Outlier Detection Techniques for Wireless Sensor Networks: A Survey

Yang Zhang, Nirvana Meratnia, and Paul Havinga

***Abstract*—In the field of wireless sensor networks, those 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 nature of sensor data and specific requirements and limitations of the wireless sensor networks. This survey provides a comprehensive overview of existing outlier detection techniques specifically developed for the wireless sensor networks. Additionally, it presents a technique-based taxonomy and a comparative table 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, outlier identity, and outlier degree.**

***Index Terms*—Outlier, outlier detection, wireless sensor networks, taxonomy.**

I. Introduction

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. In many of these applications, real-time data mining of sensor data to promptly make intelligent decisions is essential [1].

Data measured and collected by WSNs is often unreliable. The quality of data set may be affected by noise and error, missing values, duplicated data, or inconsistent data. The low cost and low quality sensor nodes have stringent resource constraints such as energy (battery power), memory, computational capacity, and communication bandwidth. The limited resource and capability make the data generated by sensor nodes unreliable and inaccurate. Especially when battery power is exhausted, the probability of generating erroneous data will grow rapidly [2]. On the other hand, operations of sensor nodes are frequently susceptible to environmental effects. The vision of large scale and high density wireless sensor network is to randomly deploy a large number of sensor nodes (up to hundreds or even thousands of nodes) in harsh and unattended environments. It is inevitable that in such environments some sensor nodes malfunction, which may result in noisy, faulty, missing and redundant data. Furthermore, sensor nodes are vulnerable to malicious attacks such as denial of service attacks, black hole attacks and eavesdropping [3], in which data generation and processing will be manipulated by adversaries.

The above internal and external factors lead to unreliability of sensor data, which further influence quality of raw data and aggregated results. Since actual events occurred in the physical world, e.g., forest fire, earthquake or chemical spill, cannot be accurately detected using inaccurate and incomplete data [4], it is extremely important to ensure the reliability and accuracy of sensor data before the decision-making process.

Due to the fact that outliers are one of the sources to greatly influence data quality, in this survey we provide a comprehensive overview of the research done in the field of outlier detection in WSNs, evaluate and compare existing outlier detection techniques specifically developed for WSNs, and identify potential areas for further research. To the best of our knowledge, this survey is the first attempt to provide a comparative table 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, outlier identity, and outlier degree.

The contributions of this survey can be summarized as:

- describing the fundamentals of outlier detection in WSNs (Section II).
- identifying important criteria associated with the classification of outlier detection techniques for WSNs (Section III).
- providing a technique-based taxonomy to categorize existing outlier detection techniques developed for WSNs (Section IV).
- addressing the key characteristics and brief description of current outlier detection techniques using the presented taxonomy (Section V).
- comparing existing techniques and introducing a comparative table to select the suitable technique based on data and outlier characteristics (Section VI).

II. Fundamentals of Outlier Detection in Wireless Sensor Networks

This section describes fundamentals of outlier detection in WSNs, including definitions of outliers, various causes of

Manuscript received 20 November 2008; revised 27 February 2009.

Y. Zhang, N. Meratnia and P.J.M. Havinga are with the Department of Computer Science, University of Twente, Drienerlolaan 5, 7522NB Enschede, The Netherlands (e-mail: {zhangy, meratnia, havinga}@cs.utwente.nl).

Digital Object Identifier 10.1109/SURV.2010.021510.00088

outliers, motivation of outlier detection, and challenges of outlier detection in WSNs.

A. What is an Outlier?

The term *outlier*, also known as *anomaly*, originally stems from the field of *statistics* [5]. The two classical definitions of outliers are:

(Hawkins [6]): *"an outlier is an observation, which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"*.

(Barnett and Lewis [7]): *"an outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data"*.

In addition, a variety of definitions depending on the particular method outlier detection techniques are based upon exist [8]. Each of these definitions signify the solutions to identify outliers in a specific type of data set.

In WSNs, outliers can be defined as, *"those measurements that significantly deviate from the normal pattern of sensed data"* [9]. This definition is based on the fact that in WSN sensor nodes are assigned to monitor the physical world and thus a pattern representing the normal behavior of sensed data may exist. Potential sources of outliers in data collected by WSNs include *noise* and *errors*, *actual events*, and *malicious attacks*. Noisy data as well as erroneous data should be eliminated or corrected if possible as noise is a random error without any real significance that dramatically affects the data analysis [10]. Outliers caused by other sources need to be identified as they may contain important information about events that are of great interest to the researchers.

B. Motivation of Outlier Detection in WSNs

Outlier detection also known as *anomaly detection* or *deviation detection*, is one of the fundamental tasks of *data mining* along with predictive modelling, cluster analysis and association analysis [10]. 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* [11]. Outlier detection has been widely researched in various disciplines such as statistics, data mining, machine leaning, information theory, and spectral decomposition [9]. Also, it has been widely applied to numerous applications domains such as fraud detection, network intrusion, performance analysis, weather prediction, etc [9].

Recently, the topic of outlier detection in WSNs has attracted much attention. According to potential sources of outliers as mentioned earlier, the identification of outliers provides data reliability, event reporting, and secure functioning of the network. Specifically, outlier detection controls the quality of measured data, improves robustness of the data analysis under the presence of noise and faulty sensors so that the communication overhead of erroneous data is reduced and the aggregated results are prevented to be affected. Outlier detection also provides an efficient way to search for values that do not follow the normal pattern of sensor data in the network. The detected values consequently are treated as events indicating change of phenomenon that are of interest. Furthermore, outlier detection identifies malicious sensors that always generate outlier values, detects potential network attacks by adversaries, and further ensures the security of the network. Here, we exemplify the essence of outlier detection in several real-life applications.

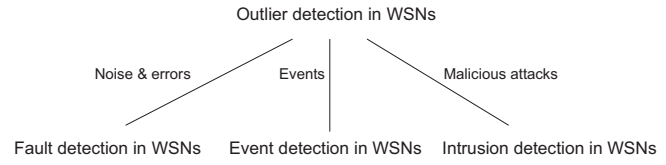


Fig. 1. Three outlier sources in WSNs and their corresponding detection techniques

- *Environmental monitoring*, in which sensors such as temperature and humidity are deployed in harsh and unattended regions to monitor the natural environment. Outlier detection can identify when and where an event occurs and trigger an alarm upon detection.
- *Habitat monitoring*, in which endangered species can be equipped with small non-intrusive sensors to monitor their behavior. Outlier detection can indicate abnormal behaviors of the species and provide a closer observation about behavior of individuals and groups.
- *Health and medical monitoring*, in which patients are equipped with small sensors on multiple different positions of their body to monitor their well-being. Outlier detection showing unusual records can indicate whether the patient has potential diseases and allow doctors to take effective medical care.
- *Industrial monitoring*, in which machines are equipped with temperature, pressure, or vibration amplitude sensors to monitor their operation. Outlier detection can quickly identify anomalous readings to indicate possible malfunction or any other abnormality in the machines and allow for their corrections.
- *Target tracking*, in which sensors are embedded in moving targets to track them in real-time. Outlier detection can filter erroneous information to improve the estimation of the location of targets and also to make tracking more efficiently and accurately.
- *Surveillance monitoring*, in which multiple sensitive and unobtrusive sensors are deployed in restricted areas. Outlier detection identifying the position of the source of the anomaly can prevent unauthorized access and potential attacks by adversaries in order to enhance the security of these areas.

It should be noted that several research topics have been developed for identifying specific sources of outliers occurred in WSNs. As illustrated in Figure 1, these topics include fault detection ([12], [13]), event detection ([4], [14], [15]) and intrusion detection ([16], [17]).

C. Outlier Detection in Event Detection Domain

Related work in outlier detection has also been found in *event detection* domain of WSNs. These event-based applications require sensor nodes to report event to the sink node in a timely manner once an event is detected. Event detection techniques are different than data-driven and query-driven

techniques, where nodes regularly report sensor readings to the sink node or respond to queries periodically issued by the sink node. A complex event combing two or more atomic events requires multiple types of sensors collaborating to detect an event [18]. Martincic and Schwiebert [4] employ a cell-based network architecture to locally detect events based on collaboration among neighboring nodes. Krishnamachari and Iyengar [15] propose a distributed Bayesian protocol to detect event regions in presence of faulty sensors. Ding et al. [14] attempt to identify event boundaries since detection of event boundary may become more important than detection of event region because of unreliability of sensor measurements.

The main differences between event detection and outlier detection are included as:

- outlier detection techniques have no a priori knowledge of trigger condition or semantic of any event, while event detection techniques hold the trigger condition or semantic of certain event issued by the sink node.
- outlier detection aims at identifying anomalous readings by comparing sensor measurements with each other, while event detection aims at specifying a certain event by comparing sensor measurements with the trigger condition or pre-defined pattern.
- outlier detection techniques need to prevent normal data to be classified as outlier and thus keeping the detection rate high and false alarm rate low, while event detection techniques need to prevent erroneous data which conform to the event condition or pattern to influence reliability of the detection.

On the other hand, the common characteristic of outlier detection and event detection is that they employ *spatio-temporal correlations* among sensor data of neighboring nodes to distinguish between events and errors. This is based on the fact that noisy measurements and sensor faults are likely to be stochastically unrelated, while event measurements are likely to be spatially correlated [15].

Due to the fact that not all outliers have to be identified in event detection applications, outlier detection techniques have not really been used in the literature of event detection domain, although they may be suitable. In this paper, we focus on addressing outlier detection in WSNs, excluding the discussion on the detections of specific outlier sources and events.

D. Challenges of Outlier Detection in WSNs

Extracting useful knowledge from raw sensor data is not a trivial task [19]. The context of sensor networks and the nature of sensor data make design of an appropriate outlier detection technique more challenging. According to the following reasons, conventional outlier detection techniques might not be suitable for handing sensor data in WSNs.

- *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. They are usually computationally expensive and require much memory for data analysis and storage. Thus, a challenge for outlier detection in WSNs is how to minimize the energy consumption while using a reasonable amount of memory for storage and computational tasks.
- *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 [20]. Most of traditional outlier detection techniques using centralized approach for data analysis cause too much energy consumption and communication overhead. Thus, a challenge for outlier detection in WSNs is how to minimize the communication overhead in order to relieve the network traffic and prolong the lifetime of the network.
- *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. Furthermore, direct computation of probabilities is difficult [21]. Most of traditional outlier detection techniques that analyze data in an offline manner do not meet the requirement of handling distributed stream data. The techniques based on the a priori knowledge of the data distribution also cannot be suitable for sensor data. Thus, a challenge for outlier detection in WSNs is how to process distributed streaming data online.
- *Dynamic network topology, frequent communication failures, mobility and heterogeneity of nodes*. A sensor network deployed in unattended environments over extended period of time is susceptible to dynamic network topology and frequent communication failures. Moreover, sensor nodes may move among different locations at any point in time, and may have different sensing and processing capacities. Each sensor node may even be equipped with different number and types of sensors. Such dynamicity and heterogeneity increase the complexity of designing an appropriate outlier detection technique for WSNs.
- *Large-scale deployment*. Deployed sensor networks can have massive size (up to hundreds or even thousands of sensor nodes). The key challenge of traditional outlier detection techniques is to maintain a high detection rate while keeping the false alarm rate low. This requires the construction of an accurate normal profile that represents the normal behavior of sensor data [19]. This is a very difficult task for large-scale sensor network applications. Also, traditional outlier detection techniques do not scale well to process large amount of distributed data streams in an online manner.
- *Identifying outlier sources*. The sensor network is expected to provide the raw data sensed from the physical world and also detect events occurred in the network. However, it is difficult to identify what has caused an outlier in sensor data due to the resource constraints and dynamic nature of WSNs. Traditional outlier detection technique often do not distinguish between errors and events and regard outlier as errors, which results in loss of important hidden information about events. Thus, a challenge of outlier detection in WSNs is how to

identify outlier sources and make distinction between errors, events and malicious attacks.

Thus, the main challenge faced by outlier detection techniques for WSNs is to satisfy the mining accuracy requirements while maintaining the resource consumption of WSNs to a minimum [21]. In other words, the main question is how to process as much data as possible in a decentralized and online fashion while keeping the communication overhead, memory and computational cost low [1].

III. Classification Criteria of Outlier Detection Techniques for WSNs

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 in Section VI.

A. 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 [21]. The type of input data determines which outlier detection techniques can be used to analyze the data. Outlier detection techniques usually consider the two following aspects of sensor data.

1) Attributes: A data measurement can be identified as outlier when its attributes have anomalous values [10]. 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. However, each sensor node may be equipped with multiple sensors and also certain correlations may exist among attributes of sensor data. Thus, outlier detection techniques for WSNs should be able to analyze *multivariate data* and identify whether the attributes together display anomaly. This is simply because sometimes none of the attributes individually may have an anomalous value [22]. Analysis of multivariate data, on the one hand, improves the accuracy of outlier detection techniques, while on the other hand increases computational complexity.

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 [23]. Attributes of multivariate sensor data may induce certain correlation, e.g., the readings of humidity and barometric pressure sensors are related to the readings of the temperature sensors. Capturing the attribute correlations helps to improve the mining accuracy and computational efficiency. On the other hand, sensor data tends to be correlated in both time and space, especially for those data collected from environmental monitoring applications [24]. Existence of *temporal correlation* implies that the readings observed at one time instant are related to the readings observed at the previous time instants, while existence of *spatial correlation* implies that the readings from sensor nodes geographically close to each other are expected to be largely correlated [25]. Capturing the spatio-temporal correlations helps to predict the trend of sensor readings and also to distinguish between errors and events.

B. Type of Outliers

Compared to a centralized approach, in which the entire data is processed in a central place, outliers in WSNs can be analyzed and identified at 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 [2]. Depending on the scope of data used for outlier detection, outlier may be either *local* or *global*.

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. Two variations for local outlier identification exist in WSNs. One is that each node identifies the anomalous values only depending on its historical values. The alternative is that in addition to its own historical readings, each sensor node collects readings of its neighboring nodes to collaboratively identify the anomalous values. Compared with the first approach, the second approach takes advantage of the spatio-temporal correlations among sensor data and improves the accuracy and robustness of outlier detection.

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. Depending on the network architecture, the identification of global outliers can be performed at different levels in the network [26]. In a centralized architecture, all data is transmitted to the sink node for identifying outliers. This mechanism consumes much communication overhead and delays the response time. In aggregate/clustering-based architecture, the aggregator/clusterhead collects the data from nodes within its controlling range and then identifies outliers. While this mechanism optimizes response time and energy consumption, it has the same problem as of centralized approach if the aggregator/clusterhead has a large number of nodes under its supervision. In addition, it should be mentioned that individual nodes can identify global outliers if they have a copy of global estimator model obtained from the sink node [2].

C. Identity of Outliers

There are three sources of outliers occurred in WSNs: (1) noise and 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.

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 [4]. Erroneous data is normally represented as an arbitrary change and is extremely different from the rest of the data. Since such errors influence data quality, they need to be identified and corrected if possible as data after correction may still be usable for data analysis. Only when the outliers are too erroneous to

correct, they are discarded in order to save transmission power and energy consumption.

2) Events: An event is defined as a particular phenomena that changes the real-world state, e.g., forest fire, chemical spill, 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 [27]. Thus, outlier detection techniques need to make use of data of neighboring nodes and spatial similarity of the sensor data. This is based on the fact that the sensor faults are likely to be spatially unrelated, while event measurements are likely to be spatially correlated [15].

D. Degree of Being an Outlier

Outlier detection techniques not only identify data that does not conform with normal pattern of sensor data, but also provide specific methods to compute the degree of which data measurements deviate from the normal pattern of sensor data. In WSNs, outliers are measured in two scales, i.e., *scalar* and *outlier score* [9].

1) Scalar: The scalar scale is a zero-one classification measure, which classifies each data measurement into normal or outlier class. Thus, the output of techniques of scalar scale, which neither do differentiate between different outliers nor provide a ranked list of outliers, is a set of outliers and a set of normal measurements.

2) Outlier Score: Techniques of the outlier score scale assign an outlier score to each data measurements depending on the degree of which the measurement is considered as an outlier and provide a ranked list of outliers. An analyst may choose to either analyze top *n* outliers having the largest outlier scores or use a cut-off threshold to select the outliers. Such threshold is often not easy to choose and is usually user-specified and fixed. The optimal solution in WSNs is to learn the threshold and to constantly modify it with updates of arrived streaming data.

E. Availability of Pre-Defined Data

A straightforward solution to identify outliers is to construct a profile of normal pattern of the data and then use the normal profile to detect outliers. The observations whose characteristics differ significantly from the normal profile are declared as outliers [19]. Based on their assumption on availability of pre-defined data, outlier detection techniques can be classified into three basic categories, namely, *supervised*, *unsupervised* and *semi-supervised* learning approaches [10].

Both supervised and semi-supervised approaches require pre-classified normal or abnormal data to characterize all anomalies or non-anomalies in the training phase. The test data is compared against the learned predictive model for normal or abnormal classes. One should note that the pre-classified data is neither always available nor easy to obtain in many real-life WSNs applications and also new types of normal or abnormal data may not be included in the pre-labelled data. On the contrary, unsupervised approaches require no pre-labelled data, but they use certain measure criteria to identify outliers. For example, in the *distance-based* approaches, the normal profile refers to the average distance between every data measurements to its corresponding k^{th} closest neighbor. If the distance from a given data measurement to its k^{th} closest neighbor is significantly bigger than the average, then the data measurement is considered as an outlier [19]. Compared to supervised and semi-supervised approaches, unsupervised approaches are more applicable to WSNs.

IV. Taxonomy Framework for Outlier Detection Techniques Designed for WSNs

Related work on taxonomy framework for outlier detection techniques for general data has been addressed in various literature. Markou and Singh [28] and [29] present an extensive review of novelty detection techniques based in statistical and neural network fields. Hodge and Austin [5] address outlier detection methodologies from perspective of three fields of computing, i.e., statistics, neural networks and machine learning. Chandola et al. [9] classify anomaly detection techniques in terms of various application domains and several knowledge disciplines. Zhang et al. [8] provide a taxonomy for outlier detection techniques with respect to multiple type of data sets. Although there may be some overlaps between these taxonomies and the one presented here, existing taxonomies are not directly applicable to WSNs due to the nature of sensor data and specific requirements and limitations of WSNs. Additionally, recently, many outlier detection techniques specifically developed for WSNs have emerged. This calls for a taxonomy addressing techniques and requirements of WSNs specifically. In this section, we provide a technique-based taxonomy framework to categorize these techniques designed for WSNs.

As illustrated in Figure 2, outlier detection techniques for WSNs can be categorized into *statistical-based*, *nearest neighbor-based*, *clustering-based*, *classification-based*, and *spectral decomposition-based* approaches. Statistical-based approaches are further categorized into *parametric* and *non-parametric* approaches based on how the probability distribution model is built [28]. *Gaussian-based* and *non-Gaussian-based* approaches belong to parametric approaches, and *kernel-based* and *histogram-based* approaches belong to non-parametric approaches. Classification-based approaches are categorized as *Bayesian network-based* and *support vector machine-based* approaches based on type of classification model that they use. Bayesian network-based approaches are further categorized into *naive Bayesian network*, *Bayesian belief network*, and *dynamic Bayesian network* based on the degree of probabilistic independencies among variables. Spectral decomposition-based approaches use *principle component analysis* for outlier detection.

V. Outlier Detection Techniques for WSNs

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 IV. Furthermore, we provide a brief evaluation for each of these disciplines.

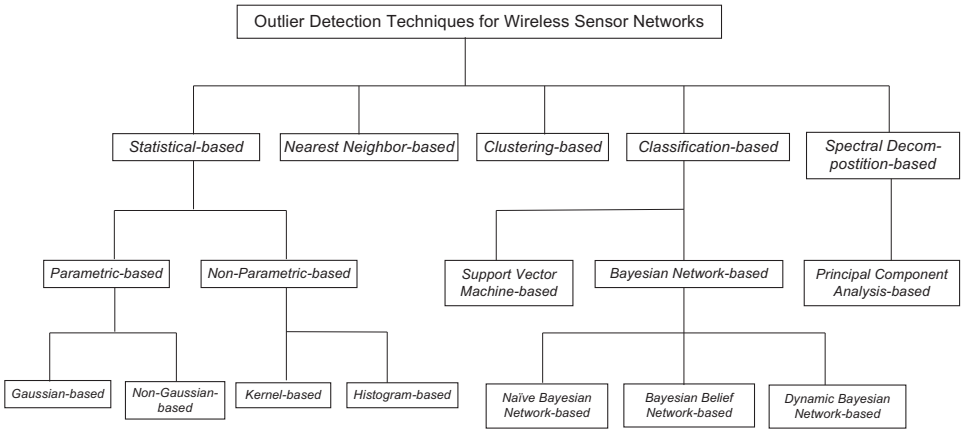


Fig. 2. Taxonomy of outlier detection techniques for WSNs

A. 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. A data instance is declared as an outlier if the probability of the data instance to be generated by this model is very low. The modelling techniques can work in an unsupervised mode, where a statistical model can be determined if it fits majority of the observations while small amounts of outliers exist in the data. The statistical-based approaches are categorized into parametric and non-parametric based on how the probability distribution model is built.

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.

- *Gaussian-based models*. Wu et al. [30] 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. A node is considered as an outlying node if the absolute value of its reading's deviation degree is sufficiently larger than a pre-selected threshold. The technique of event boundary detection is based on the previous results of outlying sensor identification and determines a node as an event node if the absolute value of the node's deviation degree in one geographical region is much larger than that in another region. Accuracy of these outlier detection techniques is not relatively high due to the fact that they ignore the temporal correlation of sensor readings.

  Bettencourt et al. [31] present a local outlier detection technique to identify errors and detect events in ecological applications of WSNs. This technique can distinguish between erroneous measurements and events by using the spatio-temporal correlations of sensor data. Each node learns the statistical distribution of difference between its own measurements and each of its neighboring nodes, as well as between its current and previous measurements. The procedure can be based on a priori knowledge of data distribution or a non-parametric density estimation. A measurement is identified as anomalous if its value in the statistical significance test is less than a user-specified threshold. The detected anomalous measurement may be considered as event if it is likely to be temporally different from its previous measurements but spatially correlated. The drawback of this technique is that it relies on the choice of the appropriate values of the threshold.

  Hida et al. [32] design a local technique to make simple aggregation operations, such as MAX or AVG, more reliable under presence of faulty sensor readings and failed nodes. This technique relies on the spatio-temporal correlations of sensor data and uses two statistical tests to locally detect outliers. Each incoming sensor value is compared against the current value and the previous values of all nodes in the neighborhood. If the incoming value passes the two statistical tests, it is allowed to be aggregated as usual; otherwise (if the incoming value is outside of 2.5 standard deviations of the mean) it is declared as an outlier and will be eliminated from the analysis. Drawbacks of this technique include the fact that it only deals with one-dimensional outlier data and too much memory is required for a node to store historical values of all its neighboring nodes.
- *Non-Gaussian-based models*. Jun et al. [33] 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. Each node in a cluster first detects and corrects temporal outliers by comparing the predicted data and the sensing data. Then the clusterhead collects the rectified data from all other nodes in the cluster and further detects spatial outliers that deviate remarkably from other normal data. This technique reduces the communication cost due to local transmission and also reduces computational cost as the cluster-heads carry out most of the computation tasks. However, the SαS distribution may not be suitable for real sensor data and the cluster-based structure may be susceptible to dynamic changes of network topology.

2) Non-Parametric-Based Approaches: 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. Kernel density estimators use kernel functions to estimate the probability distribution function (pdf) for the normal instances. A new instance that lies in the low probability area of this pdf is declared as an outlier.

- *Histogramming*. Sheng et al. [34] 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. The sink uses histogram information to extract data distribution from the network and filters out the non-outliers. Outliers can be identified by recollecting more histogram information from the network. The identification of outliers is achieved by a fixed threshold distance or the rank among all outliers. Drawbacks of this technique include the fact that re-collecting more histogram information from the whole network will cause too much communication overhead and the technique only considers one-dimensional data.
- *Kernel functions*. Palpanas et al. [35] 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. Thus, each node can locally identify outliers if the values deviate significantly from the model of approximated data distribution. A value is considered as an outlier if the number of values being in its neighborhood is less than a user-specified threshold. This technique can also be extended to high-level nodes for identification of outlier in a more global perspective. The main problem of this technique is its high dependency on the defined threshold, while choice of an appropriate threshold is quite difficult and a single threshold may also not be suitable for outlier detection in multi-dimensional data. Furthermore, the technique does not consider maintaining the model while sensor data is frequently updated.

  Subramaniam et al. [2] further extend the work of Palpanas et al. [35] and solve the two previous problems of insufficiency of a single threshold for multi-dimensional data and maintaining the data model built by kernel density estimator. They propose two global outlier detection techniques for complex applications. One technique allows each node to locally identify outliers using the same technique as Palpanas et al. [35] and then transmit the outliers to its corresponding parent to be checked until the sink eventually determines all global outliers. In the other technique, each node employs more robust technique called LOCI [36] to locally detect global outliers by having a copy of global estimator model obtained from the sink. Experimental results show that these techniques achieve high accuracy in terms of estimating data distribution and high detection rate while consuming low memory usage and message transmission. A remaining problem with this technique is its inability to detect spatial outliers due to the fact that it does not consider the spatial correlations among neighboring sensor data.

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. Thus parametric approaches may be useless if sensor data does not follow the preset distribution. Non-parametric techniques are appealing due to the fact that they do not make any assumption about the distribution characteristics. Histogramming models are very efficient for univariate data but are not able to capture the interactions between different attributes of multivariate data. Also, it is not easy to determine an optimal size of the bins to construct the histogram. Kernel functions can scale well in multivariate data and are computationally cheap.

B. Nearest Neighbor-Based Approaches

Nearest neighbor-based approaches are the most commonly used approaches to analyze a data instance with respect to its nearest neighbors in the data mining and machine learning community. They use several well-defined distance notions to compute the distance (similarity measure) between two data instance ([37], [38]). A data instance is declared as an outlier if it is located far from its neighbors. Euclidean distance is a popular choice for univariate and multivariate continuous attributes.

Branch et al. [39] propose a technique based on distance similarity to identify global outliers in sensor networks. This technique attempts to reduce the communication overhead by a set of representative data exchanges among neighboring nodes. Each node uses distance similarity to locally identify outliers and then broadcasts the outliers to neighboring nodes for verification. The neighboring nodes repeat the procedure until all of the sensor nodes in the network eventually agree on

the global outliers. This technique can be flexible in respect to multiple existing distance-based outlier detection techniques. However, the technique does not adopt any network structure so that every node uses broadcast to communicate with other nodes in the network, which will cause too much communication overhead. Consequently, it does not scale well to the large-scale networks.

Zhang et al. [40] propose a distance-based technique to identify n global outliers in snapshot and continuous query processing applications of sensor networks. This technique reduces communication overhead as it adopts the structure of aggregation tree and prevents broadcasting of each node in the network [39]. Each node in the tree transmits some useful data to its parent after collecting all the data sent from its children. The sink then roughly figures out top n global outliers and floods these outliers to all the nodes in the network for verification. If any node disagrees on the global results, it will send extra data to the sink again for outlier detection. This procedure is repeated until all the nodes in the network agree on the global results calculated by the sink. This technique considers only one-dimensional data and the aggregation tree used may not be stable due to the dynamic changes of network topology.

Zhuang et al. [27] present two in-network outlier cleaning techniques for data collection applications of sensor networks. One technique uses wavelet analysis specifically for outliers such as noises or occasionally appeared errors. The other technique uses dynamic time warping (DTW) distance-based similarity comparison specifically for outliers that are erroneous and last for a certain time period. In this technique, each node transforms raw data into the wavelet time-frequency domain and identifies the high-frequency data measurements as outliers and corrects them using proper wavelet coefficients. The long segmental outliers can be detected and removed by comparing the similarity of two sensing series of the neighboring nodes within two forwarding hops. The proposed techniques take advantage of spatio-temporal correlations of sensor data for identifying outliers. A drawback of this technique, however, is its dependency of a suitable pre-defined threshold that is not obvious to define.

1) Evaluation of Nearest Neighbor-based Techniques: Nearest neighbor-based approaches do not make any assumption about data distribution and can generalize many notions from statistical-based approaches. However, these techniques suffer from the choice of the appropriate input parameters. Additionally, in multivariate data sets it is computationally expensive to compute the distance between data instances and as a result these technique lack scalability.

C. Clustering-Based Approaches

Clustering-based approaches are popular approaches within the data mining community to group similar data instances into clusters with similar behavior. Data instances are identified as outliers if they do not belong to clusters or if their clusters are significantly smaller than other clusters. Euclidean distance is often used as the dissimilarity measure between two data instances.

Rajasegarar et al. [41] propose a global outlier detection technique based on clustering technique to identify anomalous measurements in sensor nodes. This technique minimizes the communication overhead by clustering the sensor measurements and merging clusters before communicating with other nodes. Initially, each node clusters the measurements and reports cluster summaries rather than transmitting the raw sensor measurements to its parent. The parent then merges cluster summaries collected from all of its children before sending them to the sink. An anomalous cluster can be determined in the sink if the cluster's average inter-cluster distance is larger than one threshold value of the set of inter-cluster distances. Determining the parameter k (the k nearest neighbor clusters), which is used to compute the average inter-cluster distance is not always easy. The parameter of cluster width may also not be defined appropriately.

1) Evaluation of Clustering-Based Techniques: Clustering-based approaches do not require a priori knowledge of the data distribution and are capable of being used in an incremental model, i.e., new data instance can be fed into the system and being tested to find outliers. However, these techniques suffer from the choice of an appropriate parameter of cluster width. Additionally, computing the distance between data instances in multivariate data is computationally expensive.

D. Classification-Based Approaches

Classification approaches are important systematic approaches in the data mining and machine learning community. They learn a classification model 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 labelled training data and learn the classification model which fits the majority of the data instance during training. The one-class unsupervised techniques learn the boundary around the normal instances while some anomalous instance may exist and declare any new instance falling outside this boundary as an outlier. The classifier may need to update itself to accommodate the new instance that belong to the normal class. In existing outlier detection techniques for WSNs, classification-based approaches are categorized into support vector machines (SVM)-based and Bayesian network-based approaches based on type of classification model they use.

1) Support Vector Machine-Based Approaches: SVM techniques separate the data belonging to different classes by fitting a hyperplane between them which maximizes the separation. The data is mapped into a higher dimensional feature space where it can be easily separated by a hyperplane. Furthermore, a kernel function is used to approximate the dot products between the mapped vectors in the feature space to find the hyperplane.

Rajasegarar et al. [42] propose a SVM-based technique for outlier detection in sensor data. This technique uses one-class quarter-sphere SVM to reduce the effort of computational complexity and locally identify outliers at each node. The sensor data that lies outside the quarter sphere is considered as an outlier. Each node communicates only summary information (the radius information of sphere) with its parent for global outlier classification. This technique identifies outliers from the data measurements collected after a long time window

and is not performed in real-time. The technique also ignores spatial correlation of neighboring nodes, which makes the results of local outliers inaccurate.

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. 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 networks techniques capture spatio-temporal correlations among sensor nodes. Bayesian belief network techniques consider the correlations among the attributes of the sensor data. Dynamic Bayesian networks techniques consider the dynamic network topology that evolves over time, adding new state variables to represent the system state at the current time instance.

- *Naive Bayesian Network models*. Elnahrawy and Nath [24] 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. Each node locally computes the probabilities of each of its incoming readings being in all subintervals (classes) divided from the whole values interval. If the probability of a sensed reading in its class is smaller than that of being in other classes, it is considered as an outlier. The technique requires no user-specified threshold to determine outliers and can also be used to approximate the missing readings occurred in the network. It, however, does not specify how to decide a specific spatial neighborhood under the dynamic change of network topology. Also, it only deals with one-dimensional data.
- *Bayesian Belief Network models*. Janakiram et al. [23] present a technique based on Bayesian belief network (BBN) to identify local outliers in streaming sensor data. This technique uses BBN to capture not only the spatio-temporal correlations that exist among the observations of sensor nodes but also conditional dependence among the observations of sensor attributes. Each node trains a BBN to detect outliers based on behaviors of its neighbors' readings as well as its own reading. An observation is considered as outlier if it falls beyond the range of the expected class. Compared to naive Bayesian networks, this technique improves the accuracy in detecting outliers as it considers conditional dependencies among the attributes. Accuracy of a BBN depends on how the conditional dependence among the observations of sensor attributes exists. This technique may not work well in presence of the dynamic network topology change.
- *Dynamic Bayesian Network models*. Hill et al. [43] present two techniques based on dynamic Bayesian networks (DBNs) to identity local outliers in environmental sensor data streams. This technique uses DBNs to fast track changes in dynamic network topology of sensor networks. One technique assumes that there is only a measured state variable existing in the multivariate data and the current state can be determined only depending on its historical state. This technique identifies outliers by computing the posterior probability of the most recent data values in a sliding window. The data measurements that fall outside the expected value interval are considered as outliers. The other technique uses a more complex DBN including two measured state variables for outlier detection. This technique makes it possible to operate on several data streams at once.

3) Evaluation of Classification-based Techniques: Classification-based approaches provide an exact set of outliers by building a classification model to classify. 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.

E. Spectral Decomposition-Based Approaches

Spectral decomposition-based approaches aim at finding normal modes of behavior in the data by using principle components. Principal component analysis (PCA) is a technique that is used to reduce dimensionality before outlier detection and finds a new subset of dimension which capture the behavior of the data. Specifically, the top few principal components capture the build of variability and any data instance that violates this structure for the smallest components is considered as an outlier.

Chatzigiannakis et al. [26] propose a PCA-based technique to solve data integrity and accuracy problem caused by compromised or malfunctioning sensor nodes. This technique uses PCA to efficiently model the spatio-temporal data correlations in a distributed manner and identifies local outliers spanning through neighboring nodes. Each primary node offline builds a model of the normal condition by selecting appropriate principal components (PCs) and then obtains sensor readings from other nodes in its group and performs local real-time analysis. The readings that significantly vary from the modelled variation value under normal condition are declared as outliers. The primary nodes eventually forward the information about outlier data to the sink. The offline procedure of selecting appropriate PCs is computationally very expensive.

1) Evaluation of Spectral Decomposition-Based Techniques: Principal component analysis-based approaches tend to capture the normal pattern of the data using the subset of dimensions and can be applied to high-dimensional data. However, selecting suitable principle components, which is needed to accurately estimate the correlation matrix of normal patterns, is computationally very expensive.

VI. Comparative Table for Outlier Detection Techniques for WSNs

In this section, we first discuss evaluation principles for outlier detection techniques and then present a comparative table to compare existing outlier detection techniques for WSNs using the taxonomy framework proposed in Section IV.

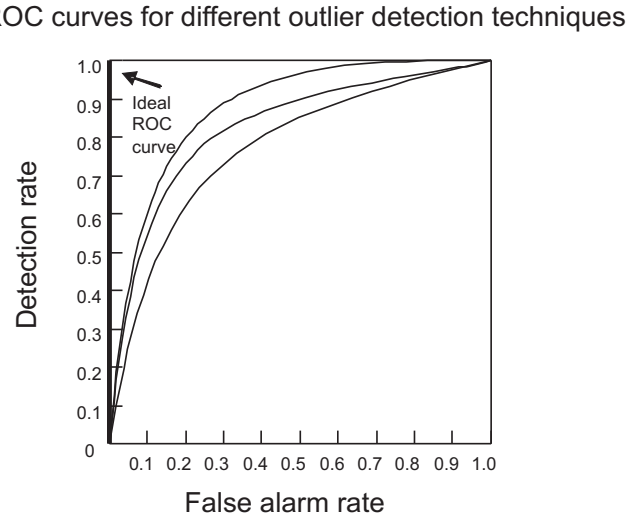


Fig. 3. ROC curves for different detection techniques [44].

We also specify shortcomings of current techniques and further highlight the required characteristics of an optimal outlier detection technique for WSNs and important research areas to focus on.

A. Evaluation of Outlier Detection Techniques

Evaluation of an outlier detection technique for WSNs depends on whether it can satisfy the mining accuracy requirements while maintaining the resource consumptions of WSNs to a minimum [21]. Outlier detection techniques are required to maintain a high *detection rate* while keeping the *false alarm rate* low. The detection rate represents the percentage of anomalous data that are correctly considered as outliers, and the false alarm rate, also known as false positive rate (FPR), represents the percentage of normal data that are incorrectly considered as outliers. A receiver operating characteristic (ROC) curves [44] usually is used to represent the trade-off between the detection rate and false alarm rate. The larger the area under the ROC curve, the better the performance of the corresponding technique. Figure 3 illustrates an example of ROC curves.

B. Shortcomings of Outlier Detection Techniques for WSNs

Table I shows characteristics of outlier detection techniques developed specially for WSNs. From the table, we realize that the existing outlier detection techniques have the following shortcomings.

- Majority of existing work do not take into account multivariate data and assume the sensor data is univariate. They ignore the fact that the attributes together can display anomaly while in some cases none of the attributes individually has an anomalous value.
- Many of techniques only consider the spatio-temporal correlations among sensor data of neighboring nodes and ignore the dependencies among the attributes of the sensor node. This in turn increases the computational complexity and reduces the accuracy of outlier detection.
- Existing techniques considering spatial correlation among sensor data of neighboring nodes suffer from the choice of appropriate neighborhood range. Techniques considering temporal correlation among sensor data suffer from the choice of the size of the sliding window.
- Little work has been done on distinguishing between events and errors. Many of existing techniques simply regard outliers as errors. Since a commonly accepted notion is that errors should be removed from the data set, important information about hidden events may be lost. In fact these techniques do not explicitly state how to deal with the identified outliers and end after identification of outliers.
- Many of these techniques use a user-specified threshold to determine outliers. However, an appropriate threshold is not easy to determine. In addition, assuming fixed thresholds is not proper considering dynamic change of WSNs characteristics.
- These techniques assume that sensor nodes are static and do not consider nodes mobility. Applying them for mobile networks or in presence of dynamic change of network topology would be challenging.

C. Requirements for outlier detection in WSNs

Having seen these shortcomings and special characteristics of WSNs, it is clear that an outlier detection technique specifically designed for WSN is required, 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. To summarize, we highlight the requirements which an optimal outlier detection approach for WSNs should meet:

- It must distributively process the data to prevent unnecessary communication overhead and energy consumption and to prolong network lifetime.
- It must be an online technique to be able to handle streaming or dynamically updated sensor data.
- It must have a high detection rate while keeping a false alarm rate low.
- It should be unsupervised as in WSN the pre-classified normal or abnormal data is difficult to obtain. Also, it should be non-parametric as no a priori knowledge about the input sensor data distribution may be available.
- It should take multivariate data into account.
- It must be simple, have low computational complexity, and be easy to implement in presence of limited resources.
- It must enable auto-configurability with respect to dynamic network topology or communication failure.
- It must scale well.
- It must consider dependencies among the attributes of the sensor data as well as spatio-temporal correlations that exist among the observations of neighboring sensor nodes.
- It must effectively distinguish between erroneous measurements and events.

TABLE I
CLASSIFICATION AND COMPARISON OF GENERAL OUTLIER DETECTION TECHNIQUES FOR WSNS

Techniques	Sensor data					Outlier type					Outlier identity	Outlier degree		
	Attribute		Correlation			Local		Global			Error/ Event	Scalar	Outlier score	
	Univariate	Mulvariate	Attribute	Spatial	Temporal	Individual	Collaboration	Individual	Aggregate	Centralized			Fixed	Flexible
Wu et al. [30]	√			√			√				√		√	
Bettencourt et al. [31]]	√			√	√		√				√		√	
Hida et al. [32]	√			√	√		√						√	
Jun et al. [33]	√			√	√	√			√					√
Sheng et al. [34]	√				√					√			√	
Palpanas et al. [35]	√				√	√			√				√	
Subramaniam et al. [2]		√			√			√	√				√	√
Branch et al. [39]	√				√								√	
Zhang et al. [40]	√				√				√				√	
Zhuang et al. [27]	√			√	√	√			√				√	
Rajasegarar et al. [41]		√			√					√				√
Rajasegarar et al. [42]		√			√	√		√					√	
Elnahrawy and Nath [24]	√			√	√		√					√		
Janakiram et al. [23]]		√	√	√	√		√					√		
Hill et al. [43]		√		√	√		√					√		
Chatzigiannakis et al. [26]		√		√	√	√								

D. Important Research Areas to Focus on

There are several important research areas related to outlier detection in WSNs, which need extra investigation. The list includes:

- Investigating applicability of Artificial Intelligence (AI) techniques for outlier detection in WSNs.
- Adaptability support.
- Investigating spatial and spatio-temporal correlations which may exist between adjacent nodes.
- Defining semantics for outliers to be able to distinguish between errors and events.
- Combining offline learning mechanisms with distributed and online outlier detection.

VII. 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 comparative table 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.

ACKNOWLEDGMENT

This work is supported by the EU's Seventh Framework Programme and the SENSEI project.

REFERENCES

[1] X. Ma, D. Yang, S. Tang, Q. Luo, D. Zhang, and S. Li, Online Mining in Sensor Networks, *IFIP international conference on network and parallel computing*, Vol. 3222, pp. 544-550, 2004.

[2] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogerakiand, and D. Gunopulos, Online Outlier Detection in Sensor Data using Non-parametric Models, *J. Very Large Data Bases*, VLDB 2006.

[3] A. Perrig, J. Stankovic, and D. Wagner, Security in Wireless Sensor Networks, *CACM*, Vol. 47, No. 6, pp. 53-57, 2004.

[4] F. Martincic and L. Schwiebert, Distributed Event Detection in Sensor Networks, *Proc. International Conference on Systems and Networks Communication*, pp. 43-48, 2006.

[5] V. Hodge and J. Austin, A Survey of Outlier Detection Methodologies, *Artificial Intelligence Review*, Vol. 22, pp. 85-126, 2003.

[6] D.M. Hawkins, *Identification of Outliers*, London: Chapman and Hall, 1980.

[7] V. Barnett and T. Lewis, *Outliers in Statistical Data*, New York: John Wiley Sons, 1994.

[8] Y. Zhang, N. Meratnia, and P.J.M. Havinga, A Taxonomy Framework for Unsupervised Outlier Detection Techniques for Multi-Type Data Sets, *Technical Report*, University of Twente, 2007.

[9] V. Chandola, A. Banerjee, and V. Kumar, Anomaly Detection: A Survey,*Technical Report*, University of Minnesota, 2007.

[10] P.N. Tan, M. Steinback, and V. Kumar, *Introduction to Data Mining*, Addison Wesley, 2006.

[11] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2006.

[12] J. Chen, S. Kher, and A. Somani, Distributed Fault Detection of Wireless Sensor Networks, *Proc. 2006 workshop on dependability issues in wireless ad hoc networks and sensor networks*, pp. 65-72, 2006.

[13] X. Luo, M. Dong, and Y.Huang, On Distributed Fault-Tolerant Detection in Wireless Sensor Networks, *IEEE Trans. Comput.*, Vol. 55, No. 1, pp. 58-70, 2006.

[14] M. Ding, D. Chen, K. Xing, and X. Cheng, Localized Fault-Tolerant Event Boundary Detection in Sensor Networks, *Proc. IEEE Conference of Computer and Communications Societies*, pp. 902- 913, 2005.

[15] B. Krishnamachari and S. Iyengar, Distributed Bayesian Algorithms for Fault-Tolerant Event Region Detection in Wireless Sensor Networks, *IEEE Trans. Comput.*, Vol. 53, No. 3, pp. 241- 250, 2004.

[16] A.P.R. Silva, M.H.T. Martins, B.P.S. Rocha, A.A.F. Loureiro, L.B. Ruiz, and H.C. Wong, Decentralized Intrusion Detection in Wireless Sensor Networks, *Proc. 1st ACM international workshop on Quality of service & security in wireless and mobile network*s, pp. 16-23, 2005.

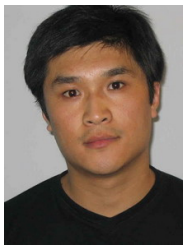
[17] V. Bhuse, and A. Gupta, Anomaly Intrusion Detection in Wireless Sensor Networks, *J. High Speed Networks*, Vol. 15, No. 1, pp. 33-51, 2006.

[18] M. Zoumboulakis and G. Roussos, Escalation: Complex Event Detection in Wireless Sensor Networks, *Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, pp. 270-285, 2007.

[19] P.N. Tan, Knowledge Discovery from Sensor Data, *Sensors*, 2006.

[20] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, Wireless Sensor Networks: A Survey, *J. Computer Networks*, Vol. 38, No. 4, pp. 393-422, March, 2002.

[21] M. M. Gaber, Data Stream Processing in Sensor Networks. In J. Gama and M. M. Gaber, *Learning from Data Streams Processing Techniques in Sensor Network*, pp. 41-48. Springer Berlin Heidelberg, 2007.
[22] P. Sun, Outlier Detection in High Dimensional, Spatial and Sequential Data Sets, *Doctoral dissertation*, University of Sydney, Sydney, 2006.
[23] D. Janakiram, A. Mallikarjuna, V. Reddy, and P. Kumar, Outlier Detection in Wireless Sensor Networks using Bayesian Belief Networks, *Proc. IEEE Comsware*, 2006.
[24] E. Elnahrawy and B. Nath, Context-Aware Sensors, *Proc. EWSN*, 2004.
[25] S.R. Jeffery, G. Alonso, M.J. Franklin, W. Hong, and J. Widom, Declarative Support for Sensor Data Cleaning, *International Conference on Pervasive Computing*, pp. 83-100, 2006.
[26] V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, and B. Maglariset, Hierarchical Anomaly Detection in Distributed Large-Scale Sensor Networks, *Proc. ISCC*, 2006.
[27] Y. Zhuang and L. Chen, In-Network Outlier Cleaning for Data Collection in Sensor Networks, *Proc. VLDB*, 2006.
[28] M. Markos, S. Singh, Novelty Detection: A Review-Part 1: Statistical Approaches. *J. Signal Processing*, Vol. 83, pp. 2481-2497, 2003.
[29] M. Markos, S. Singh, Novelty Detection: A Review-Part 2: Neural Network based Approaches. *J. Signal Processing*, Vol. 83, pp. 2499-2521, 2003.
[30] W. Wu, X. Cheng, M. Ding, K. Xing, F. Liu, and P. Deng, Localized Outlying and Boundary Data Detection in Sensor Networks, *IEEE Trans. Knowl. Data Eng.*, Vol. 19, No. 8, pp. 1145-1157, 2007.
[31] L.A. Bettencourt, A. Hagberg, and L. Larkey, Separating the Wheat from the Chaff: Practical Anomaly Detection Schemes in Ecological Applications of Distributed Sensor Networks, *Proc. IEEE International Conference on Distributed Computing in Sensor Systems*, 2007.
[32] Y. Hida, P. Huang, and R. Nishtala, Aggregation Query under Uncertainty in Sensor Networks, 2003.
[33] M.C. Jun, H. Jeong, and C.C.J. Kuo, Distributed Spatio-Temporal Outlier Detection in Sensor Networks, *Proc. SPIE*, 2006.
[34] B. Sheng, Q. Li, W. Mao, and W. Jin, Outlier Detection in Sensor Networks, *Proc. MobiHoc*, 2007.
[35] T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, Distributed Deviation Detection in Sensor Networks, ACM *Special Interest Group on Management of Data*, pp. 77-82, 2003.
[36] S. Papadimitriou, H. Kitagawa, P.B. Gibbons, and C. Faloutsos, LOCI: Fast Outlier Detection using the Local Correlation Integral, *International Conference on Data Engineering*, pp. 315-326, 2003.
[37] E. Knorr and R. Ng, Algorithms for Mining Distance-Based Outliers in Large Data Sets, *International Journal of Very Large Data Bases*, pp. 392-403, 1998.
[38] S. Ramaswamy, R. Rastogi, and K. Shim, Efficient Algorithms for Mining Outliers from Large Data Sets, *ACM Special Interest Group on Management of Data*, pp. 427-438, 2000.
[39] J. Branch, B. Szymanski, C. Giannella, and R. Wolff, In-Network Outlier Detection in Wireless Sensor Networks, *Proc. IEEE ICDCS*, 2006.
[40] K. Zhang, S. Shi, H. Gao, and J. Li, Unsupervised Outlier Detection in Sensor Networks using Aggregation Tree, *Proc. ADMA*, 2007.
[41] S. Rajasegarar, C. Leckie, M. Palaniswami, and J.C. Bezdek, Distributed Anomaly Detection in Wireless Sensor Networks, *Proc. IEEE ICCS*, 2006.
[42] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. C. Bezdek, Quarter Sphere Based Distributed Anomaly Detection in Wireless Sensor Networks, *Proc. IEEE International Conference on Communications*, pp. 3864-3869, 2007.
[43] D.J. Hill, B.S. Minsker, and E. Amir, Real-Time Bayesian Anomaly Detection for Environmental Sensor Data, *Proc. 32nd Congress of the International Association of Hydraulic Engineering and Research*, 2007.
[44] A. Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, and V. Kumar, A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection, *SIAM Conference on Data Mining*, 2003.



Yang Zhang is currently a PhD student in the pervasive system group at the University of Twente in the Netherlands. He received his B.S. and M.S. degrees in computer science and technology from the University of Jiangsu, China, in 2002 and 2004, respectively. He received his second M.S. degree in Telematics from the University of Twente, the Netherlands in 2006. His research interests include distributed data mining, outlier detection and event detection in sensor networks. He is involved as publicity co-chair and reviewer for conferences and workshops.

Nirvana Meratnia is an assistant professor in the pervasive system group at the university of Twente. She received her PhD in 2005 from the database group at the same university. Her research interests are in the area of distributed data management in wireless sensor networks, smart and collaborative objects, ambient intelligence, and context-aware applications. She is actively involved as program committee member and reviewer for conferences and workshops.

Paul Havinga is an associate professor in the pervasive system group at the University of Twente in the Netherlands. He received his PhD on the thesis entitled 'Mobile Multimedia Systems' in 2000 and was awarded with the 'DOW Dissertation Energy Award' for this work. His research interests include large-scale, heterogeneous wireless systems, sensor networks, energy-efficient architectures and protocols, ubiquitous computing, and personal communication systems. This research has resulted in over 180 scientific publications in journals and conferences. He is the editor of several journals and magazines and regularly serves as program committee chair, member, and reviewer for conferences and workshops.