Artif Intell Rev (2015) 43:193–228
DOI 10.1007/s10462-012-9370-y

Characteristics and classification of outlier detection techniques for wireless sensor networks in harsh environments: a survey

Nauman Shahid · Ijaz Haider Naqvi · Saad Bin Qaisar

Published online: 16 November 2012

Abstract Wireless sensor networks (WSNs) have received considerable attention for multiple types of applications. In particular, outlier detection in WSNs has been an area of vast interest. Outlier detection becomes even more important for the applications involving harsh environments, however, it has not received extensive treatment in the literature. The identification of outliers in WSNs can be used for filtration of false data, find faulty nodes and discover events of interest. This paper presents a survey of the essential characteristics for the analysis of outlier detection techniques in harsh environments. These characteristics include, input data type, spatio-temporal and attribute correlations, user specified thresholds, outlier types(local and global), type of approach(distributed/centralized), outlier identification(event or error), outlier degree, outlier score, susceptibility to dynamic topology, non-stationarity and inhomogeneity. Moreover, the prioritization of various characteristics has been discussed for outlier detection techniques in harsh environments. The paper also gives a brief overview of the classification strategies for outlier detection techniques in WSNs and discusses the feasibility of various types of techniques for WSNs deployed in harsh environments.

Keywords Wireless sensor networks · Harsh environments · Outlier detection · Event detection

N. Shahid (✉) · I. H. Naqvi
Department of Electrical Engineering, School of Science and Engineering,
Lahore University of Management Sciences, Sector U, DHA, Lahore Cantt 54792, Pakistan
e-mail: naumanshahid6@gmail.com

I. H. Naqvi
e-mail: ijaznaqvi@lums.edu.pk

S. B. Qaisar
NUST School of Electrical Engineering and Computer Science, Islamabad, Pakistan
e-mail: saad.qaisar@seecs.edu.pk

1 Introduction

A wireless sensor network (WSN) consists of a large number of sensor nodes distributed over a large area, with some powerful sink nodes which gather readings of sensor nodes. The sensor nodes are equipped with a magnitude of capabilities like sensing, processing and wireless communication. Each node is equipped with a wireless radio transceiver, a small microcontroller, a power source and many types of sensors such as temperature, humidity, light, heat, pressure, sound, vibration, etc. Over time, WSNs have been used for a multitude of applications. Extensive work has been dedicated for various applications of WSNs (Zhang et al. 2007a; Garca-Hernndez et al. 2004; Mainwaring et al. 2002; Cardell-Olivera et al. 2005; Ni et al. 2003; Akyildiz et al. 2002; George 2010).

A wide variety of applications of WSNs are related to personal, industrial, business and military domains, such as environmental monitoring, object tracking, health monitoring, battlefield observation, industrial safety and control, etc. (Garca-Hernndez et al. 2004; Mainwaring et al. 2002; Cardell-Olivera et al. 2005; Ni et al. 2003; Akyildiz et al. 2002; Dereszynski and Dietterich 2011; Bahrepour et al. 2010c,b; Phua et al. 2010d; George 2010). For instance, in *Event Detection and Reporting Applications*, the networks usually wait for an outlier or event to occur (Xue et al. 2006), while being inactive for the remaining time. In case of occurrence of an event, the transmission of information towards the sink should occur with severe latency requirements. *Data Gathering and Periodic Reporting Applications* are mostly used for monitoring the environmental conditions like temperature, pressure, humidity, water ingress, light concentration, structural integrity and etc. The nodes of such a network collect some data and broadcast it periodically to the sink, which can perform some computations on the gathered data. *Sink Initiated Querying applications* (Garca-Hernndez et al. 2004; Madden et al. 2002) is another set of WSN applications in which the sink queries a set of sensors for their measurements. For instance, if an anomalous behavior of a node is reported to the sink, it can query some specific sensor nodes to gather further information to determine the source of event. *Tracking based Applications* are mostly used for military and surveillance purposes to detect and track a target. The sensor nodes notify the sink promptly whenever a successful detection occurs.

1.1 Harsh environments

Various applications of WSNs have been studied extensively in literature (Garca-Hernndez et al. 2004; Mainwaring et al. 2002; Cardell-Olivera et al. 2005; Ni et al. 2003; Akyildiz et al. 2002; Dereszynski and Dietterich 2011; Bahrepour et al. 2010c,b; Phua et al. 2010d). This article, however, builds on "Event detection & reporting" applications of WSNs in "Harsh Environments", an aspect which has not been given extensive treatment in the past. Harsh environments are defined as "high stress environments which offer severe monitoring and communication challenges" (Misra et al. 2010). Underground oil, gas, coal, salt mines, forests and volcanic sites (Garca-Hernndez et al. 2004) are typical examples of such environments. These environments have to be monitored continuously for stable, safe, secure and reliable operation. For instance, an underground coal mine should be monitored regularly to ensure the reliability of mine structure, predict any disastrous conditions and more importantly for the safety of personnel working inside.

The communication in such environments is characterized by high signal attenuation, electrical interference and multiple reflections or echoes. Further, several environmental factors like dynamic changes in the underground topology, instability in mine structures, ionized air, humid and warm conditions, gaseous hazards and noise also affect the communication system,

as discussed in Tutorial on Wireless Communications and Electronic Tracking (2009),Dario et al. (2005),Akyildiz et al. (2003). As a consequence of all these factors, communication systems may suffer from limited bandwidth, intermittent link connectivity, high distortion, high packet loss rates, unacceptable packet reception ratio, jitter and delay (Misra et al. 2010). As a safety assurance solution deployed in any harsh environment is highly dependent on monitoring and communication, so these constraints may pose severe challenges in event detection applications.

1.2 Difference between outlier and event

An 'event', also referred to as a 'disastrous condition' or 'hazardous situation', from the perspective of harsh environments, may be characterized by an unexpected change in environmental conditions. Fire, unexpected rise in gaseous concentrations, earthquake and etc. are a few examples of events that may be encountered in harsh environments. 'Outlier' is another term that is closely associated with 'event'. An ' outlier' is an observation that differs significantly from the normal set of readings. The definition of Grubbs (1969), quoted in Barnett and Lewis (1994) states "An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs". Further definitions can be found in Tan et al. (2006). In the context of WSN we may say: "Those measurements that significantly deviate from the normal pattern of sensed data" (Chandola et al. 2009). Figure 1 shows outliers in two dimensional data. An outlier has a low probability that it originates from the same statistical distribution as the other observations in the data set.

Potential sources of 'outliers' in data collected by WSNs include noise and errors, malicious attacks and most importantly *events* (John 1995; Han and Kamber 2006). Of all these sources, events can be characterized as the most important sources of outliers in a harsh environment (Dereszynski and Dietterich 2011; Bezdek et al. 2011). In fact, an 'event' can actually be described as a sequence of 'outliers or erroneous readings' in a streaming data set (Shahid et al. 2012b,a; Shahid and Naqvi 2011). For example, a sensor reading indicating a high temperature independent of its surrounding sensors is an outlier, whereas, a continuous stream of high temperature readings (a stream of outliers) on a group of sensors located close together, indicates the presence of an event.

1.3 How to differentiate between an outlier and event detection scheme

From the discussion in Sect. 1.2 it follows that an *event detection* scheme should essentially be derived from an *outlier detection* scheme (Shahid et al. 2012a; Bahrepour et al. 2009c; Zhang et al. 2012) (We note here that although event detection schemes can be derived from outlier detection schemes, but outliers and events are entirely different entities. Further discussion on outliers and events can be found in Sect. 3). The main differences between event detection and outlier detection are summarized in Table 1 (Zhang et al. 2010).

1.4 Motivation for outlier and event detection

Various accidents have occurred in underground mines worldwide in the previous years. Three accidents occurred in 1891, 1956 and 1958 in different mines within the Springhill coal field due to fire, explosion and earth quake. A total of 238 human lives were lost in the three incidents. Arguably the worst ever mine disaster in the world took place on April 26, 1942 in Benxihu Colliery, located at Benxi, Liaoning. In a very recent accident in November 2009,

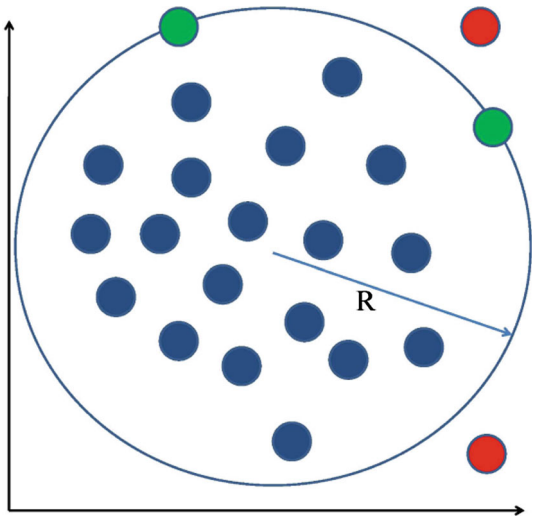

Fig. 1 Outliers in 2D data. The *points*, outside the circle of radius 'R' are outliers

Table 1 Important differences between outlier & event detection techniques for WSNs in harsh environments

	Outlier detection	Event detection
1	No a priori knowledge of trigger or semantic of an event	Hold the trigger condition or semantic of certain event issued by the sink node
2	Outliers identified by the comparison of sensor measurements with each other	Events identified by the comparison of sensor measurements with the trigger conditions or other sensors in the network
3	Need to prevent normal data to be classified as outlier and thus keeping the detection rate high and false alarm rate low	Need to prevent erroneous data which conform to the event condition or pattern to influence reliability of the detection

at least 104 miners were killed. The accident was caused by a methane explosion followed by a coal dust explosion. A methane blast at a Kemerovo coal mine killed 21 miners in 2005. Since 1978, 20 mining accidents have occurred in Poland. Even with the down trend in the fatalities and accidents, 21,351 people were injured between the year 1991 and 1999. In 1972, 91 people lost their lives in Sunshine silver mine at Kellogg Idaho. In 2006, 47 out of 72 miners lost their lives in coal mining. Majority of these deaths occurred at Kentucky and West Virginia. As recent as 5th April 2010, 29 valuable lives were claimed in Upper Big Branch mine disaster at Raleigh County, Montcoal, West Virginia. Numerous examples of such accidents can be found from history which caused severe financial impact and resulted in a great loss of human lives (http://connekt.seecs.nust.edu.pk/SAHSE.php, http://www.humanite.fr/2006-03-10_Societe_-Catastrophe-de-Courrieres-une-expression-impropre, http://www.genuki.org.uk/big/eng/LAN/Haydock/WoodPitExplosion.html, http://www.msha.gov/MSHAINFO/FactSheets/MSHAFCT8.HTM). Therefore, outlying data should be analyzed, as it may indicate a system moving towards a state of natural disaster or event. Thus, *outlier detection* in harsh environments is essential for *event detection*, data quality assurance and control (QA/QC), fault detection, intrusion detection (Chen et al. 2006; Luo et al. 2006; Silva et al. 2005; Bhuse and Gupta 2006), focused data collection, (Zoumboulakis and Roussos 2007;

Krishnamachari and Iyengar 2004; Ding et al. 2005; Ding and Cheng 2009) and adaptive system monitoring (Misra et al. 2010).

1.5 What is this survey about?

To ensure a high performance of outlier and event detection algorithms on WSNs deployed for monitoring of harsh environments, these algorithms should satisfy a certain set of characteristics. In general, outlier detection algorithms for WSNs must be energy conserving, robust to topological changes in the network, pose minimal communication overhead and use distributed approach or partially distributed approaches. All these characteristics ensure minimum energy utilization by WSNs. Moreover, they should have a high detection rate and low false alarm rate. However, the deployment of such a network in harsh environments, underground mines for instance, requires the algorithms to satisfy an additional set of characteristics. For example, the algorithms should be unsupervised, simple and computationally efficient, consider multivariate data and spatio-temporal-attribute data correlations, perform the separation of erroneous measurements and events, and should perform real-time/online detection of anomalous data.

In this survey paper, we discuss the characteristics that should be possessed by an outlier and event detection technique to be feasible for *WSNs in harsh environments*. This work is different from other survey papers (Chandola et al. 2009; Hodge and Austin 2004; Rajasegarar et al. 2008a), in that, it summarizes the analysis criteria of outlier detection techniques into a single discussion, *specifically in the context of harsh environments*. Further, a summary of the prioritization of various characteristics extremely essential for harsh environments has been presented in the form of a flow chart. This paper enables the reader to clearly differentiate between the characteristics that are essential for outlier detection techniques for simple WSNs (those deployed in non-harsh environments, e.g, in an open field, room) and those for WSNs deployed in harsh environments. Taxonomy based state-of-the-art classification scheme of outlier detection schemes for WSNs has been presented and the suitability of various types of techniques has been determined for harsh environments based on the discussed characteristics. This paper may also enable a reader to test the feasibility of any state-of-the-art outlier detection technique, previously presented in the literature, for harsh environments. Further, this paper may also strongly motivate the researchers in this field to devote their future research in developing optimal outlier and event detection algorithms which satisfy the discussed characteristics. Upto the best of our knowledge, this is the first detailed work in context of outlier detection techniques for WSNs in harsh environments.

Rest of this paper is organized as following. Section 2 deals with a description of methods for the analysis of various techniques in harsh environments. Section 3 briefly describes the state-of-the-art taxonomy based classification criteria for outlier detection techniques used in WSNs. Section 4 gives a detailed description of the characteristics of outlier detection techniques for WSNs. Section 5 defines harsh environment, explains the challenges imposed by harsh environments on outlier detection techniques and discusses various characteristics of outlier detection techniques for WSNs deployed in harsh environments. Section 6 explains the relationship between various characteristics presented in Sects. 4 and 5. Section 7 gives the prioritization of various characteristics to determine the feasibility of outlier detection technique for harsh environments. Section 8 briefly describes the feasibility of various types of outlier detection techniques in WSNs (introduced in Sect. 3) for harsh environments. Section 9 deals with the future work and research possibilities and Sect. 10 concludes the paper.

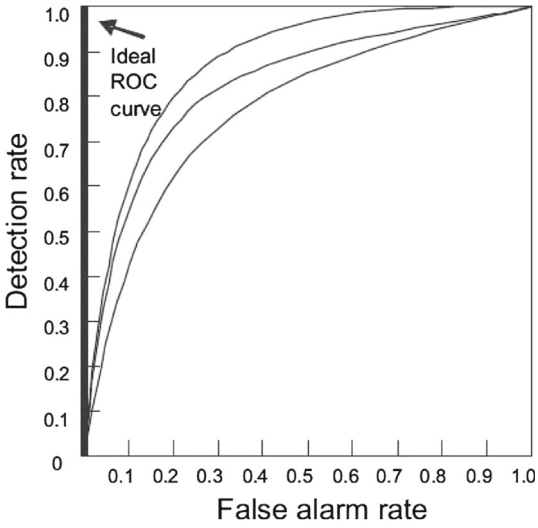

Fig. 2 Receiver operating characteristics curve (ROC) for various outlier detection techniques. Ideal ROC has maximum area enclosing it (Ganguly 2008)

2 Methods for analysis of outlier detection techniques

Before going into the detailed analysis of characteristics essential for outlier detection techniques in harsh environments, we discuss three commonly used methods for evaluation of these algorithms. These quantitative measures are mostly used by researchers to present the accuracy of their developed outlier and event detection algorithms.

1. Detection Rate
2. False Alarm Rate / False Positive Rate
3. Receiver Operating Characteristic (ROC) Curve

The effectiveness of any outlier detection technique can be evaluated quantitatively depending on the number of outliers correctly identified: known as the *detection rate* and the fraction of normal data incorrectly considered as outliers: known as *false alarm rate*. The receiver operating characteristic (ROC) curve (Lazarevic et al. 2003) represented in the form of a 2-D graph is usually used to represent the trade-off between detection rate and false alarm rate. An example ROC curve is shown in Fig. 2. The effectiveness of an outlier detection technique depends on the capability to maintain a high detection rate while keeping the false alarm rate low and large area under ROC curve (Ganguly 2008).

3 Classification of outlier detection techniques for WSNs

This section briefly presents the state-of-the-art classification criteria of outlier detection techniques for WSNs. Various types of techniques presented in this section will then be discussed in term of their feasibility for harsh environments in Sect. 5, where we will introduce the characteristics of optimal outlier detection techniques for harsh environments. Thus, based on the methodology used for outlier detection, the techniques have been classified into the following four types. A detailed explanation of this classification criteria can be found in Zhang et al. (2007a). Here we present only a brief analysis of various types.

3.1 Statistical based techniques

These techniques require an underlying data distribution model for the detection of outliers. 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, based on the distance measure. These techniques can further be classified as parametric or non-parametric. Parametric techniques assume availability of the knowledge about underlying data distribution, i.e., the data is generated from a known distribution. Distribution parameters are then estimated from the available data. These techniques are based on either a Gaussian based model or a non-Gaussian model. 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. Most widely used approaches in this respect are histogram, kernel density and wavelet based approaches. Some of the statistical based techniques considered in this paper are Dereszynski and Dietterich (2011), Zhang et al. (2012), Wu et al. (2007), Yozo et al. (2004), Jun et al. (2005), Bettencourt et al. (2007), Sheng et al. (2007), Palpanas et al. (2003), Subramaniam et al. (2006)

3.2 Nearest neighbor based techniques

Nearest neighbor-based approaches had been the most commonly used approaches to analyze a data instance with respect to its nearest neighbors in the data mining and machine learning community in the past. They use several well-defined distance notions to compute the distance (similarity measure) between two data instances. A data instance is declared as an outlier if it is located far from its neighbors. Euclidean distance is a popular choice for univariate data, whereas, multivariate continuous attributes are handled by Mahalanobis distance metric. Some of the nearest neighbor based techniques considered in this paper are presented in Branch et al. (2006), Zhang et al. (2007b), Zhuang and Chen (2006). These techniques have not been the focus of research community recently due to the limitations that will be discussed in the forth-coming sections.

3.3 Clustering based techniques

Grouping similar data instances into clusters with similar behavior is known as clustering. Clustering algorithms can be either centralized or distributed. In centralized clustering algorithms each node transmits its entire data to the gateway/ central node which then performs data clustering. This approach is however communication inefficient. In a distributed clustering approach, all the nodes are able to perform clustering of the sensed data vectors and then send specific parameters of clustered data to the gateway node to reduce communication overhead. The nodes then use some distance measure from the nearest cluster to identify outliers (Bezdek et al. 2011; Rajasegarar et al. 2008a, 2010b, 2012; Moshtaghi et al. 2011a,b,c; Bezdek et al. 2010; Suthaharan et al. 2010a,b).

3.4 Classification based techniques

Classification based techniques learn a classification model using the set of data instances during the training phase and then classify the data instance to one of the training classes

during the testing phase. These techniques can be either supervised or unsupervised. The one-class unsupervised techniques learn the boundary around normal instances during training while some anomalous instance may exist and declare any new instance lying outside this boundary as an outlier. The boundary may be defined as a sphere or quarter-sphere. However this type of classifier may need to train itself according to the new arriving normal data sets. In existing outlier detection methodologies for WSNs, classification based approaches are categorized into support vector machine (SVM) based and Bayesian network based approaches depending upon the type of classification model that is used (Bahrepour et al. 2010b; Shahid et al. 2012a,b; Shahid and Naqvi 2011; Luo et al. 2006; Elnahrawy and Nath 2004; Janakiram et al. 2006; Hill et al. 2007; Zhang et al. 2009a,b; Rajasegarar et al. 2006, 2007, 2008a,b, 2010a).

4 Characteristics of outlier detection techniques for WSNs

This section identifies and discusses several important aspects of state-of-the-art outlier detection techniques specially developed for WSNs and those presented in Sect. 3. These characteristics can be used as metrics to determine the feasibility of different outlier detection techniques for non-harsh environments. To be feasible for harsh environments, the outlier detection techniques should satisfy an additional set of characteristics which will be discussed in the next section. Following is a detailed description of various characteristics and their significance for WSNs.

4.1 Energy efficiency

WSNs have been found to be of key importance in monitoring applications. Specifically, the monitoring of remote and isolated environments has been a primary application of WSNs. Consider for example the case of a WSN deployed in a forest for fire detection, where a large number of sensor motes are randomly placed at various locations. Due to the physical constraints posed by a large forest, the replacement of the batteries of the sensor motes may present severe problems. Therefore, it is essential that once the WSN has been deployed in such an environment, the battery life should be made as long as possible by conserving the amount of energy consumed in computations and communication. Outlier and event detection techniques for WSNs can be made energy efficient by ensuring the following two characteristics.

4.1.1 Low computational and communication complexity

Sensor nodes in a WSN have limited power and a major portion of energy is consumed in communication and computation, so techniques should be computation and communication efficient. It has been proved that the communication cost of a sensor node is several orders of magnitude higher than the computation cost (Gupta and Kumar 2000; Shnayder et al. 2004). Over the recent years, significant work has been dedicated to reduce communication overhead in a WSN by increasing the computations at individual nodes of the network without compromising the performance. These efforts have led to a significant improvement in the battery life of WSNs deployed in remote environments. Various outlier and event detection techniques proposed in the literature require computations at individual nodes of the network followed by communication between the nodes of the network.

- Statistical based techniques usually pose low communication and computational complexity as compared to other techniques as they require an underlying data distribution and simply declare the most remote points as outliers (Dereszynski and Dietterich 2011; Zhang et al. 2012; Wu et al. 2007; Yozo et al. 2004; Jun et al. 2005; Bettencourt et al. 2007; Sheng et al. 2007; Palpanas et al. 2003; Subramaniam et al. 2006).
- Clustering based techniques have more computational complexity as compared to statistical based techniques as they require the computation of a distance metric for every data sample. The degree of deviation is used to declare a data sample as an outlier or normal. However, the communication complexity is comparable to that of statistical based techniques, as only a few cluster parameters need to be broadcasted between various nodes of the network (Bezdek et al. 2011; Shahid et al. 2012a,b; Zhang et al. 2012; Moshtaghi et al. 2011a; Bezdek et al. 2010; Rajasegarar et al. 2010a; Giatrakos et al. 2010a,b).
- Classification based techniques pose a computational complexity that is greater than that of clustering and statistical based techniques. Specifically, the SVM based techniques require the solution of a quadratic or linear optimization problem at every time instant. Bayesian based techniques have been reported to pose more computational complexity because they associate a probability measure between each pair of attributes at every time instant. Some of the most recent techniques which have resulted in a magnitude of reduction in communication complexity without a loss of performance are presented in Shahid et al. (2012a,b); Rajasegarar et al. (2010a); Bahrepour et al. (2009a).
- Nearest neighbor based techniques have the greatest computational complexity as compared to all other state-of-the-art techniques because they require the computation of multi-variate euclidian distance between each pair of data samples (Branch et al. 2006; Zhang et al. 2007b; Zhuang and Chen 2006). Their communication complexity is comparable to other techniques.

4.1.2 Distributed computations

An outlier and event detection technique for WSNs requires computations as well as communication between various nodes of the network, therefore, depending on these two requirements, the techniques can be classified as:

- Centralized
- Distributed

In a *centralized approach*, all the data received at individual nodes is transmitted to central node. The central node is responsible for processing the entire data received from the network and determine outliers or events. This technique provides a global perspective of the entire network to the central node (Sheng et al. 2007), but, it requires excessive multi-hop communication between various nodes of the network. Since communication cost is several orders of magnitude higher than the computation cost, so, the centralized approaches to outlier and event detection are energy inefficient and cannot be used for WSNs (Gupta and Kumar 2000; Shnayder et al. 2004).

An essential characteristic of outlier and event detection techniques that has been given primary importance in the recent years is the 'distributed nature' of technique. In a *distributed approach*, as the name implies, the outlier or event identification process is divided between all the nodes of the network. Thus, each node maintains a record of its newly arrived data and then performs some processing on it to determine the sufficient statistics of gathered data. The nodes then broadcast this data to a cluster head of the network. The cluster head is responsible for processing the sufficient statistics received from all the nodes and determine

global statistics. The global statistics are then broadcasted to all the nodes in the network. The global statistics give a perspective of the whole network to all nodes of the network. Various distributed outlier and event detection techniques have been presented in the literature.

- Most of the statistical based techniques proposed in the literature require centralized computations (Wu et al. 2007; Yozo et al. 2004; Jun et al. 2005; Bettencourt et al. 2007; Sheng et al. 2007; Palpanas et al. 2003; Subramaniam et al. 2006). A very few state-of-the-art techniques are distributed (Dereszynski and Dietterich 2011; Zhang et al. 2012).
- More recently, the clustering based techniques have opened a new era of research for distributed outlier detection techniques. Clustering is used to identify the outliers at individual nodes of the network. The aggregator nodes in the network then collect data from the neighboring nodes, broadcasting the cluster parameters upto the sink node. This process forms a global perspective of the whole network at the sink node which then helps in identifying the outliers (Bezdek et al. 2011; Subramaniam et al. 2006; Rajasegarar et al. 2010b, 2012; Moshtaghi et al. 2011a,b,c; Bezdek et al. 2010; Suthaharan et al. 2010a,b).
- Classification based techniques have also been used to perform distributed computations. SVM based techniques compute the radius of sphere or quarter-sphere at each node of the network and classify the data instances lying outside the boundary as outliers (Shahid et al. 2012a,b; Shahid and Naqvi 2011; Rajasegarar et al. 2006, 2007, 2008a,b, 2010a; Elnahrawy and Nath 2004; Janakiram et al. 2006; Hill et al. 2007; Zhang et al. 2009a,b).
- Nearest neighbor based techniques perform centralized computations mostly.

Following are a few advantages of distributed approaches:

- These techniques do not require the broadcast of entire data to the central node of the network, thus, resulting in a significant reduction in the communication complexity and data traffic in the network.
- These techniques require each node to process its own data and broadcast only a few parameters of the data to the central node. This reduces the amount of computations to be performed at the central node.

4.2 Outlier identity

There are three potential outlier sources in WSNs:

1. sensor errors
2. Events
3. Malicious attacks

Outlier detection techniques are meant to identify the outliers and perform further operations on the determined outliers or simply discard them. This article involves a discussion about outliers in the context of errors and events only.

4.2.1 Sensor errors

Errors may occur due to various sources, such as a sensor misbehavior or sensor fault. In a distributed approach, errors are normally characterized by a relatively large probability of occurrence as compared to events. Various outlier detection techniques should be able to differentiate between sensor errors, sensor misbehavior and events. The deployment of WSNs in harsh environments, makes the measurements prone to noise as well. Thus, it is typical for a noise measurement to be encountered as a sensor measurement. Most of the outlier

and event detection techniques presented in the literature focus on differentiating between outliers and events (Bezdek et al. 2011; Shahid et al. 2012a,b; Zhang et al. 2012; Moshtaghi et al. 2011a; Bezdek et al. 2010; Rajasegarar et al. 2010a), however, the characterization of outliers as sensor faults, misbehavior or noise has not been dealt. Recently some of the work has been dedicated to detecting sensor errors (Chen et al. 2006; Sharma et al. 2010), however, this domain needs to be explored more by the research community.

4.2.2 Events

Some of the measurements may be encountered which are neither errors nor noise, but still significantly deviate from the normal data set. These observations are called events. Event detection is the most important characteristics that should be present in a technique if it is being used for WSNs deployed in an environment with a non-zero probability of event occurrence. Typical examples of events may include fire, flood, earthquake, volcanic eruption and etc. Following are a few features that should be possessed by an event detection algorithm. Significant work has been dedicated for each of these features in the past.

- An event detection technique should identify the event with a high probability, i.e, a high detection rate and a very low false positive rate.
- It should be able to carry out an analysis of the detected event. Often the WSNs are deployed in unreachable environments which are being monitored from remote locations. Once an event is detected, the algorithm should initiate an analysis of the event to determine the type of event, i.e, the identification of an event as fire, flood, volcanic eruption, explosion and etc. Significant work has been done in the past which deals with each of the events independently, for example various algorithms proposed in the literature focus on fire detection, however, such algorithms cannot detect any other type of event in the monitored environment (Misra et al. 2010; Bahrepour et al. 2008, 2009c,b, 2010c,b,a; Keally et al. 2010; Zoumboulakis and Roussos 2007). A very few algorithms tend to determine the class or type of event. Significant work has been dedicated for classifying the intrusion in a WSN, such classification algorithms should also be used for event type identification in WSNs (Abe 2010q; Liu et al. 2007, 2010; Hao et al. 2009; Xu 2009; Xu et al. 2007; Aly 2005; Keerthi et al. 2008).
- An event is defined as a sequence of outliers in streaming data, thus, an event detection strategy may also be derived from an outlier detection technique (Bahrepour et al. 2009c). Some of the algorithms that perform event detection by modifying the outlier detection algorithms have been presented in Bezdek et al. (2011), Shahid et al. (2012a,b), Zhang et al. (2012), Rajasegarar et al. (2010a).
- Additionally, an event detection technique should also be able to define an event region and an event boundary for a complete specification of an event (Xue et al. 2006; Luo et al. 2006; Krishnamachari and Iyengar 2004; Ding et al. 2005; Ding and Cheng 2009; Wu et al. 2007; Suthaharan et al. 2010b).

Following are a few methods discussed in the literature for event detection in WSNs.

- Some of the Statistical based techniques incorporate an event detection strategy (Wu et al. 2007; Bettencourt et al. 2007), whereas most of the techniques do not perform event detection (Dereszynski and Dietterich 2011; Zhang et al. 2012; Yozo et al. 2004; Jun et al. 2005; Sheng et al. 2007; Palpanas et al. 2003; Subramaniam et al. 2006).
- Clustering based techniques perform event detection by comparing the cluster parameters of various nodes in the network. If an outlier is detected at any node, it can invoke

the decision of all other nodes in the neighborhood. Majority voting can then be used to determine if the outlier is an event (Bezdek et al. 2011; Subramaniam et al. 2006; Rajasegarar et al. 2010b, 2012; Moshtaghi et al. 2011a,b,c; Bezdek et al. 2010; Suthaharan et al. 2010a,b).

- Classification based techniques also determine the presence or absence of an outlier by majority voting of a group of nodes in a particular neighborhood (Bahrepour et al. 2010b; Shahid et al. 2012a,b; Rajasegarar et al. 2008a, 2010a; Zhang et al. 2009a,b).
- Nearest neighbor based techniques do not perform event detection. Some of the nearest neighbor based outlier detection techniques can be extended to event detection, but they have high computational complexity.

4.2.3 Outlier handling strategy

A complete outlier and event detection algorithm should be able to perform the following steps:

1. *Outlier labeling*: Outlier labeling means the detection of outliers from the data set. Primarily, the algorithm should be able to classify the streaming data samples as normal or outlier. This is performed by all outlier detection algorithms.
2. *Outlier Cause*: Once an outlier has been detected, the algorithm should begin a root cause analysis to determine the source of a outlier, i.e, event, sensor fault or sensor misbehavior.
3. *Event Identification*: If an event is determined to be the source of outlier, another analysis should be initiated to determine the type of event. A very few algorithms presented in the literature are able to perform this complete analysis.
4. *Outlier Accommodation*: If no root cause for an outlier can be determined, and a retest can be justified, the potential outlier should be recorded for future evaluation as more data become available. Removing data points on the basis of statistical analysis without an assignable cause is not acceptable. Robust or non-parametric statistical methods are alternate methods for analysis. Robust statistical methods such as weighted least-squares regression minimize the effect of an outlier observation (Han and Kamber 2006). Robust outlier detection techniques should be employed when the number of outliers is large, so that the resulting data distribution is not skewed, however non-robust techniques can be employed when the number of outliers is small.

Generally, the outlier detection techniques can also be classified based on the number of outliers they determine. However, the techniques should be able to determine and analyze multiple outliers in a data set.

1. *Single Outliers*: Only a few of the earlier techniques identified single outliers, like Grubbs test. Most of the state-of-the-art outlier detection techniques can identify multiple outliers.
2. *Multiple Outliers*: Multiple outliers can be identified by most of the techniques. Graph based techniques enable a visual identification of multiple outliers, similarly threshold based techniques also enable multiple outlier detection, as all the data points beyond a particular threshold are considered as outliers.

4.3 In-susceptibility to dynamic changes in network topology

Sensor networks often undergo the deletion process of sensor nodes as a consequence of which some new nodes should take the charge of dead nodes. This process of addition and deletion of nodes is known as ‘dynamic change in the network topology’. Outlier and event

detection techniques should be robust to such changes in the topology. Following are a few reasons for dynamic changes in the network topology.

- For a WSN deployed in a remote and huge environment, the battery of some of the sensor nodes may drain out earlier than the other nodes in the same network. This may lead to a loss of information towards the sink node if the faulty node lies in the multi-hop path towards the sink. Such a condition is known as a 'network hole' (Schieferdecker et al. 2011). In such cases another sensor node should replace the faulty node to route the information to sink node and perform outlier and event detection.
- A disastrous event may lead to destruction of a few nodes in the network. This situation may also produce a 'network hole'.

The performance of Outlier and event detection algorithms may suffer significantly due to dynamic changes in the network topology. Thus the techniques should be robust to such changes in the network. Only a few techniques have been presented in the literature that are robust to topology changes. Bayesian based techniques are the only examples of such techniques (Krishnamachari and Iyengar 2004). Statistical, clustering and nearest neighbor based techniques do not incorporate this feature.

5 Characteristics of outlier detection techniques for WSNs deployed in harsh environments

Various characteristics were discussed in the previous section for outlier and event detection techniques in WSNs. These characteristics are only suitable for WSNs deployed in non-harsh environments, for example, office buildings and homes. WSNs deployed in such environments are small scale and do not pose severe monitoring and communication challenges. However, the deployment of WSNs in harsh environments like underground mines, volcanic sights, forests etc. pose several additional challenges. Thus, outlier and event detection strategies for such environments should satisfy an additional set of features.

5.1 Constraints posed by harsh environments and their effect on the performance of outlier and event detection techniques

Harsh environments, as described in Misra et al. (2010), are characterized by severe monitoring and communication challenges. Thus the constraints posed by such environments on a WSN can be divided into two major categories.

1. Communication constraints
2. Environmental monitoring constraints

The first set of constraints are caused due to extreme path loss, signal absorption, spreading, rapidly changing time-varying channels, large propagation delay, noise and fading characteristics. All these characteristics of a wireless channel in a harsh environment effect the communication between various nodes in the network. The performance of outlier detection algorithms is weakly related with communication constraints, as most of the outlier detection techniques do not require significant communication between the nodes. Thus, the outliers can be determined via processing on individual nodes of the network. The performance of event detection algorithms, however, depends on the nature of communication between the nodes. State-of-the-art outlier detection techniques have reduced the amount of communication between the nodes by orders of magnitude. Thus, instead of entire data, only a

few parameters related to the measured data are exchanged between the nodes. Hence, the performance of outlier and event detection algorithms is not significantly affected by the communication constraints.

The second set of constraints, also known as monitoring constraints, are directly related to the quality of data measured in a WSN. These constraints are imposed by dynamic changes in the network topology, instability in mine structure and dynamic changes in the data distribution of various attributes being measured in a harsh environment. Noise associated with the raw data samples can be filtered before processing but the dynamic nature of data distributions is the key feature of harsh environments that poses significant challenges for outlier and event detection techniques. Thus, the monitoring constraints play a key role in testing the performance of outlier and event detection techniques in such an environment. Due to these constraints and to ensure a high performance of detection techniques, they should satisfy an additional set of characteristics when being used for WSNs deployed in harsh environments.

Following discussion presents a list of characteristics which should be satisfied by the techniques for harsh environments. We note that these characteristics should be satisfied in addition to those discussed in the previous section for optimal performance. We also note that the requirement of most of these characteristics arises due to non-stationary nature of data distributions.

5.2 Ability to process complex data

Sensor data is normally viewed as large sensor streams that are handled by the sensor nodes. These sensor streams consist of continuous data that is sensed by the sensors. Different outlier detection techniques can be analyzed in terms of type of input data they can handle.

In addition to the classification of data as univariate and multivariate, data can also be classified as numeric and symbolic. Numeric data is normally used in quantitative approaches or graph based approaches, whereas statistical techniques use symbolic data mostly. Numeric data may be continuous-valued, discrete (ordinal) or categorical (unordered numeric). Similarly symbolic data may be either ordered symbolic or unordered symbolic. We describe several common types of data sets based on the characteristics and attributes of data. They are divided into simple and complex data sets (Zhang et al. 2007a). In the context of WSNs deployed in harsh environments, multivariate and streaming data sets may be referred to as 'Complex data sets'.

5.2.1 Simple data set

The simple data set belongs to a commonly used data set, where the data has no complex semantics and usually is represented by low dimensional real-valued attributes (Zhang et al. 2007a). Such types of data sets are often encountered in WSNs deployed in non-harsh environments.

5.2.2 Multivariate data

Some of the outlier detection algorithms are able to handle only univariate data whereas some can handle multivariate data as well. A data value is said to be an outlier if its attributes have anomalous values. If an algorithm is only able to handle univariate data then outlier detection is simple as it only needs the identification of a single attribute being different from that attribute of other data. Various algorithms consider only univariate data such as

Yozo et al. (2004), Jun et al. (2005), Bettencourt et al. (2007). However for accurate determination of outliers, the underlying technique must consider multivariate data. WSNs are usually deployed in high stress environments, where different environmental conditions show correlations between their attributes. If an algorithm is unable to handle multiple attributes and their correlations, then it may not be suitable for harsh environments. WSNs may also be deployed in environments where individual attributes do not show any correlation and hence no outlier identification may result even in the presence of significant outliers. Therefore multi-variate data handling is important and essential for outlier detection techniques in harsh environments. However the use of multivariate techniques also increases computational complexity. Most of the earliest statistical based outlier detection techniques consider only univariate data, thus making them inefficient for use in harsh environments. A large number of state-of-the-art techniques can handle multi-variate data. Some examples of such techniques can be found in Bahrepour et al. (2010b); Bezdek et al. (2011); Shahid et al. (2012a,b); Rajasegarar et al. (2008a, 2010a,b, 2012); Subramaniam et al. (2006); Moshtaghi et al. (2011a,b,c); Bezdek et al. (2010); Suthaharan et al. (2010a,b); Zhang et al. (2009a,b); Yang et al. (2008); Wang et al. (2006); Tax and Duin (1999).

5.2.3 Streaming data set

A data stream is a large data that is arriving continuously in the ordered sequence. Such data is usually unlimited in size and occur in many real-time applications. For example, a huge amount of data of the average daily temperature are collected to the base station in wireless sensor networks continually. As harsh environments are characterized by frequent changes in the data distributions, so, the outlier and event detection schemes for such environments should be able to handle streaming data in an online manner. Such a technique would be able to analyze and handle various changes in the distribution and more frequent outliers and events resulting due to sudden changes. Only a few state-of-the-art techniques are able to handle streaming data (Bahrepour et al. 2010b; Shahid et al. 2012a,b; Moshtaghi et al. 2011a,b,c; Rajasegarar et al. 2010a; Zhang et al. 2009a,b).

5.3 Unsupervised data model

A straightforward approach to identify the outliers is to construct the normal profile of the data and then use the normal data to detect outliers. The observations whose characteristics differ significantly from normal data are classified as outliers (Rajasegarar et al. 2008a). Based on the type of data available from sensors, the techniques are classified as Hodge and Austin (2004):

5.3.1 Unsupervised

This approach assumes that errors or faults are separated from the normal data. It processes the data as a static distribution, and flags the most remote points as potential outliers, without the need of a pre-defined normality/abnormality model (Hodge and Austin 2004). It requires that all data be available before processing. Once the system possesses a sufficiently large database with good coverage, it can then compare new items with the existing data. This approach also involves Diagnosis/labeling and accommodation. Diagnostic approach highlights the outlying points. Once detected, the system may remove these outlier points from future processing of the data distribution. If the outlier points are not removed then they can be accommodated within the data set. Accommodation incorporates the outliers

into the distribution model and provides a classification method that is robust to outliers. These robust approaches can withstand outliers in the data and generally form a boundary of normality around the majority of the data to represent the normal behavior. In contrast, non-robust classifier methods produce representations which are skewed (Hodge and Austin 2004). A certain measure criteria, for example distance based approaches, is used to measure the outlier (Hodge and Austin 2004). Various clustering based techniques presented in the literature form unsupervised data model to determine outliers and events (Bezdek et al. 2011; Rajasegarar et al. 2008a, 2010b, 2012; Subramaniam et al. 2006; Moshtaghi et al. 2011a,b,c; Bezdek et al. 2010; Suthaharan et al. 2010a,b).

5.3.2 Supervised

These techniques require the modeling of both normality and abnormality and require pre-labeled data. The normal points could be classified into a single class or subdivided into distinct classes according to the requirements of system to provide a simple normal/abnormal classification (Hodge and Austin 2004). This approach cannot be used for on-line classification, where the classifier learns the classification model with the arrival of every new data sample and then classifies new data samples against the learned model. Statistical based techniques are mostly supervised (Dereszynski and Dietterich 2011; Zhang et al. 2012; Wu et al. 2007; Yozo et al. 2004; Jun et al. 2005; Bettencourt et al. 2007; Sheng et al. 2007; Palpanas et al. 2003; Subramaniam et al. 2006).

5.3.3 Semi-supervised

This approach models only normality or abnormality (Hodge and Austin 2004; Rousseeuw and Leroy 1996). It is also known as novelty detection or novelty recognition. It is known as semi-supervised, as the normal/abnormal class is taught, but the algorithm learns to recognize the other class. It is suitable for dynamic data, as it only learns one class which provides the model of normality or abnormality. It can learn the model incrementally as new data arrives, tuning the model to improve the fit, as each new exemplar becomes available. These techniques aim to define a boundary of the normal or abnormal class. This boundary may be hard, where a point lies wholly within or wholly outside the boundary, or soft, where the boundary is graduated depending on the underlying detection algorithm (Hodge and Austin 2004). A soft bounded algorithm can estimate the degree of *'outlierness'*. A few classification based techniques are good examples of semi-supervised techniques (Bahrepour et al. 2010b; Shahid et al. 2012a,b; Rajasegarar et al. 2010a; Zhang et al. 2009a,b).

Harsh environments are typically characterized by temporally varying and unknown data distributions. Since a normal or abnormality model of the data is difficult to obtain in a harsh environment and unsupervised techniques adapt themselves to changing data distributions and identify the remote points as outliers, hence, an outlier detection technique for harsh environments should be unsupervised. Figure 3 presents a summary of supervised, unsupervised and semi-supervised approaches for outlier detection.

5.4 Data correlations

Following types of dependencies may exist between the data at each node:

1. Dependencies among the attributes of sensor node.
2. Dependency of sensor node readings on its history.
3. Dependency of sensor node reading on its neighboring nodes.

5.4.1 Attribute correlations

The dependency between various attributes of a harsh environment is called *attribute correlations*. In a harsh environment such as an underground mine, pressure, temperature and humidity may be correlated. For example, a sudden rise in temperature may be characterized by a sudden fall in humidity. Thus in this case the temperature and humidity attributes have a negative correlation. Attribute correlations play an important role in outlier and event detection in WSNs and have been reported to increase the outlier detection rates significantly as compared to the techniques which are devoid of such correlations (Shahid et al. 2012a). Various methods have been proposed to incorporate attribute correlations in an outlier detection technique. Some of the methods are described below.

- Classification based techniques, which classify the data samples from a sensor node as an outlier or normal based on a semi-supervised data model can be divided into Support Vector Machine (SVM) and Bayesian based techniques. SVM based techniques incorporate attribute correlations by formulating a SVM problem across various attributes of measured data (Shahid et al. 2012a,b; Shahid and Naqvi 2011), whereas, bayesian based approaches do so by assigning various probability measures to the covariance between the attributes (Krishnamachari and Iyengar 2004; Ozdemir and Xiao 2011; Hassan 2011).
- Clustering based techniques, which detect outliers by forming clusters of normal data, incorporate attribute correlations by using a specific distance measure known as 'Mahalanobis distance' (Bezdek et al. 2011; Rajasegarar et al. 2008a, 2010b, 2012; Moshtaghi et al. 2011a,b,c; Bezdek et al. 2010; Suthaharan et al. 2010a,b). This distance measure determines the deviation of individual data samples from the cluster center by taking into account the covariance matrix of multi-variate data.
- Statistical and nearest neighbor based techniques do not incorporate attribute correlations.

5.4.2 Temporal correlations

In addition to the attribute correlations between various attributes of measured data in a WSN, the measured data at current time instant is also dependent on the historical measurements. Such a dependency is known as 'Temporal Correlation' in various attributes of the streaming data. We note that temporal correlations are independent of attribute correlations. Thus, various attributes of a multi-variate data may show varying temporal correlations. Since harsh environments are characterized by frequent changes in data distributions with time, so temporal correlations play an important role in outlier detection. Almost all of the outlier detection techniques proposed in literature for WSNs possess this property. Multiple methods have been adopted to incorporate temporal correlations. Some of them have been discussed below.

- Simple statistical based techniques use temporal correlations to determine the presence or absence of an outlier. A sudden change in the data distribution reduces the temporal correlations and this helps in outlier detection in streaming data (Ross et al. 2009).
- Classification based techniques, such as SVM based techniques incorporate temporal correlations by enclosing the data collected from WSN in a geometric figure such as a sphere or a quarter-sphere in such a way that outlying points remain outside the geometry (Bahrepour et al. 2010b; Shahid et al. 2012a,b; Shahid and Naqvi 2011; Yozo et al. 2004; Rajasegarar et al. 2008b, 2010a; Zhang et al. 2009a,b; Yang et al. 2008; Wang et al. 2006; Tax and Duin 1999; Gomez-Verdejo et al. 2011).

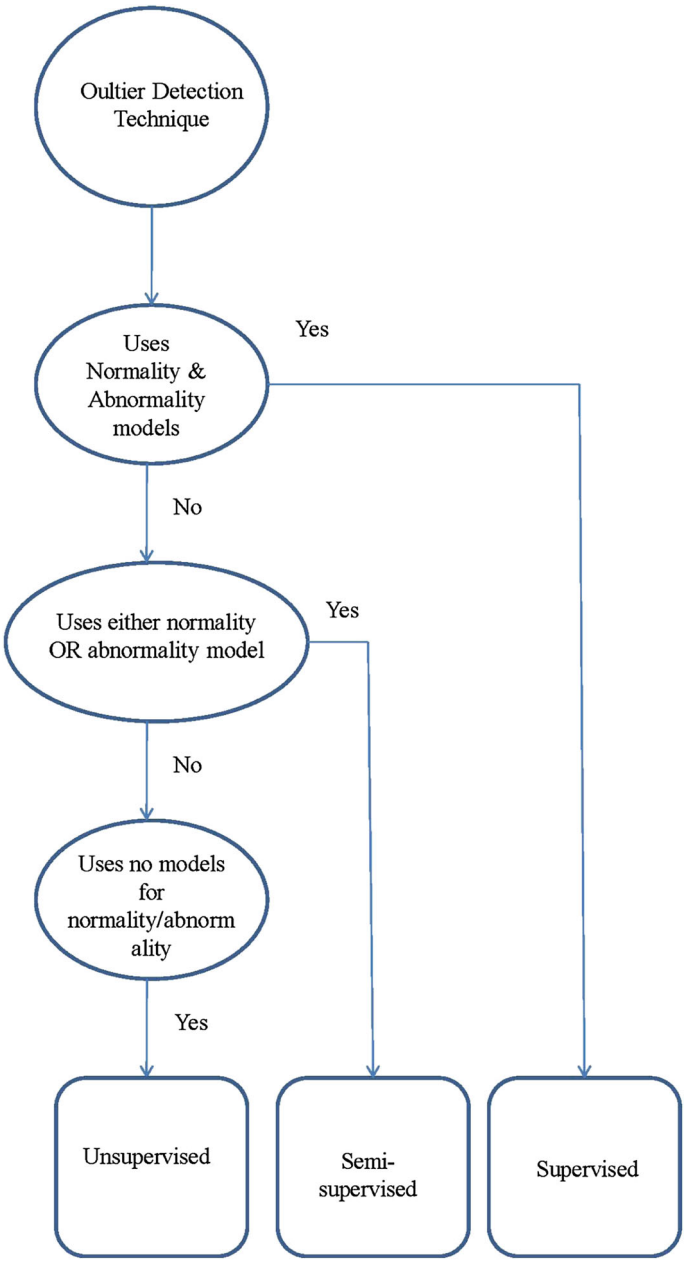

Fig. 3 Summary of the classification of outlier detection techniques into supervised, unsupervised and semi-supervised approaches. Unsupervised approaches are used for harsh environments, however, some semi-supervised techniques can also be used

- Clustering based techniques incorporate temporal correlations by enclosing the most recent samples of collected data in a cluster. The cluster parameters such as mean and covariance are then updated with the arrival of every new data sample (Bezdek et al. 2011; Rajasegarar et al. 2010b, 2012; Moshtaghi et al. 2011a,b,c; Bezdek et al. 2010; Suthaharan et al. 2010a,b).
- Nearest neighbor based techniques do not incorporate temporal correlations.

5.4.3 Spatial correlations

The dependency between the data at various nodes is called spatial correlation, which implies that the data values at a particular node are related to the data samples of the neighboring nodes. Existence of spatial correlations implies that the readings from sensor nodes that are geographically close to each other are expected to be largely correlated. Some of the state-of-the-art event detection techniques utilize spatial correlations between the data of all nodes in a neighborhood. Capturing spatial correlations provides a global picture of the entire neighborhood or region and also helps to distinguish between errors and events (Zhang et al. 2007a). Following is a list of few methods that have been used in literature to incorporate spatial correlations.

- Statistical based techniques perform a comparison of various statistical measures between the nodes in a particular neighborhood (Dereszynski and Dietterich 2011; Zhang et al. 2012).
- Recently, the classification based techniques, specifically the SVM based techniques have been used for event detection in WSNs. These techniques use spatial correlations to differentiate between outliers and events. The underlying assumption for event detection is that an event is characterized as a sequence of outliers on a node (Shahid et al. 2012a,b; Shahid and Naqvi 2011; Rajasegarar et al. 2008b, 2010a; Zhang et al. 2009a,b; Yang et al. 2008. The detection of an outlier on a particular node is followed by a consensus of all other nodes in a neighborhood. The decision of consensus then determines the presence or absence of an event in that region. The consensus can be based on a number of metrics. Some of the techniques simply use a majority voting scheme; i.e, if an outlier is detected on more than half of the nodes in the network then it is declared to be an event (Shahid and Naqvi 2011). Some other techniques use a more intelligent measure, for instance, the quarter-sphere based techniques compare the radii of quarter-spheres of various nodes. An event is present in the region if the median deviation of data samples at all nodes is greater than the median radius.
- Clustering based techniques incorporate spatial correlations by using different cluster merging strategies. Each node of the region transmits its cluster parameters to the central node. The central node then uses various merging strategies to merge these clusters (Rajasegarar et al. 2012; Moshtaghi et al. 2011a,c). The merged cluster parameters are then transmitted back to all nodes of the region.
- Nearest neighbor based techniques do not incorporate spatial correlations.

5.4.4 Spatio-temporal-attribute correlations

As discussed above, temporal and attribute correlations play an important role in the detection of outliers at a particular node of the network, whereas, spatial correlations are helpful in determining the presence of an event in the region. An optimal outlier and event detection technique for harsh environment should incorporate all three types of data dependencies.

These dependencies when considered together are known as spatio-temporal-attribute correlations. A technique meant for outlier detection in harsh environments should exploit these dependencies in the following manner.

- Initially the technique should use temporal-attribute correlations to identify the outlier at a particular node of the network. Attribute correlations play an essential role in outlier detection and significant performance enhancement has been reported in Shahid et al. (2012a).
- Once an outlier has been determined at a particular node of the network, it should invoke a spatial consensus to determine the presence of outliers at other nodes. This process exploits spatial correlations between geographically separated nodes of the network to determine the presence of an event.

5.5 In-susceptibility to non-stationary data

An outlier detection algorithm should be able to work in an inhomogeneous environment, i.e, an environment in which the data distribution depicts spatio-temporal variations. Thus, the algorithms should also be robust to non-stationarity in an environment. Different types of non-stationarities might be shown by the data distribution in a harsh environment.

5.5.1 In-susceptibility to temporal non-stationarity OR in-homogeneity

Changes in the data distribution with time is known as temporal non-stationarity or in-homogeneity. This type of non-stationarity can be explained by the example of a WSN deployed in a desert. Such a WSN will report high values of temperature during day time followed by a sudden drop in the temperature after sunset. An outlier or event detection technique that is susceptible to temporal changes in the data may declare the drop in temperature as an outlier. Thus, the incorporation of this characteristic is essential for WSNs in harsh environments. Various methods have been used in literature to incorporate this characteristic in outlier and event detection techniques.

- Statistical based techniques are a special type of outlier and event detection techniques which determine various statistics related to the measured data samples and then use these statistics to determine outlier and events. These techniques incorporate temporal changes by updating the sufficient statistics with the arrival of every new data sample. Some examples of the statistics used by these techniques include, linear sum, linear sum of squares, mean, covariance and etc. Parametric statistical based techniques update the parameters, such as mean and standard deviation of the estimated distributions with every new data sample. A very few statistical based techniques incorporate temporal non-stationarity. One example is the technique proposed in Bettencourt et al. (2007).
- Classification based techniques are also unsupervised and non-parametric. Hence, they update their estimated normal data boundary with the arrival of every new data sample and incorporate temporal non-stationarity of data (Bahrepour et al. 2010b; Shahid et al. 2012a,b; Rajasegarar et al. 2008a, 2010a; Zhang et al. 2009a,b).
- Clustering based techniques also incorporate temporal non-stationarity as they do not require any underlying data distribution and are unsupervised. The cluster parameters

are updated with the arrival of every new data sample (Bezdek et al. 2011; Subramaniam et al. 2006; Rajasegarar et al. 2010b, 2012; Moshtaghi et al. 2011a,b,c; Bezdek et al. 2010; Suthaharan et al. 2010a,b).
- Nearest neighbor based techniques do not incorporate this feature.

5.5.2 In-susceptibility to spatial non-stationarity OR in-homogeneity

Spatial non-stationarity of data arises due to varying nature of data distributions being measured at geographically separated nodes in the network. For example, a sensor node of the network may be exposed to sunlight, whereas another node of the same network may be deployed in water. This characteristic effects the performance of event detection. A spatially changing data distribution should not point to an event until an actual event has occurred. Various methods have been used in literature to incorporate this characteristic in outlier and event detection techniques.

- Statistical and nearest neighbor based techniques do not incorporate this characteristic and cannot be used for event detection.
- Classification based techniques incorporate this characteristic by using a spatial consensus of all nodes in the neighborhood before declaring an outlier as an event (Bahrepour et al. 2010b; Shahid et al. 2012a,b; Rajasegarar et al. 2008a, 2010a; Zhang et al. 2009a,b).
- Clustering based techniques also incorporate this characteristic by using various cluster merging strategies. The clusters to be merged belong to spatially separated nodes of the network. Some of the state-of-the-art cluster merging strategies have been presented in Bezdek et al. (2011); Moshtaghi et al. (2011a,b,c)

5.6 Online data processing

The temporal and spatial variation of data distribution in a harsh environment implies that outlier detection should be performed as soon as the new data arrives. Online processing of streaming data also ensures the exploitation of temporal and attribute correlations which assist in improving the outlier detection performance. Supervised techniques first need to learn the data model, therefore it is difficult to make them online. However, unsupervised techniques learn the data distribution model as new data arrives, so outlier detection can be performed along with the newly sensed data. Semi-supervised techniques can also be extended to online techniques once they have learned the normality/abnormality model. The built model can then be updated as new data arrives. Various state-of-the-art techniques incorporate incorporate online processing of data.

- Statistical based techniques update the statistical measures associated with the data at every time instant based on the temporal correlations.
- Classification based techniques, for instance, the SVM based techniques update the radius of sphere or quarter-sphere with the arrival of every new data sample (Shahid et al. 2012a,b; Shahid and Naqvi 2011; Rajasegarar et al. 2008b, 2010a; Zhang et al. 2009a,b; Yang et al. 2008).
- Clustering based techniques update the cluster parameters like means and covariance with the arrival of every new data sample. Thus the temporal-attribute correlations are taken into account at each time instant (Rajasegarar et al. 2012; Moshtaghi et al. 2011a,c).

5.7 Independence from user-specified thresholds

Various types of user dependent data may have to be input to algorithm. The non-stationary nature of data associated with WSNs deployed in harsh environments imposes the condition of adaptive thresholds. Manual thresholds for various algorithms may not be suitable for dynamically changing data distributions. Various types of thresholds may have to be input, some of which are explained below.

- *Distance Thresholds*: Some of the algorithms depend on the user defined parameters for correct identification of outliers. Distance based algorithms defined by certain distance notions (Knorr and Ng 1998) are a typical example, as most of them depend on a user specified threshold, in terms of distance value to identify outliers.
- *Clustering based thresholds*: The clustering based algorithms require a certain deviation threshold. This type of threshold is used to compare the mahalanobis distance of the newly arrived data sample with the deviation criteria. Any data sample beyond the deviation is declared as an outlier (Rajasegarar et al. 2012; Moshtaghi et al. 2011a,c).
- *Classification based parameters*: The classification based algorithms also require some user defined parameters, for instance, a user defined regularization parameter is required in SVM based algorithms. This parameter is used to set an upper bound on the number of outliers identified by the technique (Shahid et al. 2012a,b; Shahid and Naqvi 2011; Rajasegarar et al. 2008b, 2010a; Zhang et al. 2009a,b; Yang et al. 2008) .
- *Nearest Neighbor Thresholds*: The parameter k (number of nearest neighbors) for various nearest neighbor based algorithms may have to be specified. These techniques do not incorporate a lot of characteristics of optimal outlier detection techniques for harsh environments, therefore, they have not been considered recently in the literature.

Some of the formal and earlier methods of outlier detection such as Grubbs test and tietjen moore test (Akyildiz et al. 2002) require the specification of fixed number of 'outliers to be detected', whereas Ramaswamy et al. (2000) requires the specification of an upper bound on the suspected number of outliers. In general, user specified thresholds tend to reduce the accuracy of algorithm, since a fixed threshold may not be sufficient to identify all outliers. Some algorithms require adaptive thresholds depending on the type of input data. For example, if an algorithm uses multivariate data, then a single threshold may not be applicable for all the attributes.

5.8 Local & global outlier detection

Compared to a centralized approach where all the outliers are determined at the central node, outlier detection in a distributed approach can be done at the network nodes individually as well as at the sink node. This is the concept of multi-level outlier detection (Zhang et al. 2010). In multilevel outlier detection each node can determine the outliers locally using the sensed data stream. Moreover the central node or the sink node can also perform outlier detection via a global estimation model. Depending upon the type of outliers, outlier detection techniques can be classified as local or global. A simple classification of different types of outliers is given below.

5.8.1 Local outliers OR first order outliers

The existence of these types of outliers mean that some of the observations at a sensor node are anomalous with respect to the rest of data, as shown in Fig. 4a. The Local Outliers are also

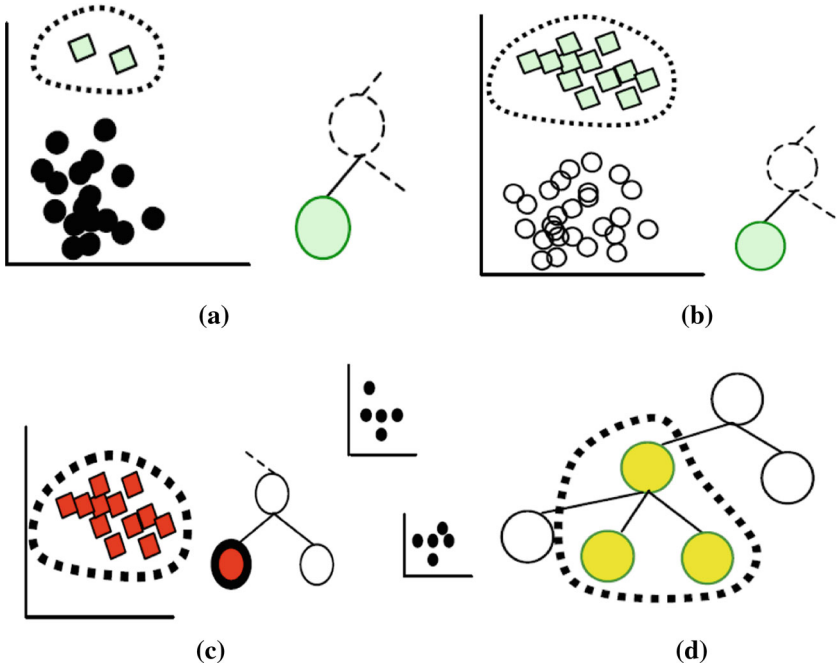

Fig. 4 **a** First order outliers. Some of the measurements are anomalous with respect to others. In the plot, *squares* represent the abnormal measurements. **b** First order epoch outliers (Type-4 local outliers). **c** Second order external outliers. All measurements of a sensor node are anomalous with respect to neighboring nodes. **d** Third order external outliers. A subset/subtree of nodes is anomalous with respect to neighboring nodes in the network (Zhang 2010)

known as First Order Outliers. The First order outliers are further classified into following categories: Type 1 or Incidental absolute errors/outliers are isolated (one-time spike) or very short sequence of extreme high or low values. For example a temperature of 0 degrees in a desert during the day time. These outliers can be identified by using a pre-defined threshold. Type 2 or Clustered Absolute Outliers are a continuous sequence of Type 1 outliers. Type 3 or Random Errors/outliers are indicated by observations not falling within the threshold of the normal data. These random errors last for a very short period of time. Type 4 or Long-Term Errors/outliers are a continuous sequence of Type 3 outliers (Zhang 2010; Ch et al. 2007). Type 4 outliers are also called First Order Epoch Outliers as shown in Fig. 4b. Here a subset of data measurements (the squares) over some contiguous time in the window differs from the general trend at the node. This could be the result of a temporary change in sensor environment (Rajasegarar et al. 2010b; Ch et al. 2007; Bezdek et al. 2010). Figure 5 shows various types of local outliers that may be encountered in a data set collected from a WSN.

Techniques for the detection of local outliers reduce the communication overhead since the data stream used to determine the outliers does not have to be communicated to the sink node. There are various versions of local outlier detection techniques

1. *Individual Approach*: Some of the techniques solely consider the determination of local outliers only, without any communication between the neighboring nodes, they, however, do consider temporal correlations (Palpanas et al. 2003).

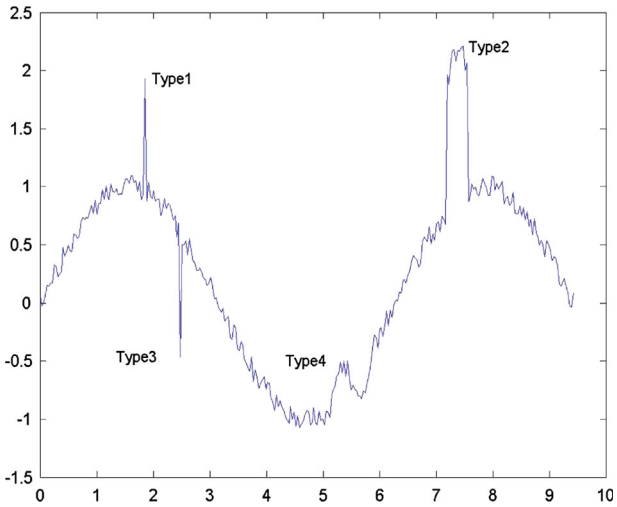

Fig. 5 Various types of outliers in a data set Zhang (2010)

2. *Aggregation Approach*:
   (a) Some techniques determine local outliers in collaboration with the neighboring nodes, like Chintalapudi and Govindan (2003). This approach requires only a small communication overhead as only a small amount of information has to be communicated to the neighboring nodes.
   (b) Some techniques consider temporal correlations as well as spatial correlation, like Yozo et al. (2004), Jun et al. (2005). This approach increases the accuracy and robustness of outlier detection technique (Zhang et al. 2010).

5.8.2 Global outliers OR higher order outliers

There are two different types of global outliers (Zhang 2010; Rajasegarar et al. 2010b; Suthaharan et al. 2010a; Bezdek et al. 2010). First, all of the data at a sensor node may be anomalous with respect to neighboring nodes. These types of outliers are called second order external outliers. In this case, a sensor node will be identified as an anomalous node as shown in Fig. 4c. Second, a set or a subtree of sensor nodes in the network may be anomalous, as shown in Fig. 4d. These are known as third order external outliers (Zhang 2010). Second and third order outliers are collectively known as Higher order (HO) external outliers (Rajasegarar et al. 2010b). Identification of global outliers can be performed at different levels in a hierarchical network, depending upon the network architecture. There are three approaches to global outlier identification.

1. *Centralized approach*: In a centralized approach, all the received data or the important parameters of data distribution are transmitted to central node that determines the global outliers. This technique provides a global perspective of the entire network to the central node (Sheng et al. 2007).
2. *Individual approach*: In some of the model based techniques, the central node, after determining the global distribution model, communicates the estimate of model to the individual nodes of the network, so that each node can determine the outliers (Rajasegarar et al. 2007).

3. *Aggregation/distributed approach*: In a distributed approach, clustering is used to identify outliers. The aggregators in the network collect data from the neighboring nodes, form a data distribution model and then identify outliers (Subramaniam et al. 2006).

The distributed approach is communication efficient as compared to the centralized approach, as the sensor nodes have limited energy and a significant fraction of available sensor power is used for communication between sensor nodes (Gupta and Kumar 2000; Shnayder et al. 2004).

5.9 Degree of outlierness

In WSNs, outliers are measured in two scales. (1) scalar and (2) outlier score (Ganguly 2008). Outlier score identifies degree to which the sensor reading deviates from the normal data.

1. *Scalar*: This outlier scale classifies a data measurement either as normal data or anomalous. This is a simple zero-one classification of data. This method neither differentiates between outliers, nor provides a ranked list of outliers. Such a decision about any data sample is also known as a 'hard decision'. Various methods proposed in the literature provide such decision about the data samples. Some of the classification and clustering based algorithms which provide a hard decision are presented in Shahid et al. (2012a,b), Shahid and Naqvi (2011), Rajasegarar et al. (2008b, 2010a, 2012), Moshtaghi et al. (2011a,c), Zhang et al. (2009a,b), Yang et al. (2008).
2. *Outlier Score*: These types of techniques not only classify a sensor reading as outlier or normal data, but also associate a score with the outlier. This score defines the degree of outlierness of the sensor measurement. This type of decision is known as 'soft decision'. Thus an analyst can pick the top k outliers based on the outlier score. Many techniques such as distance based and k-nearest neighbors associate an outlier score with data points. These techniques depend upon a threshold to identify outliers. However the threshold is user defined and not easy to choose. The optimal solution in WSNs is to constantly modify the threshold with the streaming data. The 'hard decision' based algorithms can be modified to give a 'soft decision' by associating suitable probability measures with the deviations of data samples.

A soft decision based algorithm can be used to obtain information about the degree of deviation of newly arrived data samples from normal data. This information can be helpful in generating warning alarms in a harsh environment so that appropriate actions can be taken to prevent the disaster scenario.

5.10 Summary of the characteristics

From the above discussion of the characteristics of outlier detection techniques for WSNs, we can conclude that an outlier detection technique meant for harsh environments should be unsupervised, distributed, online, threshold independent, computation and communication efficient and insusceptible to spatial and temporal non-stationarity and in-homogeneity. Moreover it should consider spatio-temporal and attribute correlations of a multivariate streaming data set and perform efficient analysis to detect and identify local/global outliers and events, with a remarkably high detection rate and a low false positive rate. Table 2 gives a summary of important characteristics of outlier detection techniques in WSNs for harsh environments.

Table 2 Description of important characteristics of outlier detection techniques for WSNs in harsh environments

	Characteristics	Description
1	Unsupervised	No requirement of normality/abnormality model
2	Multivariate data	Ability to deal with multiple attributes
3	Streaming data	Ability to deal with continuously arriving data
4	Spatial correlations	Consider the correlations between data of sensors in locality
5	Temporal correlations	Consider the time correlations of data
6	Attribute correlations	Consider the correlations between various attributes
7	Threshold independence	Does not require the specification of any thresholds, like, distance,nearest neighbor, number of outliers, etc.
8	Local outliers detection	Ability to detect the outliers locally at each node
9	Global outliers detection	Ability to detect global outliers of the network
10	Multiple outliers detection	No specification of an upper bound the number of outliers that can be detected
11	Distributed approach	Ability to detect the outliers without transmission of entire data to central node
12	Event detection	Ability to detect events of interest in the network
13	Event Identification	Ability to identify the type of event occurring in the network
14	Low computational complexity	No involvement of intensive computations
15	Low communication complexity	Minimum possible communication in between the nodes
16	Online approach	Ability to operate on streaming data
17	In-susceptibility to Spatial Non-stationarity	In-susceptibility to spatial changes in the data distribution
18	In-susceptibility to temporal non-stationarity	In-susceptibility to changing data distributions over time
19	In-susceptibility to in-homogeneity	In-susceptibility to non-homogeneous distributions at different nodes
20	In-susceptibility to dynamic topology	In-susceptibility to dynamic changes in the network topology

6 Relationship between various characteristics of outlier detection techniques for harsh environments

Various characteristics of outlier and event detection techniques for WSNs deployed in harsh environments (discussed in Sects. 4, 5) are related to each other. A summary of relationship between various characteristics is summarized below and also given in Table 3.

- An outlier and event detection technique for harsh environments should be unsupervised, i.e, it should not consider any labeled input–output data for training phase. The property of being unsupervised ensures that the technique does not require a training phase. Since the training phase is not required, so, the data can be processed in an online manner. The online processing property of the technique further makes it feasible for streaming data.

Table 3 Relationship between various characteristics of outlier detection techniques for WSNs in harsh environments

Unsupervised	Online processing of data
	Temporal correlations
	Streaming data
	In-susceptibility to temporal non-stationarity
Multi-variate data	Attribute correlations
Distributed computations	Low communication complexity
	Energy efficiency
	Local outlier detection
	Global outlier detection
	Spatial correlations
	In-susceptibility to spatial non-stationarity
	Event detection
Independence from user defined thresholds	In-susceptibility to temporal non-stationarity

The online processing also introduces the possibility of exploiting temporal correlations of the streaming data, as the statistical parameters related to the data can be updated with the arrival of every new data sample. Thus, the technique can be made in-susceptible to temporal non-stationarity of the data.

- The consideration of multi-variate data introduces the possibility of incorporating attribute correlations. The attribute correlations enhance the outlier and event detection algorithms significantly.
- Distributed computations at each node of the network reduce the communication overhead significantly and also make an outlier detection technique robust to the communication constraints posed by harsh environments. This also reduces the probability of a sensor measurement being corrupted by noise. Distributed computations at all the nodes of the network also assist in local outlier detection in the network. The information about local outliers can then be broadcasted to the central node to determine global outliers and events in the whole network. The broadcast of information between various nodes of the network can introduce the possibility of exploiting spatial correlations which can further assist in the event detection process. This also helps in making the technique in-susceptible to spatial non-stationarity.
- The independence of an outlier or event detection technique from user-defined thresholds means that it can adapt its thresholds with dynamically changing data distributions. This property ensures the in-susceptibility of the technique to temporal non-stationarity.

7 Prioritization of characteristics of outlier detection techniques for WSNs in harsh environments

The discussion in the previous section on characteristics of outlier detection techniques for harsh environments depicts that some of the characteristics are more essential for a technique to be feasible for harsh environments as compared to others. On the basis of the discussion carried out in the previous sections we propose a prioritization of characteristics of outlier detection techniques for suitability in harsh environments which is shown in Fig. 6. As evident from the figure, some of the characteristics like multivariate data, spatio-temporal and attribute correlation, Insusceptibility to temporal and spatial non-stationarity and dynamic topology changes and event identification are essential characteristics of outlier detection techniques

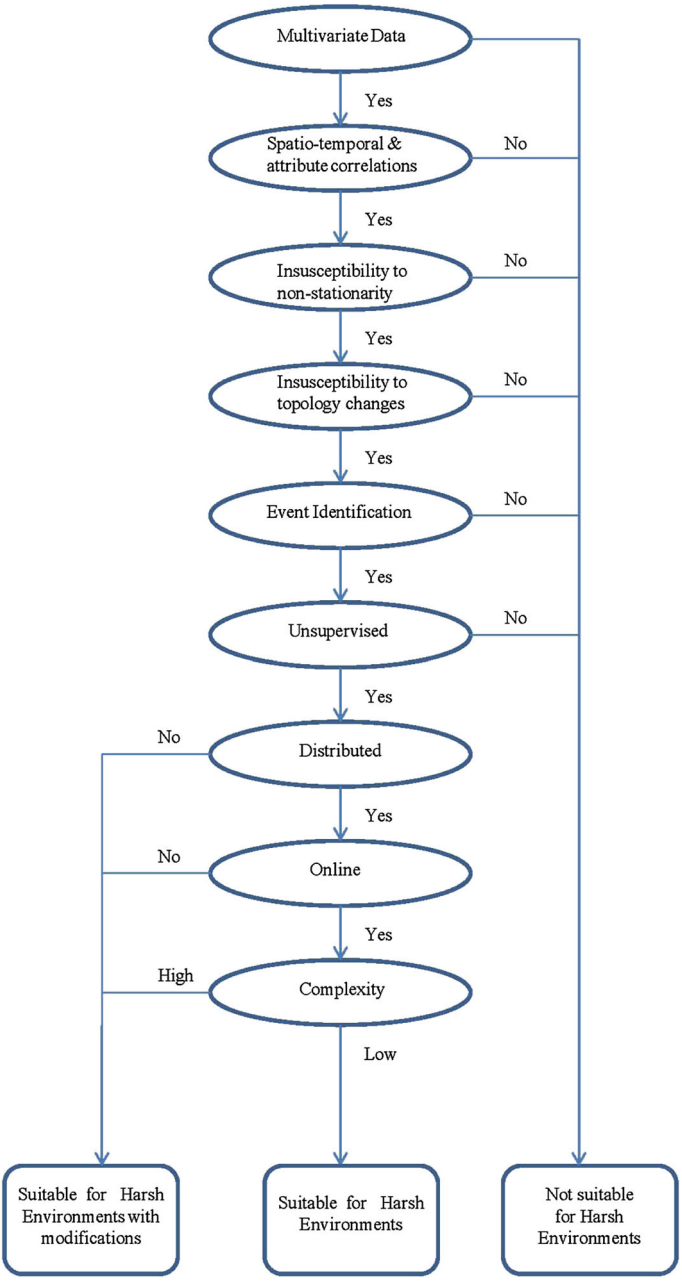

Fig. 6 Prioritization of the characteristics of outlier detection techniques for harsh environments

for harsh environments, whereas the characteristics like distributed and online approach and complexity play a secondary role to the suitability of outlier detection techniques for harsh environments.

The most important characteristic of an outlier and event detection technique for harsh environments is the consideration of *multi-variate* data. WSNs deployed in harsh environ-

ments usually measure multiple attributes, therefore, the technique for such environments should operate on multiple attributes. A technique that operates on only one of the attributes may not be able to detect outliers spread among various attributes. Thus we propose 'multi-variate' data consideration to be the most important characteristic for outlier detection in harsh environments.

The next essential characteristic for an outlier detection technique is harsh environments is the consideration of relation between various attributes, data samples at different time instants and the data of spatially separated nodes in the network. These associations are known as *attribute, temporal and spatial correlations*. Attribute and temporal correlations are extremely essential for local outlier detection at individual nodes of the network, whereas spatial correlations are essential for global outlier and event detection in harsh environments.

The consideration of spatio-temporal-attribute correlations makes it possible for an outlier detection technique to become robust to *temporal and spatial non-stationarities*. Further, it is also important to make outlier detection techniques in-susceptible to the changes in the network topology which may be caused by instability of mine structure.

If an outlier detection technique considers multi-variate data, spatio-temporal-attribute correlations and in-susceptibility to non-stationarity and dynamicity, it should be able to detect events in the environment. Therefore, *event detection* is the next important characteristic that should be possessed by a technique meant for harsh environments. Further, event type identification should also be performed by the technique along with event detection to determine the set of attributes involved in the event.

The *unsupervised* nature of a technique is the next important property that should be considered for harsh environments. Although, this property holds significant importance as it ensures that data can be processed in an online manner and temporal-attribute correlations can be exploited but some of the semi-supervised techniques, such as a few SVM based techniques have been reported to provide a high outlier and event detection performance. Semi-supervised techniques form a model of normality or abnormality and do not necessarily require a training phase like unsupervised techniques. Therefore, this property cannot be prioritized over previously mentioned characteristics.

As shown in Fig. 6 the consideration of multi-variate data, spatio-temporal-attribute correlations, in-susceptibility to temporal and spatial non-stationarity, event detection and unsupervised nature are a few characteristics which are very important for outlier detection techniques in harsh environments. Any technique that does not satisfy any of these characteristics is unsuitable for harsh environments.

Distributed computations play an important role in local and global outlier and event detection in harsh environments. However, this characteristic has been given a low priority as compared to others because this feature helps in reducing communication complexity of technique. As described in Sect. 5, local and global outliers can also be computed using centralized approaches. Thus, this feature does not directly affect the performance of a technique. Further, as evident from Fig. 6, any centralized (non-distributed) technique can also be used for outlier and event detection in harsh environments with suitable modifications.

Online data processing is also an important feature for WSNs deployed in harsh environments. However, this property has been given a low priority as batch data processing techniques may also be able to exploit temporal and spatial correlations and non-stationarities. A typical example of batch processing technique that possesses most of the characteristics for harsh environments is proposed in Rajasegarar et al. (2008b). Thus, if a technique is not online, it can still be used for outlier and event detection with modifications. Similarly,

computational and communication complexity do not affect the performance of outlier and event detection techniques. Any technique with a high complexity can be used for harsh environments with modifications.

An outlier detection technique that possesses all of the above mentioned characteristics is feasible for deployment in harsh environments without any modification. This is also shown in Fig. 6.

8 Feasibility of state-of-the-art techniques for harsh environments

State-of-the-art outlier detection techniques for WSNs have been classified into statistical based, clustering based, nearest neighbor based and classification based techniques. In this section we discuss the feasibility of these techniques for harsh environments based on the characteristics discussed in the previous section. Sections 4 and 5 incorporated a brief discussion on various types of outlier detection techniques for each of the mentioned characteristics. Thus, from the discussion carried out in the previous sections we can draw following conclusions about the feasibility of these techniques.

8.1 Feasibility of statistical & nearest neighbor based techniques

Statistical based approaches depend upon the data distribution model 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. But in real life scenarios, we do not have information about the type and distribution of sensing data, thus, these techniques are not suitable for real time systems, especially for harsh environments. Non-parametric techniques are appealing due to the fact that they do not make any assumption about the distribution characteristics, however threshold dependence of these techniques makes them inefficient. Further, only a few techniques have been proposed recently which incorporate temporal non-stationarities and online data processing strategies. The attribute correlations, which paly a vital role in outlier and event detection have not yet been introduced in these techniques. Due to certain limitations and absence of a few characteristics in these techniques (explained in Sects. 4 , 5) they cannot be considered suitable for outlier and event detection for WSNs deployed in harsh environment. Future research should focus on incorporating attribute correlations and event detection strategies in these techniques to make them feasible for harsh environments.

Nearest neighbor based approaches do not make any assumption about the data distribution and can generalize many notions from statistical based approaches. These techniques suffer from the choice of input parameters. Moreover it is difficult and computationally expensive to make distance calculations in multivariate data sets. Hence these techniques lack scalability. Optimized versions of nearest neighbor based and distance based approaches can be used for outlier detection in harsh environments because they can make the computations efficient. Moreover, from the discussion carried out in Sects. 4, 5, these techniques do not incorporate spatio-temporal-attribute correlations, insusceptibility to non-stationarity, online data processing and various other characteristics that are essential for techniques meant for harsh environments. Recent research on outlier detection techniques for WSNs has not been focussed on these techniques. Thus, nearest neighbor based techniques are unsuitable for harsh environments.

8.2 Clustering & classification based techniques

Clustering based techniques have opened a new era of research in unsupervised outlier and event detection techniques for WSNs in harsh environments. These technique are computationally efficient, require less communication, incorporate data non-stationarities, exploit spatio-temporal-attribute coorrelations to detect local and global outliers as well as events in a region. These techniques can also perform online processing on the streaming sensor data. Similarly, classification based techniques are also unsupervised or semi-supervised and possess a number of characteristics discussed in Sects. 4, 5. On the basis of discussion carried out in Sect. 5, it can be concluded that these techniques are the most feasible techniques for outlier and event detection in WSNs deployed in harsh environments.

9 Open research areas

The discussions carried out in this paper concludes that clustering and classification based techniques are most suitable for outlier detection in WSNs deployed in harsh environments, whereas statistical and nearest neighbor are infeasible. The clustering and classification based techniques, although suitable for our applications, still carry a lot of research potential.

- Both the clustering and classification based techniques have not been studied for event type identification. This involves the identification of key attributes involved in an event, so that a manual analysis does not have to be performed. Clustering based techniques provide more scope in this context and various computationally efficient modifications can be incorporated in them. Specifically, the 'mahalanobis distance' measure used in these techniques can play a vital role in this process.
- The mahalanobis distance measure used in clustering based techniques for outlier detection can only be used for multi-variate normal data sets. Suitable distance measures should be formulated for these techniques which can incorporate deviation measures for data sets beyond normal distributions (Ekström 2011).
- Clustering and classification based techniques do not provide a soft decision for outliers in a data set, i.e, they do not provide a ranked list of outliers based on the degree of deviation. These techniques can easily be modified for this purpose by associating a certain probability measure with the deviation of outliers from normal data.
- Clustering based techniques have been mostly explored for outlier detection. However, their use for event detection is limited and can be focussed on by future researchers. These techniques should be modified to incorporate robust event detection strategies which can clearly differentiate between an event and a change in data distribution.
- Although significant work has been focussed in recent years to reduce the computational complexity of classification based techniques but future research should focus more in this domain. Certain techniques need to be developed which can detect outliers and events in an online manner without solving complex optimization problems at each time instant. A simpler method for this is to introduce suitable distance or deviation measures based on kernel functions used for classification (Somorjai et al. 2011).

10 Conclusion

In this paper we presented various characteristics of state-of-the-art outlier detection techniques for WSNs, essential for operation in harsh environments. We have developed a

mechanism to identify what features an outlier detection algorithm must incorporate in such environments. Some of these characteristics include input data type, spatio-temporal and attribute correlations, outlier types(local and global), type of approach (distributed or centralized), outlier identification (event or error), outlier degree, susceptibility to dynamic topology, non-stationarity and inhomogeneity etc. To the best of our knowledge, these aspects, in the domain of harsh environments have never been discussed before. Flow charts summarizing all the discussions on the characteristics, including the prioritization of various characteristics of outlier detection techniques for WSNs, in harsh environments has also been presented. Moreover, the feasibility of various types of outlier detection techniques have been discussed for harsh environments.

References

Abe S (2010) Support vector machines for pattern classification. Springer, New York

Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) Wireless sensor networks: a survey. Comput Netw 38:393–422

Akyildiz IF, Akan zgr B, Akan OB, Chen C, Fang J, Su W (2003) Interplanetary internet: state-of-the-art and research challenges. Comput Netw 43:75–112

Aly M (2005) Survey on multiclass classification methods. Neural Netw 1–9

Bahrepour M, Meratnia N, Havinga PJM (2008) Automatic fire detection: a survey from wireless sensor network perspective. Centre for Telematics and Information Technology University of Twente, Enschede, technical report TR-CTIT-08-73, Dec 2008. http://eprints.eemcs.utwente.nl/14624/

Bahrepour M, Meratnia N, Havinga PJM (2009a) Sensor fusion-based event detection in wireless sensor networks. In: SensorFusion, Toronto, Canada. IEEE, Los Alamitos, pp 1–8

Bahrepour M, Meratnia N, Havinga PJM (2009b) Use of ai techniques for residential fire detection in wireless sensor networks. In: AIAI 2009 workshop proceedings, Greece, vol 475, July 2009, pp 311–321. ceur-ws.org

Bahrepour M, Zhang Y, Meratnia N, Havinga PJM (2009c) Use of event detection approaches for outlier detection in wireless sensor networks. In: Proceedings of symposium on theoretical and practical aspects of large-scale wireless sensor networks, the 5th international conference on intelligent sensors, sensor networks and information processing 2009 (ISSNIP 2009), Melbourne, Australia. IEEE Press, Victoria, Dec 2009, pp 439–444

Bahrepour M, Meratnia N, Havinga PJM (2010a) Fast and accurate residential fire detection using wireless sensor networks. Environ Eng Manag J 9(2):215–221

Bahrepour M, Meratnia N, Poel M, Taghikhaki Z, Havinga PJM (2010b) Distributed event detection in wireless sensor networks for disaster management. In: International conference on intelligent networking and collaborative systems, INCoS 2010, Thessaloniki, Greece. IEEE Computer Society, USA, pp 507–512

Bahrepour M, van der Zwaag BJ, Meratnia N, Havinga P JM (2010c) Fire data analysis and feature reduction using computational intelligence methods. In: Phillips-Wren G, Jain LC, Nakamatsu K (eds) Proceedings of the second KES international symposium on advances in intelligent decision technologies, IDT 2010, Baltimore, Maryland, USA, series smart innovation, systems and technologies, vol 4. Springer, Berlin/Heidelberg, July 2010, pp 289–298

Barnett V, Lewis T (1994) Outliers in statistical data. Wiley, Lonon

Bettencourt LMA, Hagberg AA, Larkey LB (2007) Separating the wheat from the chaff: practical anomaly detection schemes in ecological applications of distributed sensor networks. In: Computing distributed in sensor systems (DCOSS 2007), Santa Fe, NM, USA, June 2007, pp 223–239

Bezdek J, Havens T, Keller J, Leckie C, Park L, Palaniswami M, Rajasegarar S (2010) Clustering elliptical anomalies in sensor networks. In: 2010 IEEE international conference on fuzzy systems (FUZZ), pp 1–8

Bezdek J, Rajasegarar S, Moshtaghi M, Leckie C, Palaniswami M, Havens T (2011) Anomaly detection in environmental monitoring networks [application notes]. Comput Intell Mag IEEE 6(2):52–58

Bhuse V, Gupta A (2006) Anomaly intrusion detection in wireless sensor networks. J High Speed Netw 15:33–51

Branch J, Szymanski B, Giannella C, Wolff R, and Kargupta H (2006) In-network outlier detection in wireless sensor networks. In: 26th IEEE international conference on distributed computing systems, 2006. ICDCS 2006, p 51

Cardell-Olivera R, Kranza M, Smettemb K, Mayerc K (2005) A reactive soil moisture sensor network: design and field evaluation. Int J Distrib Sens Netw 1(2):149–162
Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41:15:1–15:58. doi:10.1145/1541880.1541882
Ch V, Banerjee A, Kumar V, Chandola V (2007) Outlier detection: a survey
Chen J, Kher S, Somani A (2006) Distributed fault detection of wireless sensor networks. In: Proceedings of the 2006 workshop on dependability issues in wireless ad hoc networks and sensor networks, series DIWANS '06. ACM, New York, NY, pp 65–72. doi:10.1145/1160972.1160985
Chintalapudi K, Govindan R (2003) Localized edge detection in sensor fields. In: Proceedings of the first IEEE 2003 international workshop on sensor network protocols and applications, May 2003, pp 59–70
da Silva APR, Martins MHT, Rocha BPS, Loureiro AAF, Ruiz LB, Wong HC (2005) Decentralized intrusion detection in wireless sensor networks. In Proceedings of the 1st ACM international workshop on quality of service & security in wireless and mobile networks, series Q2SWinet '05. ACM, New York, NY, pp 16–23. doi:10.1145/1089761.1089765
Dario IA, Akyildiz IF, Pompili D, Melodia T (2005) Underwater acoustic sensor networks: research challenges. Ad Hoc Netw 3:257–279
Dereszynski E, Dietterich T (2011) Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns. ACM Trans Sens Netw 8(1):3
Ding M, Cheng X (2009) Robust event boundary detection in sensor networks—a mixture model based approach. In: IEEE INFOCOM 2009, April 2009, pp 2991–2995
Ding M, Chen D, Xing K, Cheng X (2005) Localized fault-tolerant event boundary detection in sensor networks. In: Proceedings IEEE of 24th annual joint conference of the IEEE computer and communications societies INFOCOM 2005, vol 2, pp 902–913
Ekström J (2011) Mahalanobis distance beyond normal distributions. UCLA Stat (preprint)
Elnahrawy E, Nath B (2004) Context-aware sensors. In: European workshop on wireless sensor, networks, pp 77–93
Ganguly AR (2008) Knowledge discovery from sensor data. CRC Press, Boca Raton
Garca-Hernndez CF, Ibargengoytia-Gonzlez PH, Garca-Hernndez J, PrezDaz JA (2004) Wireless sensor networks and applications: a survey. Int J Comput Sci Netw Secur 7(3):264–273
George S, Zhou W, Chenji H, Won M, Lee Y, Pazarloglou A, Stoleru R, Barooah P (2010) Distressnet: a wireless ad hoc and sensor network architecture for situation management in disaster response. IEEE Commun Mag 48(3):128–136
Giatrakos N, Kotidis Y, Deligiannakis A (2010a) Pao: power-efficient attribution of outliers in wireless sensor networks. In: Proceedings of the seventh international workshop on data management for sensor networks. ACM, pp 33–38
Giatrakos N, Kotidis Y, Deligiannakis A, Vassalos V, Theodoridis Y (2010b) Taco: tunable approximate computation of outliers in wireless sensor networks. In: Proceedings of the 2010 international conference on management of data. ACM, pp 279–290
Gomez-Verdejo V, Arenas-Garcia J, Lazaro-Gredilla M, Navia-Vazquez A (2011) Adaptive one-class support vector machine. IEEE Trans Signal Process 59(6):2975–2981
Gupta P, Kumar P (2000) The capacity of wireless networks. IEEE Trans Inf Theory 46(2):388–404
Han J, Kamber M (2006) Data mining: concepts and techniques. Morgan Kaufmann, Los Altos
Hao P, Chiang J, Lin Y (2009) A new maximal-margin spherical-structured multi-class support vector machine. Appl Intell 30(2):98–111
Hassan A, et al (2011) A heuristic approach for sensor network outlier detection. Int J Res Rev Wirel Sens Netw 1(4):66–72
Hill DJ, Minsker BS, Amir E (2007) Real-time bayesian anomaly detection for environmental sensor data. In: Proceedings of the 32nd conference of IAHR, 2007
Hodge VJ, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 22:85–126. doi:10.1007/s10462-004-4304-y
http://connekt.seecs.nust.edu.pk/SAHSE.php
http://www.genuki.org.uk/big/eng/LAN/Haydock/WoodPitExplosion.html
http://www.humanite.fr/2006-03-10_Societe_-Catastrophe-de-Courrieres-une-expression-impropre
http://www.msha.gov/MSHAINFO/FactSheets/MSHAFCT8.HTM
Janakiram D, Adi Mallikarjuna Reddy V, Phani Kumar A (2006) Outlier detection in wireless sensor networks using bayesian belief networks. In: Communication system software and middleware, 2006. Comsware 2006, pp 1–6
John GH (1995) Robust decision trees: removing outliers from databases. In: In knowledge discovery and data mining. AAAI Press, Menlo Park, pp 174–179
Jun MC, Jeong H, Kuo C-CJ (2005) Distributed spatio-temporal outlier detection in sensor networks

Keally M, Zhou G, Xing G (2010) Watchdog: confident event detection in heterogeneous sensor networks. In: 2010 16th IEEE on real-time and embedded technology and applications symposium (RTAS), pp 279–288
Keerthi S, Sundararajan S, Chang K, Hsieh C, Lin C (2008) A sequential dual method for large scale multi-class linear svms. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 408–416
Knorr EM, Ng RT (1988) Algorithms for mining distance-based outliers in large datasets, pp 392–403
Krishnamachari B, Iyengar S (2004) Distributed bayesian algorithms for fault-tolerant event region detection in wireless sensor networks. IEEE Trans Comput 53(3):241–250
Lazarevic A, Ozgur A, Ertoz L, Srivastava J, Kumar V (2003) A comparative study of anomaly detection schemes in network intrusion detection. In: Proceedings of the third SIAM international conference on data mining
Liu S, Liu Y, Wang B (2007) An improved hyper-sphere support vector machine. In: Third international conference on natural computation, 2007. ICNC 2007, vol 1. IEEE, pp 497–500
Liu C, Yang Y, Tang C (2010) An improved method for multi-class support vector machines. In: 2010 International conference on measuring technology and mechatronics automation (ICMTMA), vol 1, pp 504–508
Luo X, Dong M, Huang Y (2006) On distributed fault-tolerant detection in wireless sensor networks. IEEE Trans Comput 55(1):58–70
Madden S, Franklin MJ, Hellerstein JM, Hong W (2002) Tag: a tiny aggregation service for ad-hoc sensor networks. In: IN OSDI, 2002
Mainwaring A, Culler D, Polastre J, Szewczyk R, Anderson J (2002) Wireless sensor networks for habitat monitoring. In: Proceedings of the 1st ACM international workshop on wireless sensor networks and applications, series WSNA '02. ACM, New York, NY, pp 88–97. doi:10.1145/570738.570751
Misra P, Kanhere S, Ostry D, Jha S (2010) Safety assurance and rescue communication systems in high-stress environments: a mining case study. Commun Mag IEEE 48(4):66–73
Moshtaghi M, Havens T, Bezdek J, Park L, Leckie C, Rajasegarar S, Keller J, Palaniswami M (2011a) Clustering ellipses for anomaly detection. Pattern Recog 44(1):55–69
Moshtaghi M, Leckie C, Karunasekera S, Bezdek J, Rajasegarar S, Palaniswami M (2011b) Incremental elliptical boundary estimation for anomaly detection in wireless sensor networks. In: 2011 IEEE 11th international conference on data mining (ICDM), pp 467–476
Moshtaghi M, Rajasegarar S, Leckie C, Karunasekera S (2011c) An efficient hyperellipsoidal clustering algorithm for resource-constrained environments. Pattern Recog 44:2197–2209
Ni L, Liu Y, Lau YC, Patil A (2003) Landmarc: indoor location sensing using active rfid. In: Proceedings of the first IEEE international conference on pervasive computing and communications, 2003 (PerCom 2003), March 2003, pp 407–415
Ozdemir S, Xiao Y (2011) Outlier detection based fault tolerant data aggregation for wireless sensor networks. In: 2011 5th IEEE international conference on application of information and communication technologies, pp 1–5
Palpanas T, Papadopoulos D, Kalogeraki V, Gunopulos D (2003) Distributed deviation detection in sensor networks. SIGMOD Rec 32:77–82. doi:10.1145/959060.959074
Phua C, Lee V, Smith K, Gayler R (2010) A comprehensive survey of data mining-based fraud detection research. Arxiv, preprint arXiv:1009.6119
Rajasegarar S, Leckie C, Palaniswami M, Bezdek JC (2006) Distributed anomaly detection in wireless sensor networks. In: 10th IEEE Singapore international conference on communication systems, 2006. ICCS 2006, Oct 2006, pp 1–5
Rajasegarar S, Leckie C, Palaniswami M, Bezdek J (2007) Quarter sphere based distributed anomaly detection in wireless sensor networks. In: IEEE international conference on communications. ICC '07, June 2007, pp 3864–3869
Rajasegarar S, Leckie C, Palaniswami M (2008a) Anomaly detection in wireless sensor networks. IEEE Wirel Commun 15(4):34–40
Rajasegarar S, Leckie C, Palaniswami M (2008b) Cesvm: centered hyperellipsoidal support vector machine based anomaly detection. In: IEEE international conference on communications, 2008. ICC '08, May 2008, pp 1610–1614
Rajasegarar S, Leckie C, Bezdek J, Palaniswami M (2010a) Centered hyperspherical and hyperellipsoidal one-class support vector machines for anomaly detection in sensor networks. IEEE Trans Inf Forensic Secur 5(3):518–533
Rajasegarar S, Bezdek JC, Leckie C, Palaniswami M (2010b) Elliptical anomalies in wireless sensor networks. ACM Trans Sens Netw 6:7:1–7:28 [Online].10.1145/1653760.1653767
Rajasegarar S, Bezdek J, Moshtaghi M, Leckie C, Havens T, Palaniswami M (2012) Measures for clustering and anomaly detection in sets of higher dimensional ellipsoids. In: The 2012 international joint conference on IEEE in neural networks (IJCNN), pp 1–8

Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large data sets. SIGMOD Rec 29:427–438. doi:10.1145/335191.335437
Ross G, Tasoulis D, Adams N (2009) Online annotation and prediction for regime switching data streams. In: Proceedings of the 2009 ACM symposium on applied computing, pp 1501–1505
Rousseeuw P, Leroy A (1996) Robust regression and outlier detection. Wiley, London
Schieferdecker D, Völker M, Wagner D (2011) Efficient algorithms for distributed detection of holes and boundaries in wireless networks. Exp Algorithm 6630:388–399
Shahid N, Naqvi IH (2011) Energy efficient outlier detection in wsns based on temporal and attribute correlations. In: International conference on emerging technologies, 2011
Shahid N, Naqvi IH, Qaisar SB (2012a) Quarter-sphere SVM: attribute and spatio-temporal correlations based outlier & event detection in wireless sensor networks. In: 2012 IEEE wireless communications and networking conference: mobile and wireless networks (IEEE WCNC 2012 track 3 mobile & wireless), France, Paris
Shahid N, Naqvi IH, Qaisar SB (2012b) Real time energy efficient approach to outlier & event detection in wireless sensor networks. In: 13th IEEE international conference on communication systems 2012 (IEEE ICCS'12), Singapore, Singapore
Sharma A, Golubchik L, Govindan R (2010) Sensor faults: detection methods and prevalence in real-world datasets. ACM Trans Sens Netw 6(3):23
Sheng B, Li Q, Mao W, jin W (2007) Outlier detection in sensor networks
Shnayder V, Hempstead M, rong Chen B, Allen GW, Welsh M (2004) Simulating the power consumption of large-scale sensor network applications. In: In Sensys. ACM Press, pp 188–200
Somorjai R, Dolenko B, Nikulin A, Roberson W, Thiessen N (2011) Class proximity measures-dissimilarity-based classification and display of high-dimensional data. J Biomed Inf 44(5):775–788
Subramaniam S, Palpanas T, Papadopoulos D, Kalogeraki V, Gunopulos D (2006) Online outlier detection in sensor data using non-parametric models. In: Proceedings of the 32nd international conference on very large data bases, series VLDB '06. VLDB endowment, pp 187–198. Available on http://portal.acm.org/citation.cfm?id=1182635.1164145
Suthaharan S, Alzahrani M, Rajasegarar S, Leckie C, Palaniswami M (2010a) Labelled data collection for anomaly detection in wireless sensor networks. In: 2010 sixth international conference on intelligent sensors, sensor networks and information processing (ISSNIP), Dec 2010, pp 269–274
Suthaharan S, Leckie C, Moshtaghi M, Karunasekera S, Rajasegarar S (2010b) Sensor data boundary estimation for anomaly detection in wireless sensor networks. In: 2010 IEEE 7th international conference on mobile adhoc and sensor systems (MASS), pp 546–551
Tan P, Steinback M, Kumar V (2006) Introduction to data mining. Addison Wesley, Reading
Tax DMJ, Duin RPW (1999) Data domain description using support vectors. In: ESANN'99, pp 251–256
Tutorial on wireless communications and electronic tracking, 2009
Wang D, Yeung DS, Tsang ECC (2006) Structured one-class classification. IEEE Trans Syst Man Cybern Part B Cybern 36(6):1283–1295
Wu W, Cheng X, Ding M, Xing K, Liu F, Deng P (2007) Localized outlying and boundary data detection in sensor networks. IEEE Trans Knowl Data Eng 19(8):1145–1157
Xu T (2009) A new sphere-structure multi-class classifier. In: Pacific-Asia conference on circuits, communications and systems, 2009. PACCS'09. IEEE, pp 520–525
Xu T, He D, Luo Y (2007) A new orientation for multi-class svm. In: Eighth ACIS international conference on software engineering, artificial intelligence, networking, and parallel/distributed computing, 2007. SNPD 2007, vol 3. IEEE, pp 899–904
Xue W, Luo Q, Chen L, Liu Y (2006) Contour map matching for event detection in sensor networks. In: Proceedings of the 2006 ACM SIGMOD international conference on management of data, series SIGMOD '06. ACM, New York, NY, 2006, pp 145–156. doi:10.1145/1142473.1142491
Yang Z, Meratnia N, Havinga P (2008) An online outlier detection technique for wireless sensor networks using unsupervised quarter-sphere support vector machine. In: International conference on intelligent sensors, sensor networks and information processing, 2008. ISSNIP 2008, pp 151–156
Yozo CP, Hida Y, Huang P, Nishtala R (2004) Aggregation query under uncertainty in sensor networks, technical report
Zhang Y (2010) Observing the unobservable—distributed online outlier detection in wireless sensor networks. Ph.D. dissertation, University of Twente
Zhang Y, Meratnia N, Havinga PJM (2007a) A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets. Centre for Telematics and Information Technology University of Twente, Enschede, technical report TR-CTIT-07-79, Nov 2007. http://eprints.eemcs.utwente.nl/11366/
Zhang K, Shi S, Gao H, Li J, (2007b) Unsupervised outlier detection in sensor networks using aggregation tree. In: Proceedings of the 3rd international conference on advanced data mining and applications,

series ADMA '07. Springer, Berlin/Heidelberg, pp 158–169. [Online]. Available http://dx.doi.org/10.1007/978-3-540-73871-8_16

Zhang Y, Meratnia N, Havinga P (2009a) Adaptive and online one-class support vector machine-based outlier detection techniques for wireless sensor networks. In: Proceedings of international conference on advanced information networking and applications workshops WAINA '09, pp 990–995

Zhang Y, Meratnia N, Havinga PJM (2009b) Hyperellipsoidal svm-based outlier detection technique for geosensor networks. In: Third international conference on geosensor networks, Oxford, UK, series lecture notes in computer science, vol 5659. Springer, Berlin, July 2009, pp 31–41

Zhang Y, Meratnia N, Havinga P (2010) Outlier detection techniques for wireless sensor networks: a survey. IEEE Commun Surv Tutor 12(2):159–170

Zhang Y, Hamm NAS, Meratnia N, Stein A, van de Voort M, Havinga PJM (2012) Statistics-based outlier detection for wireless sensor networks. Int J Geogr Inf Sci 26(8):1373–1392

Zhuang Y, Chen L (2006) In-network outlier cleaning for data collection in sensor networks. In: In CleanDB, workshop in VLDB. APPENDIX 2006, pp 41–48

Zoumboulakis M, Roussos G (2007) Escalation: complex event detection in wireless sensor networks. Smart Sens Context 4793:270–285