WIRELESS NETWORKS: A COMPARISON AND CLASSIFICATION BASED ON OUTLIER DETECTION METHODS

T. KAVITHA*, A. CHANDRA*
*Department of Computer Science and Systems Engineering , Sree Vidyanikethan Engineering College
A. Rangampet, Tirupati
*kavithakrishnag19@gmail.com, chandra_adhimulam@yahoo.com

ABSTRACT
Outlier detection has been used for centuries to detect and, where appropriate, remove anomalous observations from data. Outliers arise due to mechanical faults, changes in system behavior, fraudulent behavior, human error, instrument error or simply through natural deviations in populations. Their detection can identify system faults and fraud before they escalate with potentially catastrophic consequences. It can identify errors and remove their contaminating effect on the data set and as such to purify the data for processing. The original outlier detection methods were arbitrary but now, principled and systematic techniques are used, drawn from the full gamut of Computer Science and Statistics. In the field of wireless sensor networks, measurements that significantly deviate from the normal pattern of sensed data are considered as outliers. The potential sources of outliers include noise and errors, events, and malicious attacks on the network. Traditional outlier detection techniques are not directly applicable to wireless sensor networks due to the multivariate nature of sensor data and specific requirements and limitations of the wireless sensor networks. In this paper we provide a comprehensive overview of existing outlier detection techniques specifically developed for the wireless sensor networks. Additionally, it presents a technique-based taxonomy and a decision tree to be used as a guideline to select a technique suitable for the application at hand based on characteristics such as data type, outlier type and outlier degree.

1. INTRODUCTION
In many data analysis tasks a large number of variables are being recorded or sampled. One of the first steps towards obtaining a coherent analysis is the detection of outlaying observations. Although outliers are often considered as an error or noise, they may carry important information. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modeling and analysis. An exact definition of an outlier often depends on hidden assumptions regarding the data structure and the applied detection method. Yet, some definitions are regarded general enough to cope with various types of data and methods. Hawkins (Hawkins, 1980) defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Barnett and Lewis (Barnett and Lewis, 1994) indicate that an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs, similarly, Johnson (Johnson, 1992) defines an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data. Other case-specific definitions are given below. Outlier detection methods [1] have been suggested for numerous applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, attack performance analysis, and other data-mining tasks. A wireless sensor network (WSN) typically consists of a large number of small, low-cost sensor nodes distributed over a large area with one or possibly more powerful sink nodes gathering readings of sensor nodes. The sensor nodes are integrated with sensing, processing and wireless communication capabilities. Each node is usually equipped with a wireless radio transceiver, a small microcontroller, a power source and multi-type sensors such as temperature, humidity, light, heat, pressure, sound, vibration, etc. The WSN is not only used to provide fine-grained real-time data about the physical world but also to detect time-critical events. A wide variety of applications of WSNs includes those relating to personal, industrial, business, and military domains, such as environmental and habitat monitoring, object and inventory tracking, health and medical monitoring, battlefield observation, industrial safety and control, to name but a few.

2. FUNDAMENTALS OF OUTLIER DETECTION IN WSN’S
This section describes fundamentals of outlier detection in WSNs, including definitions of outliers, various causes of outliers, motivation of outlier detection, and challenges of outlier detection in WSNs.

2.1 What is an Outlier?
The term outlier [1], also known as anomaly, originally stems from the field of statistics (Hodge and Austin, 2003). The two classical definitions of outliers are:
2.2 Motivation of Outlier Detection in WSNs [4]:
Outlier detection also known as anomaly detection or deviation detection, is one of the fundamental tasks of data mining along with predictive modeling, cluster analysis and association analysis. Compared with these other three tasks, outlier detection is the closest to the initial motivation behind data mining, i.e., mining useful and interesting information from a large amount of data (Han and Kamber, 2006). Outlier detection has been widely researched in various disciplines such as statistics, data mining, machine learning, information theory, and spectral decomposition. Also, it has been widely applied to numerous applications domains such as fraud detection, network intrusion, performance analysis, weather prediction, etc.

Here, we exemplify the essence of outlier detection in several real-life applications.

- **Environmental monitoring**, in which sensors such as temperature and humidity are deployed in harsh and unattended regions to monitor the natural environment.
- **Habitat monitoring**, in which endangered species can be equipped with small non-intrusive sensors to monitor their behavior.
- **Health and medical monitoring**, in which patients are equipped with small sensors on multiple different positions of their body to monitor their well-being.
- **Industrial monitoring**, in which machines are equipped with temperature, pressure, or vibration amplitude sensors to monitor their operation.
- **Target tracking**, in which sensors are embedded in moving targets to track them in real-time.
- **Surveillance monitoring**, in which multiple sensitive and unobtrusive sensors are deployed in restricted areas.

As illustrated in Figure 1, these topics include fault detection (Chen et al., 2006; Luo et al., 2006), event detection (Krishnamachari and Iyengar, 2004; Martinovic and Schwiebert, 2006; Ding et al., 2005) and intrusion detection (Silva et al., 2005; Bhuse and Gupta, 2006).

![Fig. 1. Three outlier sources in WSNs and their corresponding detection techniques](image)

2.3 Challenges of Outlier Detection in WSNs:
Extracting useful knowledge from raw sensor data is not a trivial task (Tan, 2006). The context of sensor networks and the nature of sensor data make design of an appropriate outlier detection technique more challenging. Due to the following reasons, conventional outlier detection techniques might not be suitable for handling sensor data in WSNs [2].

- **Resource constraints**: The low cost and low quality sensor nodes have stringent constraints in resources, such as energy, memory, computational capacity and communication bandwidth. Most of traditional outlier detection techniques have paid limited attention to reasonable availability of computational resources.
- **High communication cost**: In WSNs, the majority of the energy is consumed for radio communication rather than computation. For a sensor node, the communication cost is often several orders of magnitude higher than the computation cost (Akyildiz et al., 2002).
- **Distributed streaming data**: Distributed sensor data coming from many different streams may dynamically change. Moreover, the underlying distribution of streaming data may not be known a priori.

3. CLASSIFICATION CRITERIA OF OUTLIER DETECTION TECHNIQUES FOR WSN’S
This section identifies and discusses several important aspects of outlier detection techniques specially developed for WSNs. These aspects will be used as metrics to compare characteristics of different outlier detection techniques.
3.1 Input Sensor Data:
Sensor data can be viewed as data streams, i.e., a large volume of real-valued data that is continuously collected by sensor nodes (Gaber, 2007). The type of input data determines which outlier detection techniques can be used to analyze the data (Chandola et al., 2007). Outlier detection techniques usually consider the two following aspects of sensor data.

3.1.1 Attributes: A data measurement can be identified as outlier when its attributes have anomalous values (Tan et al., 2006). An outlier in univariate data with a single attribute can be easily detected if the single attribute is anomalous with respect to that attribute of other data.

3.1.2 Correlations: There are two types of dependencies at each sensor node, i.e., (i) dependencies among the attributes of the sensor node, and (ii) dependency of sensor node readings on history and neighboring node readings (Janakiram et al., 2006).

3.2 Type of Outliers [5]:
Compared to centralized approach, in which the entire data is processed in a central place, outliers in WSNs can be analyzed and identified in many different nodes in the network. This multi-level outlier detection in WSNs makes local models generated from data streams of individual nodes totally different than the global one (Subramaniam et al., 2006). Depending on the scope of data used for outlier detection, outlier may be either local or global.

3.2.1 Local Outliers: Due to the fact that local outliers are identified at individual sensor nodes, techniques for detecting local outliers save communication overhead and enhance the scalability. Local outlier detection can be used in many event detection applications, e.g., vehicle tracking, surveillance monitoring.

3.2.2 Global Outliers: Global outliers are identified in a more global perspective. They are of particular interest since analysts would like to have a better understanding of overall data characteristics in WSNs.

3.3 Identity of Outliers:
There are three sources of outliers occurred in WSNs: (1) errors, (2) events, and (3) malicious attacks. The sort of outliers caused by malicious attacks is concerned with the issue of network security and is out of the scope of this paper. For outliers resulted from different sources, outlier detection techniques are desired to specify the identity of these outliers and deal further with them.

3.3.1 Errors: An error refers to a noise-related measurement or data coming from a faulty sensor. Outliers caused by errors may occur frequently, while outliers caused by events tend to have extremely smaller probability of occurrence.

3.3.2 Events: An event is defined as particular phenomena that change the real-world state, e.g., forest fire, air pollution, etc. This sort of outlier normally lasts for a relatively long period of time and changes historical pattern of sensor data. However, faulty sensors may also generate similar long segmental outliers as events and therefore it is hard to distinguish the two different outlier sources only by examining one sensing series of a node itself (Zhuang and Chen, 2005).

4. TAXONOMY FRAMEWORK FOR OUTLIER DETECTION TECHNIQUES FOR WSN’S
Recently, many outlier detection techniques specifically developed for WSNs have emerged. In this section, we provide a technique-based taxonomy framework to categorize these techniques. As illustrated in Figure 2, outlier detection techniques for WSNs can be categorized into statistical-based, nearest neighbor-based, classification-based, and spectral decomposition-based approaches.

5. OUTLIER DETECTION TECHNIQUES FOR WSN’S
In this section, we classify outlier detection techniques designed for WSNs based on the discipline from which they adopt their ideas and address the key characteristics and performance analysis of each outlier detection technique using the taxonomy framework presented in Section 4. Furthermore, we provide an evaluation for each of these disciplines.

5.1 Statistical-Based Approaches:
Statistical-based approaches are the earliest approaches to deal with the problem of outlier detection. The statistical outlier detection techniques are essentially model-based techniques. They assume or estimate a statistical (probability distribution) model which captures the distribution of the data and evaluate data instances with respect to how well they fit the model.

5.1.1 Parametric-Based Approaches:
Parametric techniques assume availability of the knowledge about underlying data distribution, i.e., the data is generated from a known distribution. It then estimates the distribution parameters from the given data. Based on type of distribution assumed, these techniques are further categorized into Gaussian-based models and non-Gaussian-based models. In Gaussian models, the data is assumed to be normally distributed (Chandola et al., 2007).
-- Gaussian-based models. Wu et al. (2007) present two local techniques for identification of outlying sensors as well as identification of event boundary in sensor networks. These techniques employ the spatial correlation of the readings existing among neighboring sensor nodes to distinguish between outlying sensors and event boundary. In the technique for identifying outlying sensors, each node computes the difference between its own reading and the median reading from its neighboring readings. Then it standardizes all differences from its neighborhood. Bettencourt et al. (2007) present a local outlier detection technique to identify errors and detect events in ecological applications of WSNs.

-- Non-Gaussian-based models. Jun et al. (2006) present a statistical-based technique, which uses a symmetric α-stable (SαS) distribution to model outliers being in form of impulsive noise. The technique utilizes the spatio-temporal correlations of sensor data to locally detect outliers.

Fig. 2. Taxonomy of outlier detection techniques for WSNs.

Each node in a cluster first detects and corrects temporal outliers by comparing the predicted data and the sensing data.

5.1.2 Non-Parametric-Based Approaches [6]: Non-parametric techniques do not assume availability of data distribution. They typically define a distance measure between a new test instance and the statistical model and use some kind of thresholds on this distance to determine whether the observation is an outlier. Two most widely used approaches in this category are histograms and kernel density estimator. Histogramming models involve counting frequency of occurrence of different data instances (thereby estimating the probability of occurrence of a data instance) and compare the test instance with each of the categories in the histogram and test whether it belongs to one of them. A new instance that lies in the low probability area of this pdf is declared as an outlier (Chandola et al., 2007).

-- Histogramming. Sheng et al. (2007) present a histogram-based technique to identify global outliers in data collection applications of sensor networks. This technique attempts to reduce communication cost by collecting histogram information rather than collecting raw data for centralized processing.

-- Kernel functions. Palpanas et al. (2003) propose a kernel-based technique for online identification of outliers in streaming sensor data. This technique requires no a priori known data distribution and uses kernel density estimator to approximate the underlying distribution of sensor data.

5.1.3 Evaluation of Statistical-Based Techniques: Statistical-based approaches are mathematically justified and can effectively identify outliers if a correct probability distribution model is acquired. Moreover, after constructing the model, the actual data on which the model is based on is not required. However, in many real-life scenarios, no a priori knowledge of the sensor stream distribution is available.

5.2 Classification-Based Approaches

Classification approaches are important systematic approaches in the data mining and machine learning community. They learn a classification model[3] using the set of data instances (training) and classify an unseen instance into one of the learned (normal/outlier) class (testing). The unsupervised classification-based techniques require no knowledge of available labeled training data and learn the classification model which its the majority of the data instance during training.

5.2.1 Support Vector Machine-Based Approaches: SVM techniques separate the data belonging to different classes by fitting a hyper plane between them which maximizes the separation. The data is mapped into a higher dimensional feature space where it can be easily separated by a hyper plane.

5.2.2 Bayesian Network-Based Approaches: Bayesian network-based approaches use a probabilistic graphical model to represent a set of variables and their probabilistic independencies. They aggregate information from different variables and provide an estimate on the expectancy of an event to belong to the learned class.
(Chandola et al., 2007). They are categorized as naive Bayesian network, Bayesian belief network, and dynamic Bayesian network approaches based on degree of probabilistic independencies among variables.

-- **Naive Bayesian Network models.** Elnahrawy and Nath (2004) present a Bayesian model-based technique to discover local outliers and detect faulty sensors. This technique maps the problem of learning spatio-temporal correlations to the problem of learning the parameters of the Bayesian classifier and then uses the classifier for probabilistic inference.

### 5.2.3 Evaluation of Classification-based Technique:
Classification-based approaches provide an exact set of outliers by building a classification model to classify [7]. However, a main drawback of SVM-based techniques is their computational complexity and the choice of proper kernel function. Learning the accurate classification model of a Bayesian network is challenging if the number of variables is large in deployed WSNs.

**CONCLUSION**

In this paper, we address the problem of outlier detection in WSNs and provide a technique-based taxonomy framework to categorize current outlier detection techniques designed for WSNs. We also introduce the key characteristics and brief description of current outlier detection techniques using the proposed taxonomy framework and provide an evaluation for each technique. Furthermore, we present a decision tree to compare these techniques in terms of the nature of sensor data, characteristics of outlier and outlier detection. The shortcomings of existing techniques for WSNs clearly calls for developing outlier detection technique, which takes into account multivariate data and the dependencies of attributes of the sensor node, provides reliable neighborhood, proper and flexible decision threshold, and also meets special characteristics of WSNs such as node mobility, network topology change and making distinction between errors and events.

**REFERENCES**


