International Conference on Automation, Control, Engineering and Computer Science (ACECS'14)
Proceedings - Copyright IPCO-2014 ISSN 2356-5608, pp.37-46

Anomaly detection in WSN: critical study with new vision

Aymen ABID[#1], Abdennaceur KACHOURI [#2], Adel MAHFOUDHI[#3]
[1#] *CES laboratory, ENIS, Sfax University, Tunisia*
[2#] *ISSIG, University of Gabes, Tunisia*
[3#] *CCIT, Taif University, Taif, Saudi Arabia*
[1] aymen.abid.mail@gmail.com
[2] abdennaceur.kachouri@enis.rnu.tn
[3] a.mahfoudhi@tu.edu.sa

ABSTRACT

In Wireless Sensor Networks "WSN", intrusion, reliability of links or sensors, energy and other challenges have a serious impact about security and assurance of good information (Quality of Information) e.g. for making decision. In fact, WSN is a system that collects data of events or alerts from sensors. Therefore, to have a good decision and functioning we can analyze data to detect anomalies (intrusions, luck of energy, congestion, bad connections…) or events.

The aim of this study is to investigate on the detection of abnormalities in WSN and these components via catches being anomalies or events. With new vision, we define the anomaly problem and it detection locally in WSN based on the captured data. Also, we localize the problem in a wider area, "reliability engineering", that aims to have a safe, secure, available and maintainable system. The paper will present different techniques used for anomaly-treatment and a comparative study.

Keywords: WSN, Reliability Engineering, Information Assurance, Data Analysis, Anomaly Detection, Abnormal Behavior.

I. INTRODUCTION

The wireless sensor networks (WSN) are used in many fields such as medical, economic applications (in agriculture, industry, services, etc...), military, environmental, home automation... (Makhoul, 2008)

From disciplines that use anomaly detection approach, we find statistic, data mining, artificial intelligence (AI), machine leaning, information theory and spectral decomposition (Zhang, 2010). In fact, it's very used in many applications e.g. fraud detection, network intrusion, performance analysis, weather prediction…

Intrusion detection system (IDS) is an important component of computer security that in reality detect a kind of anomaly. In military forest fire applications, SNode are deployed in unprotected and open environment. As result, SNode are exposed to several kinds of attacks, from physical to transport layer. For example, in data link layer, the attackers try to exhaust battery by repetition of unusual retransmission. It can be used in network by comparing the current state with the normal comportment (Kumari, 2013). The challenge in (Chen,2007) is "How to detect anomaly intrusion, modeled as malicious behavior, in critical conditions?". In this work, the technical aim is to have a fast detection of known attack with decentralized implementation and to get a high ratio accuracy of network error.

The objective of use the detection method in (Suzuki,2009) is to have solutions for the difficulty to collect information after disaster, e.g. in underground malls, in order to reduce the damage caused by disaster (e.g. Kawata 1995).

As solution, (Suzuki, 2009) propose a "Robot Sensor Network System" (RSNS) using a "high mobility rescue robot with WSN".

(Nagajothi, 2012) have the goal to determinate the best path and source redundancy levels to satisfy QoS while maximizing the MTTF (Mean Time to failure), using fault tolerant QoS control algorithm.

Often, the sensor networks have a major role in decision making and significantly influence on the behavior of the system user. As a result, the poor state of sensor nodes (SNodes) will have adverse effects on the diagnosis given to the user.

This, it brings us to monitor the behavior of sensors based primarily on data they collect because this information is the axial component of a system delivered by the network of Wireless Sensors Networks to a user.

The rest of paper is organized as follows: abnormality, event and anomaly, metrics and criteria are the subject of the first section. The second presents a survey of anomaly detection techniques, completed by a comparison between some methods in the last section. Conclusion will discuss this paper and what can be done in the future.

II. DETECTION BASED ON ABNORMAL DATA IN WSN

An abnormality in a data set may be defined as a case that seems inconsistent with the rest of the data set. It is "an abnormal feature, characteristic, or occurrence" (Oxford dictionary). So the suspected data collected in a space covered by the WSN over time, is a description of a behavior that is abnormal and not standardized (scheme model, usually ...). Owing to this, in most time, abnormal data reflects:

- Spatial behavior
- Temporal behavior
- Spatial-temporal behavior

In reality, abnormal data detection in most works is called "outlier detection". In fact, a system for detecting outlier values "Outlier Detection System ODS" that we can also call it deviation detection or data cleaning procedure, is an analyze process to identify a specific data from the rest of sensed data (Ben-Gal, 2005). So thus, we notice that normal/abnormal detection (outlier) classifies data to:

- Normal/anomalous for ADS (Anomaly Detection System)
- Normal/events for EDS (Event Detection System).

1. Outlier data: anomaly and event in WSN

Outliers are patterns in data that don't conform to a well-defined notion of normal behavior (Mallick, 2009). So, the detection of anomaly can be done by the detection of data deviation from normal behavior, and data participated in this deviation are called outliers. Anomalies are considered as observations that do not correspond to a well-defined normal behavior concept. While, events are new observations that can provide a natural but unexpected behavior transformed to new models or classes (Markou, 2003a).

In literature, there is no a very discrimination between anomaly detection and outlier detection terms in WSN. Well, the term "anomaly" can signify "outlier". Yet, ADS specify the first handle of outlier detection based on data collected in WSN. The second handle is event detection because the data presenting an event can be viewed as outlier data. In other manner, events are one of the causes of outliers e.g. in spatiotemporal correlation, noisy measurements and sensor faults haven't a spatial relation but event measurements are spatially correlated. In fact, recent works try to resolve anomaly and event data simultaneously (Fawzy, 2013).

(Ghaddar, 2010) and (Lim, 2010) say that ADS includes fault detection (noise, errors…), event detection (fire, explosion, movement…) and intrusion detection (malicious attacks...). But with reliability engineering vision, an outlier data may be event or error. In this context, it's very clear that in reliability domain, anomaly detection based on collected data in WSN can be viewed as error detection through data analysis. In another way, at least, this detection based on data is a common point between the two areas outlier detection and reliability engineering. In fact, error detection in general is a technic from others used in reliability engineering to discover the existence of an error (incorrect state) especially for fault tolerance e.g. test programs algorithms (with Z/B-Language or other), likelihood control etc.. (Abid, 2010).

Reliability engineering (said also safety engineering) of a computer system is the property that allows users to place a justified confidence in the service delivered to them (Arlat, 1995). It is the failure science that (Villemeur, 1988) has defined as the ability of an entity from one device to fulfill a number of functions where required under given conditions.

So reliability engineering is the system quality centered on the concept of fault as a potential cause of a malfunctioning in computer system. The fault will cause an error that gives rise to a failure (Bennani, 2005).This quality is measured by many attributes which we quote: Reliability, Availability, Maintainability, Safety, and Security (RAMSS). It is ensured by avoidance, tolerance, prediction and elimination of failures, which are the means of safety (Arlat, 2006). These means can be used in the design phase (pre-means) or built to be used during operation of the system (post-means) (Abid, 2010).

We conclude that for ODS, anomaly in detection procedure is manifested with erroneous data. In the other case, the outlier is confirmed as an event. The error is caused by a fault (intrusion, physical handicap, network congestion or sensor ...) explained as a failure (bad decision, quality…). Also, an internal erroneous state of the system (sensor or network) can cause a failure.

The fault is identified by its nature (accidental, intentional ...), its phenomenological causes (physical faults, faults caused by man ...), the boundary of the system (internal fault or external fault.) and its occurrence phase (in the design or operation) and persistence time (temporary or permanent faults). According to their origin, nature and temporal persistence, faults can be classified into three main types: design, physical and interaction fault (Arlat, 2006).

Then, ADS cans be achieved by a diagnostic to define the origin fault in failure or before in error case. So we can say that ADS with outlier detection technics aims to filter noisy data, find faulty nodes (Fawzy, 2013). This kind of method is beneficial to ameliorate Quality of Information (QoI) (Ghaddar, 2010; Kumar, 2013) and so Quality of Service (QoS) (ABID, 2013a).

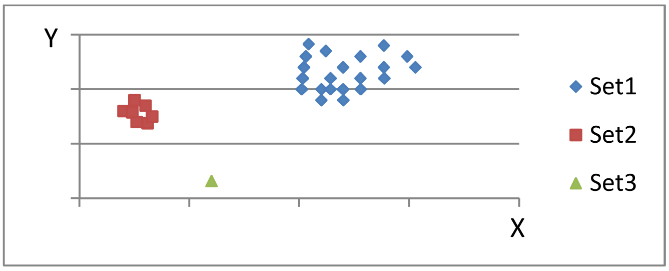

Fig.1 A 2-dimensinal repartition of values

In **figure1**, set1 is normal data set but other points are abnormal: set2 is outlier that can be a new event or also anomalous and data in set3 is anomalous (suspected to be erroneous).

2. Event detection

WSN encourages applications of event detection in many field such as monitoring hazardous gases, nuclear power, the amount of pollution in the evacuated area ... Indeed, EDS may be into three categories: the threshold detection method, the model testing method and reasoning testing method (Jing, 2013). The event can be temperature, pressure, visual, acoustic… that the detector must find it from unobserved pattern.

EDS is a system that used to detects events and principally that are not declared before. In fact, novelty detector is an outlier detector that searches the novel events. It's aimed to find new event or data. For (Tarassenko, 2009), event detection is focused in novelty detection.

Novelty can be referred by the probability of analyzing data that does not belong to the normal distribution data. This is usually generated from a distribution of underlying data estimated from sample data (Ghaddar, 2011).

For (Emmanouilidis, 2010) the detection of new events is a critical step for monitoring. The problem of data maintenance is the subject of several levels in a monitoring system status. The higher level processes the data offline, while the lower level it analyzes the collection of samples by the online sensors. In all cases the outlier state of data defined as the deviation from a model of known behavior.

3. Anomaly detection

Anomalies are viewed as observations that do not correspond to a well-defined notion of normal behaviors (Lim, 2010). Automatically, for data analysis, when we say "abnormal data" we also think to "normal data". In this context, a lot of ADS build models from normal data used to define those who are not, and even build their models too (Ghaddar, 2010).

In WSN, anomalies can be classified into three broad categories depending on nature: point, contextual and collective anomalies (Mallick, 2009; Chandola, 2009). First, if an instance-specific data can be considered abnormal compared to the rest of the data, the instance is named as a point anomaly. Second, if a data instance is anomalous within a particular context but not otherwise, then it is termed as a contextual anomaly called also conditional anomaly. E.g. a temperature time series show the monthly temperature of an area for one year. A temperature of 10° might be normal during the winter normal, but the same value during summer would be contextually abnormal. Third, if a collection of related data instances is anomalous with respect to the entire data set, it is appointed as a collective anomaly. For example, in a human electrocardiogram output, an area indicates a problem because the very low value exists for an unusually long time despite the low value in itself is not an anomaly.

However, data anomalies can be classified in three types: temporal, spatial and spatial temporal. For the first, we study the single data instance. For the second, the behavior context is focused. For the last, is the case of a collection of data does not conform to entire set of data then it is known as collective anomalies (Sahni, 2013).

In accordance to the source of the anomalies, it exist three levels:

- Node anomaly,
- Network anomaly,
- Data anomaly.

In data level, if the readings of neighboring sensors are not really compatible, then there is an anomaly to be treated. The low cost of sensors is the principal origin of data anomalies.

Even intrusion attacks may be easier in "WSN" than other networks, because of the small existing default protection in sensors. This is classified in anomaly-nodes level. In network level, wireless physical communication medium is clashed to noisy environment, e.g. applications for natural disasters. Also, malicious routing attacks are classified in this last level, e.g. flooding, selecting forwarding, sinkhole attack, Sybil attack, wormhole attack…

4. Performance Metric For ODS Evaluation

The efficiency of data analysis techniques via detection abnormality is usually evaluated by their ability to differentiate abnormal behaviors from normal. The most metric used to evaluate experiences are “Detection Rate”, “False Alarm Rate” and “False Positive Rate”. (Chitradevi, 2013) have a good definition for this metric but as more other propositions it’s explained adapting with their solution. Referenced to this last work, we define this following three metrics for efficiency detection with more generally manner:

- **Detection Rate (DR)**: is the ratio between number of correctly detected outliers and the total number of real outliers.
- **False Alarm Rate (FAR):** is the ratio between numbers of normal data (SNode) declared abnormal and the total number of real outliers.
- **False Positive Rate (FPR):** is the ratio between the total number of abnormal declared normal and the total number of normal measurements (data).

5. Conception criteria of ODS for WSN

Concerning the input sensor data, data viewed as streams, can determine the technique to use it whereas two aspects: attributes and correlations. The Data can be single or multivariate attributes. Correlation is the SNode dependency among other elements e.g. this SNode data, data of other SNodes, personal or neighbor history... The second main criterion that we canvass is the type of outlier detection if it’s a local or global. In local detection, SNode detect its outlier without the help of other. In global outlier, the SNode or a set of SNode or a supervisor detect outlier from data stream (Zhang, 2010). Another criterion defines the degree of being outlier. This degree is mentioned as a scalar or score manner. The scalar scale gives only sets of outlier and normal measurements. The score scale fixe one or many threshold to decide the score and the state of each measurement. Availability of pre-defined data is used to construct a normal pattern in order to detect abnormal behavior. This is generally required in the beginning of the ODS work, that we call ‘test phase’.

Lastly, identity of outlier is a criterion that defines the source of the outlier. Outlier source is an event or an anomaly. The cause of anomaly may be a noise in the node or in the network, a malicious attack… (Zhang, 2010; Abid, 2010). The cause of the event will be in most of time, a sudden change in environmental parameters (accident…) or new homogenous event (**figure2**).

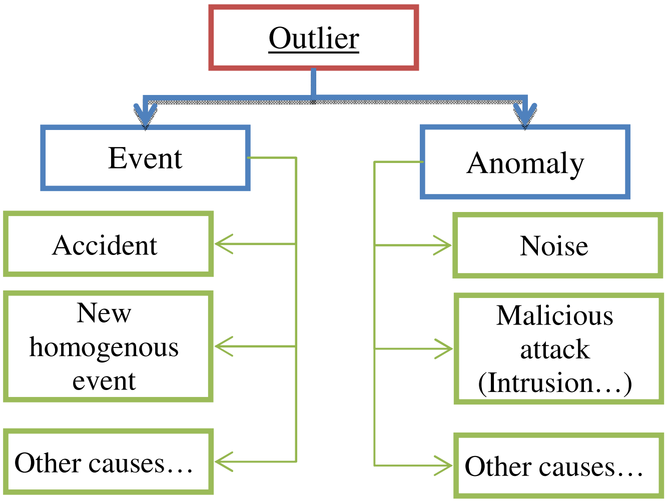

Fig.2 Outlier identities

III. Anomaly Detection techniques

(Lim, 2010) organize the detection operation in four stages: establishing the detection model, deploying the detection agent, performing the detection and declaring the anomalies. At first stage, the choice of learning approach is depends on the availability of data. As said before, supervised approach learns and constructs the normal and abnormal models using a pre-labeled data. Semi-supervised anomaly detection uses normal pre-labeled data to build the normal profile of the systems. Unsupervised anomaly detection does not require any pre-labeled data available to train the system. At second stage, the three different approaches can be used; centralized, hybrid and distributed.

At performing anomaly detection stage, the five mentioned technique above are used. Declaring anomalies stages is provided by decision algorithms.

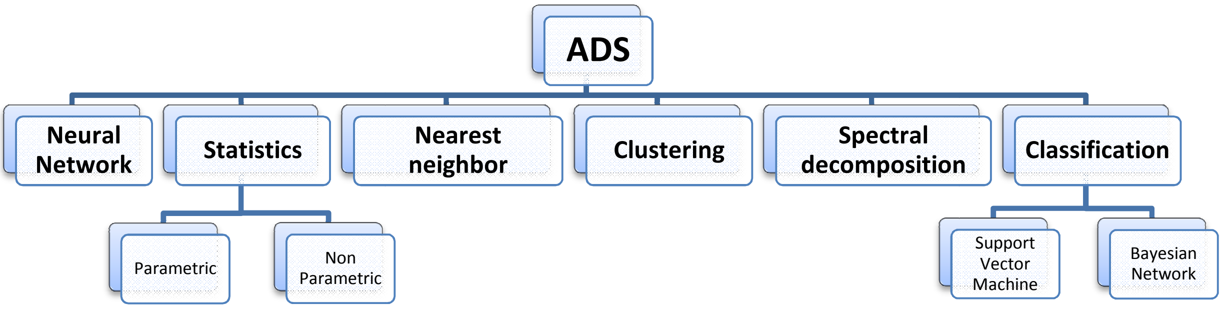

Fig.3 Organization of anomaly detection techniques for WSN

In fact, a comparison of various anomaly detection algorithms is developed in (Lim, 2010) based on anomaly type (data, network…), the data attributes (single or multiple in which a combination between attributes can be established to detect anomalies), the learning model if it exist and the manner with it's made (online or offline), the signature update (detection model is static or dynamic) and finally based on algorithm and architecture used (KNN, clustering… with centralization …).

We notice that's possible to combine techniques e.g. statistical techniques are used after a test phase with neural network.

Whilst technique of outlier detection is not intended only to anomalies (ADS), there are persons made their classification with "aberrant" vision. In this case the two terms have the same meaning e.g. (Zhang, 2010) classifies outlier to Noise & error, Events and malicious attacks, for (Fawzy, 2013) outlier has the source error or event, (Bahrepour, 2009) use the event detection approaches for outlier detection etc… But other they do it and the classification was in anomaly context (Lim, 2010; Chandola, 2009). The next improving of (Zhang, 2010) ADS classification" is that which may also be adopted for ODS and for which we add Neural Network branch.

So, in accordance with (Berkhin,2006), we support the same classification of clustering algorithms that are with more detail in this work (figure3). But, we add new branch, the Neural Network approaches.

1. Neural Network Techniques

Neural Network methods (NN) have more constraints and difficulties to use it as ODS specially for ADS (Markou, 2003a). We cite, it needs more computational resources for test phase and it will be more required for retraining of test phase. Automating the detection remains a limitation for neuron as well as statistics. Alternatively, the neural network alone is not the solution for this problem. Also, declaring of new class (so new event or errors) is also an obstacle for NN. As statistical, NN uses borders decision (hyper-planes, thresholds...) to separate normal and abnormal and to define new sets (classes, clusters...).

As said, it will be very expensive to reconstruct the NN after a detection of new set and results not are evident, even so it exist works that aims reduce the complexity of this retraining e.g. constructive neural networks. In fact, they reconfigure weights using some techniques as the cascade correlation. Also, the still problem is to connect new elements presenting to new sets without losing the historic and knowledge.

In this paragraph, we present methods that are in more detail in (Markou, 2003a). First, Multi-Layer Perceptron "MLP" is a famous approach that generally uses NN. The limitation of MLP is to define sets limits that will badly about detection. The testing phase in the NN of MLP aims to suggest a good the regression-approximation by minimizing the sum of squared errors, defined on a finite set of data. Second, an auto-association is an approach that reconstruct output same as input. The main method was Principal Component Analysis (PCA). PCA can be linear with the decomposition in eigenvector of covariance or correlation matrix from data. Also, PCA can be non-linear with multi-layer perceptron architecture for configuration functions of feedback. The detection is done when it cannot establish the input at output. Finally, Self-Organizing Map (SOM) is an unsupervised approach, alternative to statistical clustering of data with a threshold used for new detections.

2. Statistical Techniques

Statistical-based approach generally methods can be divided into two branches "parametric based approach" in which we have knowledge of the available data and "Non-parametric based approach" for which we don't know the availability of data distribution. It is a method relied on statistical modeling. The statistical techniques for ADS through collected data are an approach to detect outliers relies on this assumption: "Normal data instances occur in high probability regions of a stochastic model, while outliers occur in the low probability regions of the stochastic model". An "ADS" with this technique is defined with this principle: "An anomaly is an observation which is suspected of being partially or wholly irrelevant because it is not generated by the stochastic model assumed" (Mallick, 2009).

So, the good realization of the type of technique is related to the good check of the assumption above. The anomaly score is associated with a confidence interval that is additional information for decision-making for a test instance. Another advantage is for a good outliers distribution estimation step in data, this approach operable in unsupervised frame without labeled training data.

However, it's not always true that we have a particular distribution, e.g. high dimensional real data sets. In addition, it's not easy to choose the best statistic tool. The main challenge for this technique is if we need to have a relation and interaction between data e.g. multidimensional data with histogram approach. We notice that in the literature, there are no many people who speak and work with Stats as a pure technique; this may be because of its greedy use in the definition of algorithms based on other techniques. But exist some pure statistical techniques for ODS such as Gaussian Model Based, Regression Model Based, Histogram Based, Kernel Function Based.

In the following two paragraphs, we present some statistical methods that are in more detail in (Markou, 2003b).

The mainly preset distribution used in statistical modeling is the Gaussian form. GMM "Gaussian mixture modeling", a parametric approach, is a general distribution of modeling and estimation for density by maximizing the log likelihood. For that, optimization algorithms are required e.g. conjugate gradients or re-estimation techniques (Expectation-Maximization Algorithm EMA or other algorithm). Per contra, GMM is also helpless against multidimensional data that need a huge number of testing attributes sequence.

Parzen window method is non-parametric data density estimator. In this method, the outlier threshold is a distinct determination set. It is used in unconditional probability p(x) of a test pattern "x" based on the modeled distribution.

(Ghaddar, 2010) try to detect temporal abnormal behavior using predictive time-series based on autoregressive models (AR). Of course, futures reading of nodes are predicted from history of sensed values. ADS compare current value "Nt" with predicted value "Xt". The failure will be declared after ǀNt − Xtǀ exceeds a certain threshold. For evaluation, they generate random anomalous data in $\left[\bar{\mu} \pm {}^{\sigma}/_{2\sqrt{n}}\right]$ to have lowest confidence level, with: "σ" the population standard deviation and "n" the population size. Analyses-results were with Detection Rate (DR: 78% to 90%), False Positive Rate (FPR: 5% to 11%) False Alarm Rate (FAR: 8% to 17%).

3. Nearest neighbor techniques

It's an approach that analyzes a data instance with respect to its K nearest neighbors. For this reason, most procedures drafted with this idea are called "K-NN algorithms ". The distance notions for nearest is computed as distance (e.g. Euclidean) between devices or similarity measure between two data instance. So, a K-NN can be "a distance based-data instance" or "a relative density based-data instance". This technique for ADS is based on this assumption: "Normal data instances occur in dense neighborhoods, while outliers occur far from their closest neighbors" (Mallick, 2009). This last makes a study for a simple algorithm using the technique of distance for SNode with neighbors (K-NN). For him, k-nearest neighbor's distance of object (SNode) "Xp" can be defined as:

$$\overline{\mathbf{d_{Xp}}} := {}^{\mathbf{1}}/_{\mathbf{K}} \sum_{\mathbf{X_i \in N_p}} \mathbf{dist}(\mathbf{X_i}, \mathbf{X_p}) \qquad (1)$$

The internal distance of k-nearest neighbor of Xp is:

$$\overline{\mathbf{D_{Xp}}} := {}^{\mathbf{1}}/_{\mathbf{K(k-1)}} \sum_{\mathbf{X_i, X_j \in N_p, i \neq j}} \mathbf{dist}(\mathbf{X_i}, \mathbf{X_j}) \qquad (2)$$

Therefore, local distance-based outlier factor (LDOF) uses the relative position of an object to its neighbors to indicate the deviation degree of the object from its neighborhood system:

$$\mathbf{LDOF_k}(\mathbf{X_p}) := {}^{\overline{\mathbf{d_{Xp}}}}\big/_{\overline{\mathbf{D_{Xp}}}} \qquad (3)$$

In (Chitradevi, 2013), they addresses data reliability as an input issue seeing that in WSN data integrity is affected by the harsh environmental conditions. This causes outlier sequences that this work try to detect by two density-based outlier detection techniques for discover local outliers using k-distance neighborhood based local outlier factor (LOF) formulation: DBOD_MSS and DBOD_PMSS. Although the time complexity is the same as DBOD_PMSS DBOD_MSS, DBOD_PMSS manages detection of misrepresentation in the distribution of sensor data in both dense and sparse data while DBOD_MSS supports outlier detection in dense distribution only. This solution is unsupervised technique that pursues to calculate the degree of measurement to being an outlier (LOF). A value is reported as an outlier if its "LOF" is significantly higher compared to its local neighbors.

Their solution aims to minimize the computational time for LOF. Simulation is with real database and they demonstrate that DBOD_MSS have less LOF time but DBOD_PMSS is more robust by successfully notification of outlier appearing as small groups.

4. Clustering techniques

Clustering is a technique for grouping similar objects in which each group is called a cluster throughout their similar behavior. This grouping is not monitored by hidden data concept learning. Clustering has many advantages, which we quote: easy adaptation and integration of new elements, no supervision requirement, easy arrangement for ADS for different technics, and fast testing phase mainly for small number of cluster (Fawzy, 2013).

Also, this technique can operate in an unsupervised mode. However, it is one of the energy saving techniques to prolong the durability and extensibility of a sensor network especially with a dynamic network analysis.

However, performance of clustering based techniques is highly dependent on the efficiency of clustering algorithm in capturing the cluster structure of normal instances. Also, many ADS don't optimize anomalies detected using the clustering. The clustering can assign all instances into groups that will be give false classification for specific instance e.g. point anomaly. In addition, the computational complexity for data clustering is always high in order of O (N^2d).

Clustering has always been used in statistics. For example, in the first hand the machine learning clustering algorithms are implemented into image segmentation and computer viewing, in the second hand clustering can be considered as a problem of density estimation that is the purpose of the traditional multivariate statistical estimation (Berkhin,2006). Several clustering algorithms do not force every data instance to belong to a cluster e.g. DBSCAN, ROCK, SNN clustering, Self-Organizing Maps (SOM) and K-means clustering (Mallick, 2009).

In centralized outlier detection mechanism, clustering for outlier detection are performed after collection of data in BS or gateway. In distributed mechanism (in-network mechanism), clustering algorithm folds inside the network. Each SNode performs the clustering algorithm. After that, the parent combines their result with other. Anyway, the evolution of centralized and distributed methods can be done in the gateway.

While most studies are based on mathematical tools, (Dutta, 2013) use a fuzzy method for clustering in WSN with multi-hop protocol named “EEDS”. The cluster heads (CH) are dynamically selected with reasonable transmission energy. They reveal that by choosing SNode with more residual energy aids in optimal energy consumption to extend the network lifetime. However, they consider also the network traffic state and distance between SNodes to don't congest network by "CH" choice based on energy only. In fact, the likelihood of a SNode to be a "CH" increases with the increase in the number of neighboring nodes and the battery power, and declines with increasing of distance from the center of gravity of the cluster.

A comparison simulation with F3N and power constrained LEACH is performed and they found that "EEDS" performs better than the others. They aim in their future works to implement fuzzy logic algorithms for variables that allow the "CH" to be more sparsely distributed.

In our work (ABID, 2013a), we aim to ameliorate the quality of service QoS (availability and user confidence) and quality of information QoI (exactitude). This is with a data vote system for failure detection that can be viewed as clustering method.

5. Spectral decomposition techniques

The spectral theorem has developed in functional analysis and linear algebra based on the "eigenvalues" and "eigenvectors" in a square matrix.

The spectral theorem usually requires an algebraic formalism that is simpler than others (Cauchy...). The formalism is for forms (Quadratic, Hermitian ...), endomorphism and matrices.

The spectral theory has been success thanks to the use of his theorem in functional analysis giving conditions for expression of an operator in simpler sum.

The spectral decomposition is based on the theory and the spectral theorem. Indeed, the decomposition may be the factorization of a positive definite (Economics, 2014) or positive real matrix “A” (Dasgupta,2008), that can be written as follows according to the spectral theorem:

$$\boldsymbol{A = C\,L\,C^T} \qquad (4)$$

With:

C: Eigenvector matrix
L: Eigenvalues diagonal matrix

According to the spectral theorem, this matrix "A" may be formalized in a sum of products as following:

$$\boldsymbol{A = \sum_{i=1}^{i=n} L_i C_i C_i^T} \qquad (5)$$

With:

“i" is the ith column of C, L, CT

(Xie, 2011) propose a “Principal Component Analysis PCA” that use the “spectral decomposition” in order to justify the reduction of data dimension. The requirement was SNodes fail to support high dimensional training data in anomaly detection systems “ADS”. So, they aim to reduce data representation by finding a new Q-dimensional< P-dimensional in order to optimize the compilation. In fact, they select the first Q principal component (PC) according to:

$$\sum_{i=1}^{Q} \lambda_i \Big/ \sum_{i=1}^{P} \lambda_i \geq t \qquad (6)$$

Where:

“t” is a predefined threshold, e.g. t = 80%.

The division was defined using the “Spectral decomposition theorem” and it quantifies the cumulative proportion of the variance explained by the first “Q-PCs”. This first "PC" is only used to represent the data source as it contains most of the variance in the data. The simulation was under distrusted manner using “Intel Berkley Research Lab” data base (Madden, 2013). A same “unsupervised anomaly detection algorithm (UAD)” was used with and without this PCA to evaluate this new data decomposition based on FPR.

For them, the degradation between 10% and 20% in performance may be tolerable for the application favoring energy.

6. Classification techniques

Data classification is a popular data mining technique used to predict group membership for data tuples. The aim of classification is to discover a relationship between input attributes and output class. As technique, it is generally based on testing phase to learn a model from a set of labeled data. This principle is the same used in ODS based on this technique; training then testing phase. The primary assumption for ADS is: « A classifier that can distinguish between normal and anomalous classes can be learnt in the given feature space » (Mallick, 2009).

This technique is used to build many ADS e.g. neural networks based, Bayesian networks based, Support vector machine based, Rule based techniques.

In fact, testing phase of techniques using this method is fast since each test instance needs to be compared against the recomputed model. The multi-class techniques are making use of advanced algorithms which can distinguish between instances members of different classes.

However, this multiclass depends on availability of accurate labels for various normal classes that’s not always the case. Also, the risk of the anomaly presence in the test instance will have bad impact since a label will be assigned to this anomalous instance.

Approaches based on the classification, mainly used in data mining, can be categorized as two groups according to their approach “the Support Vector Machines” and “Bayesian Network” (Sahni, 2013).

Support vector machine (SVM) is a classification technique that configures a hyper plane in order to assign data to different classes. SVM is good as it doesn't require an explicit statistical model, it can offer an optimum solution for grouping by optimizing the margin of the decision boundary and it doesn't cope to multidimensional attributes (Bahrepour, 2009).

For homogenous WSN with SNode time synchronization, one-class quarter-sphere SVM-based technique can be proved; it's a spatiotemporal correlation-based technique (Markou, 2003a).

Naïve Bayes is widely used for element tabulation. It is called Naïve because of its emphasis on independency of the input data (Bahrepour, 2009). For (Gangrade, 2012), the uses case was the classification of Bayesian approach. Bayesian classifier is a statistical classifier based on "Bayes Theorem".

The technic used is Naïve Bayes Classifier for partitioned data. Naïve Bayes classifier is a simple Bayesian classifier that able to have a performance comparison with decision tree and selected neural network classifier. As assumption, it has the "class conditional independence": the effect of an attribute value on a given class is independent of the values of the other attributes.

The probability notation and Bayes theorem supported by (Gangrade, 2012) is:

- **The data tuple "X":**
  - For Bayesian: X is the evidence.
  - For WSN usage: measurement or a set of n attributes (an observed data tuple X)
- **P (H/X):** The objective of classification problem is always to determinate P (H/X). P (H/X) is a probability that the hypothesis H holds given the "evidence" (observed data tuple X); The posterior probability of H conditioned with X.
- **Bayes theorem:**

$$P(A\backslash B) = \frac{P(B\backslash A)P(A)}{P(B)} \qquad (7)$$

As Bayesian classification technique, (Gangrade, 2012) aims to present a classifier protocol without assumption except the final classifier or model parameters. As solution, a three-layer privacy preserving Naïve Bayes classifier that proposes a new protocol to calculate model parameters for horizontally partitioned databases is developed.

A comparison between horizontally and vertically partitioned data is done in (Gangrade, 2013). They analyze the performance of their two privacy preserving Naïve Bayes classifiers for distributed databases ("NBC: Naïve Bayes classifier" VS "3LPPHPNBC: 3-Layer Privacy Preserving Horizontally Partitioned NBC" and "NBC" VS "3LPPVPNBC"). The experiment was done with 2GB RAM processor having 500GB hard disk using open-source software Net-Beans IDE (version 6.9). Net-Beans IDE supports development of all Java applications and integrated these algorithms into Weka version 3.6. Weka is a data mining tool that is used to perform various data mining algorithms. the term of this comparison is correctly classified test data (Accuracy), Execution Time (for calculating model parameters, using Multi-party or Two-Party…).

Our recent work (Abid, 2013b), try to use Bayesian Network "BN" to auto configure the limits of the set of valid values in WSN failure detection. Values existing outside this set are suspected to be erroneous and their sensors are declared failed. A comparison is done with other work (ABID, 2013a) based on FAR metric.

IV. Comparative study between a set of approaches

We quote at the end a comparison between some work based on the accuracy (FAR, DR ...), databases and test data type, the field test and the characteristics and detection procedure (TABLE I).

In general, it exist works that treat events and anomaly (presented by erroneous data) together, and offers techniques to detect events as outliers (Bahrepour, 2009) and other are focused in detection of event (Jing, 2013) or anomaly.

There is research that tries to compare some approaches by using the same conditions. In fact, (Bahrepour, 2009) gives a comparison between some methods of classification based on Bayes (Naïve Bayes, the Fusion based approach using Naïve Bayes) or SVM (Support Vector Machine technology based) and the neural network (feed forward neural network FFNN, The Fusion based approach using FFNN).

In the table, we present the method "Neural Network Intrusion Detector NNID" (Ryan, 1998), "SmartSifter SS" (Yamanishi, 2000), "Global k-nearest-neighbor Global-KNN" and "Global nearest-neighbor Global-NN" (Branch, 2013), "In-network Knowledge Discovery approach IKD" (Fawzy, 2013), "Distance Based Anomaly Detection DB-AD" (Xie, 2011) and Quarter-Sphere SVM (Bahrepour, 2009).

V. Conclusion

WSN is mainly used in environment monitoring and event detection. In WSN, Anomaly Detection Systems "ADS" is mainly used to declare abnormal behaviors of data or activities that we should prepare reactions against it (such as be removed) as this behavior can have a bad impact about the job of WSN.

This paper address anomaly data detection system (ADS) in reliability engineering in order to build a monitor system controlling sensors state, based primarily on captured data and SNodes characteristics (energy, transmission-capacity ...). The goal is to avoid bad decisions, increase service quality measured by attributes of reliability engineering (accuracy, confidence, consistency ...) and better react against network anomalies. The basic principle is to analyze multidimensional collected data, according to time and space.

This investigation and shortcomings in existing works calls to give answers to the following questions. The first question is how we will monitor the sensors and detect their failures? And most importantly, how we will distinguish between events (new events…) and anomalies (intrusions, physical errors ...)? The second is how we will react against these faults? The third is how a solution it will be compared with the existing and what criteria will judge performance? With only general metric e.g. complexity, FAR…? Or using other? In the end, for evaluation, what is the type of error that we will consider and how we will introduce errors and new events in order to evaluate the system? We will use a synthetic base or natural base or pseudo-synthetic e.g. natural event base in that we insert errors?

TABLE I. COMPARATIVE TABLE AMONG ANOMALY DETECTION APPROACHES

Approach [Technique]	**Data Base**	**Area**	**Detection characteristics**	**Detection principle**	**Accuracy**
Neural Network Intrusion Detector [Neural Network] (Ryan, 1998)	Network commands of 100 usage pattern (e-mails, web…)	Security of computer networks	• Supervised and Centralized • Offline monitoring system preference	• Collecting training data • Training neural network to identify the user • Signal anomaly for any no validation of the user.	DR : 96% FAR : 7%
SmartSifter [Statistical] (Yamanishi, 2000)	Positive real data	Medical pathology	• Online unsupervised • Parametric complexity: $O(d^2K)$ • Non parametric complexity: $O(d^2K^2)$	• Detection based on unsupervised learning of information source with "k" Gaussian distribution	DR: 5% to 98% FPR: able to detect 82% of intrusions
Global-KNN & Global-NN [Nearest neighbor] (Branch, 2013)	Environmental phenomena from (Madden, 2013) and synthetic dataset	Localization using time difference arrival (TDOA) of sound to a sensor	• Unsupervised • Energy consummation: under 3%	• A sensor pi detects an event using: initializations, the set of points Di of pi change, transmission of single packets M to neighbors… • Correctness algorithms required for some case.	DR: 99%
In-network Knowledge Discovery [clustering] (Fawzy, 2013)	Environmental phenomena from (Madden, 2013) and synthetic dataset	Not specified	• Time execution: 10ms for 65000 epochs.	• Clustering data to groups with nearest neighbor classification. • Detect normal cluster and outlier cluster. • Classify the degree of outlier value (error or event)	DR: 100% FAR: 0.02% to 0.1%
Distance Based Anomaly Detection [Spectral Decomposition] (Xie, 2011)	Two subsets randomly picked from (Madden, 2013) with injection of 100 anomalous data generated by normal distribution	Misbehaviors (cyber-attacks, sensor faults…)	• Hierarchical Network and distributed detection • Unsupervised using Principal Component Analysis (PCA)	• Distributed normalization of values • Distributed PCA • Distributed detection with distance consideration from CH (cluster head) • MN (member node) detect if a data is normal or anomalous.	DR: 80% to 98% FPR: 20% to 30%
Quarter-Sphere SVM [Classification] (Bahrepour, 2009)	Environmental data in Grand-St-Bernard, Switzerland (LCAV, 2013) & synthetic data	harsh deployment area	• The use of Polynomial kernel function • Complexity: O(m×p); m is number of features, p is number of data vectors (number of classes).	• Each node learns the local radius "R" of the quarter sphere using its m measurements • Find a minimal R • Define R with neighbors and decide outliers	• Real base: DR : 98.05% FPR : 1.24% • Synthetic : DR: 98.53% FPR : 1.58%

VI. References

[1] Makhoul, A. (2008). Réseaux de capteurs: localisation, couverture et fusion de données (Doctoral dissertation).

[2] Ben-Gal, I. (2005). Outlier Detection in Maimon, O. and Rockach, L.(eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers.

[3] Mallick, S. (2009). A Project Report on Outlier Data Detection in Wireless Sensor Network (Doctoral dissertation, JADAVPUR UNIVERSITY).

[4] Markou, M., & Singh, S. (2003a). Novelty detection: a review—part 2: neural network based approaches. Signal processing, 83(12), 2499-2521.

[5] Fawzy, A., Mokhtar, H. M., & Hegazy, O. (2013). Outliers detection and classification in wireless sensor networks. Egyptian Informatics Journal, 14(2), 157-164.

[6] Berkhin, P. (2006). A survey of clustering data mining techniques. In Grouping multidimensional data (pp. 25-71). Springer Berlin Heidelberg.

[7] Chitradevi, N., Palanisamy, V., Baskaran, K., & Swathithya, K. (2013). Efficient Density Based Techniques for Anomalous Data Detection in Wireless Sensor Networks. Journal of Applied Science and Engineering, 16(2), 211r223.

[8] Zhang, Y., Meratnia, N., & Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. Communications Surveys & Tutorials, IEEE, 12(2), 159-170.

[9] Tarassenko, L., Clifton, D. A., Bannister, P. R., King, S., & King, D. (2009). Novelty Detection. Encyclopedia of Structural Health Monitoring. ISBN: 978-0-470-05822-0.

[10] Jing, Z. H. O. U. (2013). A Kind of Space-Time Event Detection Approach for Wireless Sensor Network.

[11] Emmanouilidis, C., & Pistofidis, P. (2010). Design requirements for wireless sensor-based novelty detection in machinery condition monitoring. In Engineering Asset Lifecycle Management (pp. 420-429). Springer London.

[12] Abid, A. (2010). Sûreté de fonctionnement par traitement de défaillance dans les réseaux de capteurs, NTSID Master, ENIS-Sfax University, Tunisia.

[13] Kumari, K., Devaraju, H., & Kumar, Y. R. (2013). A Novel Agent Based Intrusion Detection System.

[14] Chen, H., Han, P., Zhou, X., & Gao, C. (2007). Lightweight anomaly intrusion detection in wireless sensor networks. In Intelligence and Security Informatics (pp. 105-116). Springer Berlin Heidelberg.

[15] Suzuki, T. (2009). Sensor Network Deployment by Dropping and Throwing Sensor Nodes to Gather Information in Underground Spaces in a Post-Disaster Environment.

[16] Nagajothi, N., Lakshmi, A. S., Vimaladevi, V., & Rajeshwari, A. (2012). Maximizing the Lifetime of Query Based Wireless Sensor Networks Using Ftqc Algorithm.

[17] Lim, T. H. (2010). Detecting anomalies in Wireless Sensor Networks. Qualifying Dissertation, University of York.

[18] Ghaddar, A. (2011). Improving the quality of aggregation using data analysis in WSNs (Doctoral dissertation, Lille 1).

[19] Kumar, B., Rani, S., & Singh, P. (2013). A Critical Study Of Existing Approaches Based On Quality Of Information Attributes And Metrics In Wireless Sensor Network. International Journal of Data & Network Security, 4(1), 152-161.

[20] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 15.

[21] Sahni, G., & Sharma, S. (2013). Study of Various Anomalies and Anomaly Detection Methodologies in Wireless Sensor Network. International Journal, 3(5).

[22] Arlat, J., & Blanquart, J. P. (1995). A. costes, Y. Crouzet, Y. Deswarte, JC Fabre, H. Guillermin, M. Kaâniche, K. Kanoun, C. Mazet, D. Powell, C. Rabéjac, P. Thevenot, sous la direction de JC Laprie-Guide de la sûreté de fonctionnement, CEPADUES Editions, Toulouse.

[23] Villemeur, A. (1988). Sûreté de fonctionnement des systèmes industriels, collection de la Direction des Études et Recherche d'Électricité de France. Eyrolles, ISBN-10, 2(615.8).

[24] Bennani, T. (2005). Tolérance aux fautes dans les systèmes répartis à base d'intergiciels réflexifs standards (Doctoral dissertation, INSA de Toulouse).

[25] Arlat, J., Crouzet, Y., Deswarte, Y., Fabre, J. C., Laprie, J. C., & Powell, D. (2006). Tolérance aux fautes. Encyclopédie de l'informatique et des systèmes d'information. Vuibert, Paris, France, 92.

[26] ABID, A., KAANICHE, H., KACHOURI, A., & ABID, M. (2013a). Quality of service in Wireless Sensor Networks through a failure-detector with Voting Mechanism. IEEE, doi 10.1109/ICCAT.2013.6522037.

[27] Bahrepour, M., Zhang, Y., Meratnia, N., & Havinga, P. J. (2009, December). Use of event detection approaches for outlier detection in wireless sensor networks. In Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2009 5th International Conference on (pp. 439-444). IEEE.

[28] Markou, M., & Singh, S. (2003b). Novelty detection: a review—part 1: statistical approaches. Signal processing, 83(12), 2481-2497.

[29] Dutta, P. K., Naskar, M. K., & Mishra, O. P. (2013). Impact of two-level fuzzy cluster head selection model for wireless sensor network: An Energy efficient approach in remote monitoring scenarios. arXiv preprint arXiv:1308.0690.

[30] Economics (2014), About.com, Retrieved March2014 http://economics.about.com/od/economicsglossary/g/spectral.htm

[31] Dasgupta, S. (2008). Course notes, CSE 291: Topics in unsupervised learning.

[32] Xie, M., Han, S., & Tian, B. (2011, November). Highly efficient distance-based anomaly detection through univariate with PCA in wireless sensor networks. In Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on (pp. 564-571). IEEE.

[33] Madden, S. (2013). Intel Berkeley Research Lab, Retrieved October 20, 2013, http://db.lcs.mit.edu/labdata/labdata.html

[34] Gangrade, A., & Patel,(2012) R. Privacy Preserving Naïve Bayes Classifier for Horizontally Distribution Scenario Using Un-trusted Third Party. IOSR Journal of Computer Engineering (IOSRJCE) ISSN, 2278-0661.

[35] Abid, A. Maalej, M. A., Kachouri, A., Mahfoudhi, A., (2013b) A Bayesian Network for an auto configuration limits of valid values in WSN failure detection, first international conference on Reasoning and Optimization in Information Systems ROIS.

[36] Gangrade, A., & Patel, R. (2013). Performance Analysis of Privacy Preserving Naïve Bayes Classifiers for Distributed Databases. International Journal of Computer Science Issues (IJCSI), 10(2).

[37] LCAV (2013), Audiovisual Communications Laboratory - LCAV, Retrieved September 4, 2013, http://lcav.epfl.ch/page-86035-en.html

[38] Ryan, J., Lin, M. J., & Miikkulainen, R. (1998, July). Intrusion detection with neural networks. In Advances in neural information processing systems (pp. 943-949). MORGAN KAUFMANN PUBLISHERS.

[39] Yamanishi, K., Takeuchi, J. I., Williams, G., & Milne, P. (2000, August). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 320-324). ACM.

[40] Branch, J. W., Giannella, C., Szymanski, B., Wolff, R., & Kargupta, H. (2013). In-network outlier detection in wireless sensor networks. Knowledge and information systems, 34(1), 23-54.