the most recent transactions. A number of experiments are conducted to prove the algorithm’s efficiency.

Tanbeer et al. (2009), proposed a tree-based data structure called Sensor Pattern Tree (SP-Tree) to generate association rules from SnI data with one database scan. The main idea of proposed approach is to obtain the frequency of all event-detecting sensors’ data and construct a SP-Tree. Results shows SP-Tree based approach is less time consuming because it constructs the tree by scanning the database only once. Moreover, the mining phase of SP-tree is highly efficient due to the frequency-descending tree structure.

**Sequential pattern mining:** Cook et al. (2003) presented MovHome an intelligent home which perceives the state of the home through sensors and acting upon the environment through device controllers. An important component of proposed architecture is the ability to make decisions based on predicted activities. A key disadvantage is the fact that the entire action history must be stored and processed offline which is not practical for large prediction tasks over a long period of time.

Wu et al. (2001) proposed an algorithm for mining sequential alarm patterns from the alarm data of Global System for Mobile Communication (GSM). By specifying the referent urgent window d in sequential alarm pattern it predicts and controls the alarms. For example, if d is set as six hours, the sequential alarm pattern (a, b, c) indicates that a, b and c happen in order and that the time interval between a and b and between b and c is less than 6 h.

Rabatel et al. (2009) presented a strategy to detect anomalies from sensor data to improve the railway maintenance. It extracts sequence pattern from real railway data and identify the abnormal behavior or patterns. PSP algorithm has been used to identify the sequential patterns. To tackle the environments conditions a contextual knowledge based method is proposed. If new data has any anomalies this method provide the reason and seriousness of each anomaly.

Guralnik and Haigh (2002) used sequential pattern mining to learn typical behaviors of humans in their homes. Human behavior is inferred by using motion sensors, pressure pads, door latch sensors and toilet flush sensors. By analyzing from room sequence in which the person was acting complex behavioral models are identified.

**Clustering:** Aggarwal et al. (2003) proposed a framework for clustering data streams, called the ChuStream algorithm. The proposed technique divides the clustering process into two components. The online component stores summary statistics about the data streams and the offline one performs clustering on the summarized data. A number of experiments on real data sets have been conducted to prove the accuracy and efficiency of the proposed algorithm.

Ordonez (2003) proposed the improvements to the k-means algorithm to cluster binary data streams. Three variant of k-means algorithms includes: On-line k-means, Scalable k-means and Incremental k-means are proposed. The experiments were conducted on real data sets as well as synthetic data sets. They demonstrated that the proposed algorithm outperforms the scalable k-means in most of the cases.

O’Callaghan et al. (2002) proposed an algorithm called Stream and LocalSearch algorithms. The Stream algorithm starts by determining the size of the sample and
then applies the LocalSearch algorithm if the sample size is larger than a pre-specified equation result. This process is repeated for each data chunk. Finally, the LocalSearch algorithm is applied to the cluster centers generated in the previous iterations.

Papadimitriou et al. (2003) proposed an algorithm named Arbitrary Window Stream Modeling Method (AWSOM) to discover interesting patterns and trends from sensors. AWSOM is a one-pass algorithm to incrementally update the patterns. It requires only $O(N \log N)$ memory where $N$ is the length of the sequence. Wavelet coefficients are used to compress information representation and correlation structure detection and linear regression model is applied in the wavelet domain. Experiments are conducted on real and synthetic data sets.

Chen and Tu (2007) proposed a framework D-Stream to efficiently identify outliers of clusters. In this framework each input data is mapped into a grid, density of each grid is computed and grids are clustered using a density-based algorithm. This method can discover clusters with arbitrary shapes and detect many outliers around original clusters.

Classification: Aggarwal et al. (2004) have applied the idea of micro-clusters introduced in CluStream (Aggarwal et al., 2003) for On-Demand classification. The technique uses clustering results to classify data using statistics of class distribution in each cluster. A classification system is created in which the training model is quickly adapted to the changes of the underlying data stream. On-demand classification process dynamically selects the appropriate window of past training data to build the classifier. The results show high classification accuracy in an evolving data stream.

Domingos and Hulten (2000, 2001), Hulten et al. (2001) presented a general method called Very Fast Decision Tree (VFDT) to mine high speed stream data. It is using a small sample of available examples when choosing the split attribute at any given node. Only the first example to arrive at data stream need to be used to choose split attribute at the root, subsequent ones are passed through induced portion of the tree until they reach a leaf.

Chichakoui et al. (2010) proposed Decision Tree (DT) based classification technique for person identification in ubiquitous environment. In order to identify persons, the proposed approach first extracts frequent episodes from the datasets. Next, it assigns weights to these episodes. Finally, it applies Decision Tree (DT) to classify the frequent episodes. Gaber et al. (2005) proposed Algorithm Output Granularity (AOG) based on Light Weight Classification (LWClass) for on-board mining of data streams in sensor networks. The algorithm is empirically validated using synthetic streaming data and the resource-constrained environment of a common handheld computer.

Wang et al. (2003b) proposed a general framework for mining concept drifting data streams. The proposed framework use weighted classifier ensembles to mine data streams. They use synthetic and real life data streams to test their algorithm and compare between the single classifier and classifier ensembles. The proposed algorithm combines multiple classifiers weighted by their expected prediction accuracy. Also the selection of a number of classifiers instead of using all is an option in the proposed framework without losing accuracy.

**ANALYSIS OF DATA MINING TECHNIQUES FOR SENSORS’ DATA**

In previous section data mining applications used for mining sensor network are critically analyzed. As a result a qualitative analysis is presented shown in Table 1 and Table 2. The analysis shows the efficiency of data mining applications in developing sensor data mining algorithm. Furthermore, it highlights the objectives of data mining function being used, its strength and open issues in glimpse. Additionally, it directs the future research work for interested researcher in this area of research.

It can be observed from analysis that different types of data mining algorithms are developed for different type of sensors’ data. Data mining methods for the sake of association rule mining, sequential pattern mining, clustering and classification are widely used for extracting meaningful patterns from sensor data. On the basis of results drawn from analysis the amount of work done in each data mining method is also analyzed. Figure 1 shows

![Fig. 1: Mining techniques applied for each side](image)
### Table 1: Critical analysis of data mining techniques for network side

<table>
<thead>
<tr>
<th>Technique</th>
<th>Objectives</th>
<th>Strengths</th>
<th>Data source</th>
<th>Real</th>
<th>Synthetic</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent pattern</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apriori</td>
<td>Extracting rules from sensor values</td>
<td>Memory efficient</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Time consuming</td>
</tr>
<tr>
<td>Frequent patterns</td>
<td>Find spatiotemporal relationships</td>
<td>Data reduction</td>
<td>x</td>
<td></td>
<td></td>
<td>Communication overhead</td>
</tr>
<tr>
<td>Frequent patterns</td>
<td>Identify the correlated sensors</td>
<td>Message reduction</td>
<td>x</td>
<td></td>
<td></td>
<td>Increases buffer</td>
</tr>
<tr>
<td>Frequent patterns</td>
<td>Extract behavioral patterns</td>
<td>Predict future events</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Multi database scan</td>
</tr>
<tr>
<td>FP-growth</td>
<td>Infer values from previous</td>
<td>Efficient power management</td>
<td>x</td>
<td></td>
<td></td>
<td>Information missing</td>
</tr>
<tr>
<td>Sequential pattern</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLOT</td>
<td>Prediction of moving object</td>
<td>Energy efficient</td>
<td>x</td>
<td></td>
<td></td>
<td>Memory consuming</td>
</tr>
<tr>
<td>PTSPs</td>
<td>Prediction of moving object</td>
<td>Mining data recovery</td>
<td>x</td>
<td></td>
<td></td>
<td>Difficult to predict high speed objects</td>
</tr>
<tr>
<td>Relational pattern</td>
<td>Mining multi-dimensional patterns</td>
<td>Human readable patterns</td>
<td>x</td>
<td></td>
<td></td>
<td>Memory and time consuming</td>
</tr>
<tr>
<td>Clustering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACE-C</td>
<td>Save energy consumption</td>
<td>Less energy consumption</td>
<td>x</td>
<td></td>
<td></td>
<td>Data redundancy Data delay</td>
</tr>
<tr>
<td>LEACH</td>
<td>Save energy consumption</td>
<td>Self organizing clustering</td>
<td>x</td>
<td></td>
<td></td>
<td>Low scalability and stability</td>
</tr>
<tr>
<td>HEED</td>
<td>Save energy consumption</td>
<td>Prolong SN lifetime</td>
<td>x</td>
<td></td>
<td></td>
<td>Energy lost</td>
</tr>
<tr>
<td>k-means</td>
<td>Future analysis of sensory data</td>
<td>Reduced power consumption</td>
<td>x</td>
<td></td>
<td></td>
<td>Effects communication when number of sensor s are high</td>
</tr>
<tr>
<td>II-Clusters</td>
<td>Sensory data clustering</td>
<td>Best cluster quality</td>
<td>x</td>
<td></td>
<td></td>
<td>High data loss rate</td>
</tr>
<tr>
<td>General architecture</td>
<td>Prediction based monitoring</td>
<td>Prolong SN lifetime</td>
<td>x</td>
<td></td>
<td></td>
<td>Clustering over heads</td>
</tr>
<tr>
<td>CAG</td>
<td>Network structure maintenance</td>
<td>Data reduction, Energy efficient</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Sensor data loss</td>
</tr>
<tr>
<td>Classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JDBC decision tree</td>
<td>Distributed predictive model</td>
<td>Data confidentiality, Easy deployment</td>
<td>x</td>
<td></td>
<td></td>
<td>Extra Memory required, Increase power computation</td>
</tr>
<tr>
<td>NNTC</td>
<td>Data classification</td>
<td>General purpose</td>
<td>x</td>
<td></td>
<td></td>
<td>Not evaluated on real dataset</td>
</tr>
<tr>
<td>kNN, ML</td>
<td>Vehicle classification</td>
<td>High classification accuracy</td>
<td>x</td>
<td></td>
<td></td>
<td>High computation Large memory</td>
</tr>
<tr>
<td>SVM training method</td>
<td>Energy-efficient distributed algorithms</td>
<td>Incremental learning</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Computational complexity</td>
</tr>
</tbody>
</table>

### Table 2: Critical analysis of data mining techniques for central side

<table>
<thead>
<tr>
<th>Technique</th>
<th>Objectives</th>
<th>Strengths</th>
<th>Data source</th>
<th>Real</th>
<th>Synthetic</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent pattern</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSARM</td>
<td>Estimating sensors' missing</td>
<td>Estimated value accuracy values</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Ignore the sensor that reports different values</td>
</tr>
<tr>
<td>CARM</td>
<td>Missing value estimation</td>
<td>Compute exact sets of IDS</td>
<td>x</td>
<td></td>
<td></td>
<td>Inefficient for handling high speed streams</td>
</tr>
<tr>
<td>Moment</td>
<td>Discovering closed frequent remotely over sliding window</td>
<td>Incremental update</td>
<td>x</td>
<td></td>
<td></td>
<td>Memory and time consuming</td>
</tr>
<tr>
<td>FP-Stream</td>
<td>Mining time sensitive data streams</td>
<td>Incremental pruning and update of items</td>
<td>x</td>
<td></td>
<td></td>
<td>FP-stream grow larger and larger over time</td>
</tr>
<tr>
<td>FP-growth</td>
<td>Mining tree structures</td>
<td>Memory efficient</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Tree construction time and cost is high</td>
</tr>
<tr>
<td>Sequential pattern</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSAP</td>
<td>Mining sequential alarm patterns</td>
<td>Fault prediction</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Increase cost (multiple DB scan)</td>
</tr>
<tr>
<td>GSP</td>
<td>Intelligent home creation</td>
<td>Reduced operation cost</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Fixed time interval</td>
</tr>
<tr>
<td>PEP</td>
<td>Extract sequence patterns and detect anomalies</td>
<td>More effective on offline anomaly prediction</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Real time Anomaly prediction is missing</td>
</tr>
<tr>
<td>Tree projection</td>
<td>Infer human behavior human behavior</td>
<td>Automatic prediction of</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Technique is partially validated</td>
</tr>
<tr>
<td>Clustering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustream</td>
<td>Clustering large evolving data streams</td>
<td>Higher quality and scalability</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Offline clustering</td>
</tr>
<tr>
<td>Incremental k-means</td>
<td>Clustering binary data streams</td>
<td>Fast and efficient</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Only suitable for sparse binary data</td>
</tr>
<tr>
<td>Stream local search</td>
<td>Clustering data streams</td>
<td>Incremental learning</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Low clustering quality in high speed streams</td>
</tr>
<tr>
<td>AWSOM</td>
<td>Pattern and trend prediction</td>
<td>Dynamic update single pass</td>
<td>x</td>
<td></td>
<td>x</td>
<td>High complexity</td>
</tr>
<tr>
<td>D-Stream</td>
<td>Density-based clustering</td>
<td>Concept drift detection</td>
<td>x</td>
<td></td>
<td></td>
<td>High complexity</td>
</tr>
<tr>
<td>Classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LWClass</td>
<td>Classifying data streams</td>
<td>High speed less memory</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Time consuming and costly learning</td>
</tr>
<tr>
<td>Ensemble-based</td>
<td>Mining concept drifting</td>
<td>Single pass dynamic update data streams</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Low speed Storage memory problem</td>
</tr>
<tr>
<td>VKFM</td>
<td>Decision tree learning system</td>
<td>High speed less memory</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Costly learning</td>
</tr>
<tr>
<td>On-demand stream classifier</td>
<td>Dynamic classification model</td>
<td>Dynamic update high speed</td>
<td>x</td>
<td></td>
<td>x</td>
<td>High cost and time need for labeling</td>
</tr>
<tr>
<td>DT</td>
<td>Person identification in ubiquitous environment</td>
<td>Improved classification results</td>
<td>x</td>
<td></td>
<td>x</td>
<td>Increase cost (multiple DB scan)</td>
</tr>
</tbody>
</table>
specific data mining algorithm and a certain type of SNs data for the sake of solving a certain kind of problems. All these approaches shows an efforts to improve the performance of traditional mining algorithm by using additional knowledge and constraint nature of sensors for improving the quality of final results.

CONCLUSION

The emerging need for involvement of data mining algorithms in the field of SNs resulted in the development of numerous algorithms. Each one of these algorithms solves certain particular issues related to the appropriate SNs type. This study presented an overview of selected existing research approaches and classifies the past research either they use centralized or distributed data mining approach. From analysis it is observed that the techniques intended for mining sensor data at network side are helpful for taking real time decision as well as serve as prerequisite for development of effective mechanism for data storage, retrieval, query and transaction processing at central side, on the other hand centralized techniques are helpful in generating off-line predictive insights which in turn can facilitate real-time analysis. This study presented the strengths, limitations and an overall analysis of the past research which can provide insights for end-users in applying or developing an appropriate data mining techniques for sensor networks.

ACKNOWLEDGMENTS

This study was supported in part by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 20110141100662 and International S&T Cooperation Program of Hubei Province under Grant No. 2010BFA008.

REFERENCES


Han, J. and J. Gao, 2008. Research Challenges for Data Mining in Science and Engineering. In: Next Generation of Data Mining, Kargupta, H., J. Han, P.S. Yu, R. Motwani and V. Kumar (Eds.). Chapman and Hall/CRC, USA.


