Outlier Detection Techniques in WSNs A Survey

Mehwish Raza, Umm-e-Farwa

***Abstract* — WSNs carries great potential in research and has colossal impact on emerging field of data communication networks, artificial intelligence and information technology. They are used in several critical applications like remote patient monitoring system, military surveillance, radiation monitoring, smart agriculture, fire detection etc., where decision making process has high dependency on the quality of data acquired from WSN. However, the raw data collected from sensors is highly vulnerable to noise while unusual real-time events can be easily subject to malicious attacks. To resolve this, the node central to the system must implement outlier detection algorithms for smooth system progress. Data classification becomes mandatory to prevent illogical behavior of the system so that techniques like data mining and machine learning can play key role in WSN improvements. It is necessary to examine data for outliers before analyzing and making decisions, thus outlier detection provides a shielding mechanism for WSNs against erroneous data which leads to fallacious operations. In this paper, we present a review on basic outlier detection techniques in WSNs .The survey can help to evaluate different techniques and can offer suggestions for future research.**

***Index Terms* — Wireless Sensor Networks, Outlier detection, Bayesian Networks Based Approaches, Support Vector Machines**

I. INTRODUCTION

The fundamental concept of WSN (Wireless Sensor Network) is that it is an integrated network of sensor nodes and is widespread over vast area, with one or more dominant sink nodes that gather data from other sensor nodes and then route the collected data to the client. The sensor nodes possess the ability to sense; manipulate raw data to extract meaningful information and transmit data through wireless communication. Each node has four parts,

- A variety of sensors (temperature, humidity and pressure etc.)
- A microcontroller to process and store the sensor output
- A wireless transceiver to transmit and receive the sensed data
- A power source

WSN can be advantageous in scenarios that combine

- Harsh environment (for e.g. battle field, nuclear plant)
- Health Monitoring
- Air pollution Monitoring
- Large area (for e.g. .agricultural field, forest)
- Event detection (intrusion)
- Short sensor range (temperature, smoke detection etc.)
- High temporal and spatial variability

WSNs have become an imperative basis for such critical applications and are major source of data gathering. In many of these applications real time data mining of sensor data is essential to make intelligent decisions with the elimination of outliers with the implication of outlier detection techniques. Anomaly detection and deviation detection are other names for outlier detection and it constitutes a fundamental part of operational data mining along with analysis of co-occurrence of outliers and predictive modelling [1].

Statistics, data mining, machine leaning, information theory, and spectral decomposition carries great potential for research on outliers [2].

The common characteristic of outlier detection is that it employs spatio-temporal correlations among sensor data of neighboring nodes to distinguish between events and errors.

The major problems accompanied with the deployment of WSNs can be improved with the implementation of outlier detection techniques. Some of the problems are as follows:

A. Unreliable Data

Data generated by WSNs are often erratic and inaccurate due to limited resources and capabilities (such as battery power, storage memory, computational capacity and communication bandwidth) [3] .In addition the environmental factors that may result in outliers (erroneous data) are also unavoidable. Furthermore anomalous data results due to some security breaches which are a threat to data security policies. Noise and errors may affect the quality of data set and it may contain missing values, duplicate data, inconsistent values etc. Therefore it becomes prerequisite to ensure the reliability of sensed data before proceeding to the decision making process and the need for detection of outliers becomes evident.

B. Energy Utilization

Energy consumption is an important issue in WSNs. There are two types of energy consumption in WSNs. Energy Consumption due to useful sources and due to wasteful sources [4].

Useful Energy consumption can be due to

- Transmitting and receiving data
- Receiving and processing request queries from other neighboring nodes
- Forwarding request queries to other neighboring nodes

Wasteful Energy Consumption can be due to

- Retransmitting data due to packet collisions
- Idle Listening

Mehwish Raza and Umm-e-Farwa, Department of Computer and Information Systems Engineering, NED University of Engineering and Technology, Karachi, 75270, Pakistan. Email: mehwishraza@ymail.com.
Manuscript received July 13, 2016; Revised on August 25, 2016; Accepted on September 19, 2016.

 ISSN – 1999-4974

To reduce energy consumption and to improve network stability, clustering based approach found to be an effective technique [5] in which it is supposed that the normal objects or regular data belong to dense and outsized clusters, while outliers form very small clusters.

What are Outliers?

The values that vary significantly from majority of data set and fall outside the overall trend of the data set are commonly known as outliers. One of the classical characterizations of outliers is that an outlier is an observation, which differs so much from other observations as to arouse suspicions that it was generated by a different mechanism [6].

In WSNs, outliers can be defined as measurements that significantly diverge from the normal pattern of sensed data [7].

The sensors in WSN monitor the physical world and the data generated exhibit a normal behavior which forms the basis of the above definition. The deviation from normal behavior indicates the presence of outliers in dataset that can dramatically affect the process of data analysis.

Therefore it is quite important to develop an appropriate outlier detection technique with less communication and storage overhead.

II. CLASSIFICATION OF OUTLIERS

There can be different sources of outliers in dataset. If the outlier is an erroneous or noisy data, then it must be eliminated in order to ensure accurate and highly reliable data and conserve energy by reducing communication overhead. However, if the outlier is due to some event (for e.g. fire detection, chemical spills etc.) then removing outliers will result in information loss that may cost undesirable penalty. Therefore we may say the classification of outliers as either erroneous data or due to some real world event. Hence, outlier detection is an important research area that needs to be investigated in depth.

Different approaches are used by researchers for distinguishing anomalies:

- Centralized Approach
- Distributed Approach

In Centralized Approach, both clustering and outlier detection algorithm is performed at sink node. The data from source node is transmitted to the sink node through intermediate nodes so that it can be processed and analyzed. This approach can incur a large communication overhead if the sink node is far away from the source node and it may also result in network congestion. Additionally, the sink node will take a long time to resolve the anomalies. In Network or Distributed Approach, each sensor node performs a clustering algorithm to produce clusters. The outlier detection is performed at the gateway node [8].

It is evident that network/distributed approach is better than centralized approach, as it reduces the communication overhead thus saving energy and increasing network lifetime.

III. DIFFERENT OUTLIER DETECTION TECHNIQUES FOR WSNS

There are different techniques employed for outlier detection worldwide. On the basis of methodology, these techniques are categorized as follows:

A. Nearest Neighbor Based Technique

Nearest Neighborhood is one of the most commonly used approaches in data mining and machine learning for detection of outliers. It uses distance as a similarity measure between two data instances. In case of univariate data Euclidean Distance is used whereas continuous multivariate data are handled by Mahalanobis Distance Metric.

Some of the nearest neighbor techniques considered in our survey are presented below:

Reference [9] proposes that global outliers in sensor networks can be identified on the criteria of distance similarity. Set of characteristic data is exchanged between the neighboring nodes and each node performs distance similarity algorithm to identify outliers in the vicinity and then broadcasts the outliers to neighboring nodes for verification. The procedure is repeated by the neighboring nodes until all the sensor nodes in the WSN eventually approve the existence of global outliers. Multiple existing distance-based outlier detection techniques facilitate the flexibility of the proposed technique. In this technique, no particular network configuration is supposed to be adopted so every node in the network uses broadcast mode to communicate with other nodes in the network, which causes communication overhead within the network. Thus, it does not scale well and does not provide flexibility to form large-scale networks.

Paper [10] presents a distinct case of k-NN graph that is Mutual k Nearest Neighbor (Mk-NN). If k-NN directed graph has bidirectional edge both from vertices v_i to v_j and from v_j to v_i as shown in Fig.1 then these vertices are mutually linked together and the connected vertices form clusters in the data set while the connected components with unidirectional edge from a vertex to the other is defined an outlier. Outliers that exist too close to inliers or regular readings can be misclassified which is the underlying possible problem with this approach.

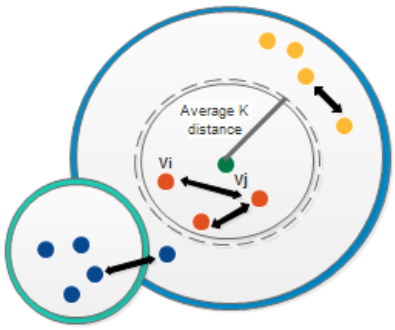

Fig. 1 Mk-NN directed graph

A. Statistical Based Technique

One of the earliest techniques that were used for outlier detection is Statistical Based Technique. This technique is

ISSN – 1999-4974

based on a model with respect to which data patterns are evaluated. If the data point is least likely to be produced by the reference model, based on distance measure, the data point is supposed to be an outlier. In short, an anomaly exists if any deviance from reference model is observed in the data set. In short, an anomaly exists if any deviance from reference model is observed in the data set. On the basis of probability distribution model built there are two types of Statistical Based Techniques which are as follows:

1) Parametric Based Approaches

In Parametric techniques, the reference model is built against the known data set distribution. The reference model then uses different parameters to evaluate the estimation of distribution parameters in the data set. Based on the type of distribution used, parametric approach is further classified into Gaussian Based models and Non Gaussian Based models.

2) Non Parametric Based Approaches

In Non Parametric Approach data distribution is not known. It measures the distance between the data instance (that is being tested) and reference model and uses some threshold value on the distance measured, to determine - whether the tested data instance is an outlier or a normal value. Non Parametric is further classified into Histogramming and Kernel Functions.

B. Classification Based Techniques

In this approach a training data set is used to learn a model (classifier), the model is then used to classify the unseen instances of data set under test. Fig.2 shows that classification based outlier detection technique usually operates in two phases as follows:

a) Training

This phase acquires a classifier model by making use of available training data.

b) Testing

This phase uses the learned model, classifies the unseen test data as either outlier or normal value.

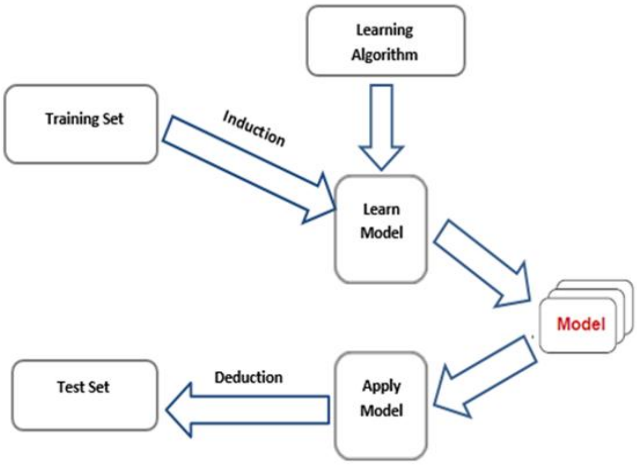

Fig. 2 Flow Diagram for Testing and Training phases

There are two main categories in classification based techniques are as follows:

1) One Class

This assumes that a single discriminative boundary lies between normal data values and outliers. The data value is declared as outlier when it does not fall within the boundary of normal instances of data. Fig.3 depicts One Class classification of outliers.

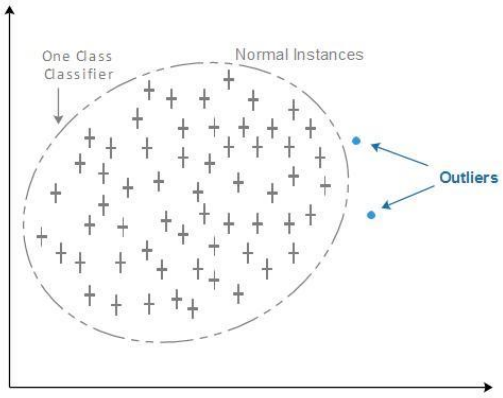

Fig. 3 One Class Representation of Outliers

2) Multi Class

This assumes that there are multiple classes to which the normal values may belong. The data value that does not belong to any of the normal class is declared as an outlier. Fig.4 depicts multi class classification of outliers. The existing classification based outlier detection techniques are further classified into Bayesian Network based approach and Support Vector Machines approach.

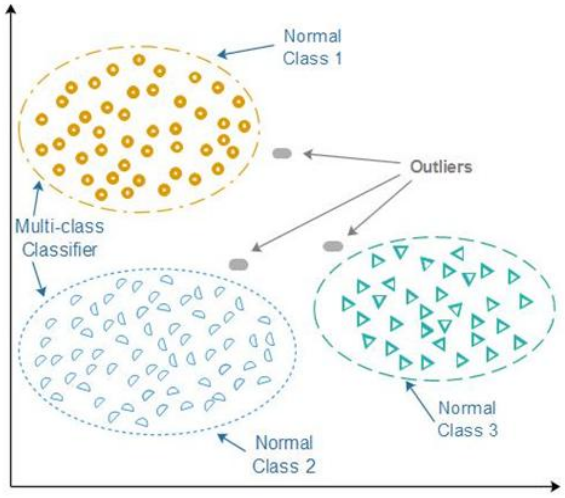

Fig. 4 Multi Class Representation of Outliers

a) Bayesian Network Based Approaches

This technique uses a Probabilistic graph model to represent a probabilistic relationship between random variables such as disease and symptoms. Using the symptoms, the technique can be used to evaluate the probabilities of different disease. The approach involves three main phases. In the first phase, learning of model is achieved using a training

ISSN – 1999-4974

data set. In the second phase testing is performed and in the third phase logical conclusions are derived and the results obtained are used for inference.

An idea of context aware sensors based on Bayesian model is proposed [11]. The technique is used for discovering outliers, detecting faulty sensor nodes, handling and approximating missing values and in network sampling. It exploits the spatiotemporal relation that exists among the sensor nodes in WSN by predicting the sensor's current reading on the basis of its own past readings and the readings of the sensors that are its neighbors. This spatial and temporal relation is termed as Contextual Information (CI). By using the Naïve Bayes method for classification, we can figure out whether the reading belongs to a class or not. If the reading falls outside the class it is acknowledged as an outlier .The technique does not require any specified threshold or reference model like in previous techniques. However it does not cover multi-dimensional data and is strictly restricted to one dimensional data. Furthermore, the conditional dependencies among the attributes are not considered by the Naïve Bayes networks.

In paper [12], Bayesian Belief Network (BBN) that covers the conditional dependencies among the observation of sensor attributes is proposed. The entire process is distributed in three phases constructing BBN, learning from BBN and inferring from BBN. The BBN is trained to capture the spatiotemporal relation among the sensor nodes and the conditional dependency that exist between the attributes of sensor nodes. The parameters that are to be learned by BBN are sensors' current reading, previous reading of the same sensor and neighbor readings. As compared to Naïve Bayesian Network, the Bayesian Belief Network provides more accuracy as it considers the conditional probability dependency as well.

Furthermore it is not restricted to one dimensional data but also covers multivariate data. However the accuracy of the technique may be affected if network topology changes over time.

b) Support Vector Machines

Support Vector Machines use supervised learning algorithms for data analysis and pattern recognition used in classification. The technique separates data points that belong to different classes by a hyper plane. Larger the distance of the hyperplane to the nearest data point, lesser will be the generalization error and more optimal will be the hyperplane. In paper [13], a SVM based distributed approach for detecting outliers in WSNs is proposed. This approach uses a one class quarter sphere SVM technique to detect anomalous data locally at each node. Outliers are considered to be the data values that lie outside the quarter sphere. Each node sends its radius information to its parent node; the parent node combines the children radii information with its own local radius information and compute global radius. The parent node then forwards the global radius to the children nodes which are instructed to relate their data values with the global radius and classify them as globally anomalous or normal.

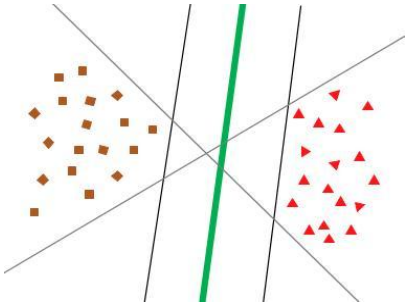

Fig. 5 Pictorial Representation of Support Vector Machine

IV. LIMITATIONS OF EXISTING OUTLIER DETECTION TECHNIQUES

- The techniques discussed above works only for static WSNs and not for dynamic WSNs.
- Most of the techniques do not take into account multivariate sensor data.
- Little work is done to handle online streaming datasets.
- There exists a need to design algorithms to distinguish between real-time events and outliers.
- Majority of the techniques do not contemplate the dependency that exists amongst the sensor attributes and neighboring sensor nodes.

V. LOCALIZATION AND OUTLIERS IN WSNS

Localization is one of the most critical research topics in wireless sensor networks (WSNs) nowadays. In many applications, it is necessary to record location information of nodes to perform cluster analysis. In localization process, the location information of all nodes is collected as primary data. However, these data entries may contain outliers that deviate from their true values of position. Thus, there rises a need to detect and handle outliers in order to achieve high localization accuracy. Network functionalities such as geographic routing and data centric storage can be analyzed by location estimation [14].

A direct solution for localization of sensor nodes in a WSN is to install global positioning systems (GPSs) in every sensor node [15]. However, installing GPS in every sensor node is impractical as it is not cost effective and the power requirement is also very high [16, 17, 18]. In recent years, a number of localization algorithms have been proposed to reduce or remove the dependence on GPS in WSNs [19]. The main idea in most localization algorithms is that a few nodes called anchors. These anchor nodes are aware of their locations (by GPS receivers or manual configuration) transmit beacons with their coordinates to help the rest nodes called unknown nodes discover their locations [16, 19]. As mentioned before, the primary data used by localization process is the position information and thus the distances between neighboring nodes is computed with this available data. However, these primary data may contain outliers that strongly deviate from their true values, which include both the outlier distances and the outlier anchors [20].

ISSN – 1999-4974

Table I Comparative Analysis between Outlier Detection Techniques

√True with subject to application - May or may not be applicable

Techniques	Sensor Data		Type of Outlier				
	Attribute		Global			Local	
	Univariate data	Multivariate data	Individual	Aggregate	Centralized	Individual	Collaborative
Nearest Neighbor Based Technique							
Branch et al. [10]	√				√		
Hautamaki[11]	-	-	-	-	-	-	-
Statistical Based Technique							
1) Bayesian Network Based Approaches							
Eiman Elnahrawy [12]	√						√
Janakiram[13]		√					√
2) Support Vector Machines							
Rajasegarar[14]		√	√			√	

A. Distance Outlier

In WSNs localization research field, the distances between neighboring nodes are accurately measured which are used to derive node locations accordingly. Typical distance-measuring techniques or ranging techniques include TOA, TDoA and RSS. However, among these distance measurements, there inevitably exist outlier distance whose distance measurement error (the difference between the true distance and the measured distance) is abnormally large [21]. Generally, the probable sources of outlier distances are as following:

1) Environmental factors: TOA may generate outlier distances with strongly enlarged estimates due to non-line-of-sight propagation. RSS is sensitive to channel noise, reflection, and interference, all of which have significant impacts on signal amplitude. The irregularity of signal attenuation remarkably increases, especially in complex indoor environments.
2) Hardware malfunction: When encountering ranging hardware malfunction distance measurements will be meaningless. In addition, incorrect hardware calibration and configuration also worsen ranging accuracy. For example, the inaccuracy of clock synchronization results in ranging errors for TDoA, and RSS suffers from transmitter, receiver, and antenna variability.
3) Malicious attacks: When a WSN is deployed in hostile environments; the localization process is becoming the target of adversary attacks. By reporting fake location or ranging results, an attacker, for example a compromised (malicious) node, can completely distort the coordinate system [21-22].

CONCLUSION

In this paper, a survey of existing outlier detection techniques in WSNs is presented. Furthermore shortcomings of the current techniques are discussed. We offer our survey to compare the techniques so as to assist researchers to make better choices. The future research directions of outlier detection techniques for wireless sensor networks localization possibly are as following:

1) It is necessary to design new outlier detection algorithms for localization in WSNs as general.
2) Detecting anchor outliers should be taken into account. More work must be done on distinguishing between anchor outliers and distance outliers. Also new techniques to detect anchor outliers must be developed.
3) Researches should propose new algorithms to detect outliers in localization algorithms for mobile sensor networks.
4) We can investigate the applicability of Artificial Intelligence (AI) techniques for outlier detection localization in WSNs.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, Anomaly Detection: A Survey,Technical Report, University of Minnesota, 2007.
[2] P.N. Tan, M. Steinback, and V. Kumar, Introduction to Data Mining, Addison Wesley, 2006
[3] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogerakiand, and D. Gunopulos, Online Outlier Detection in Sensor Data using NonparametricModels, J. Very Large Data Bases, VLDB 2006.
[4] Devasena,Dr B.Sowmya,A Study of Power & Energy Efficient Clustering Protocols in Wireless Sensor Networks, International Journal of Advance Research in Computer Science and Management Studies.
[5] Vivek Katiyar, Narottam Chand, Surender Soni, "A Survey on Clustering Algorithms for Heterogeneous Wireless Sensor Networks.
[6] D.M. Hawkins, Identification of Outliers, London: Chapman and Hall, 1980.
[7] Mohammad Abdur Razzaque and Simon Dobson,"Energy-Efficient Sensing in Wireless Sensor Networks Using Compressed Sensing",12 February 2014.
[8] V. Chandola, A. Banerjee, and V. Kumar, Anomaly Detection: A Survey, Technical Report, University of Minnesota.

 ISSN – 1999-4974

[9] J. Branch, B. Szymanski, C. Giannella, and R. Wolff, In-Network Outlier Detection in Wireless Sensor Networks, Proc. IEEE ICDCS, 2006.
[10] Ville Hautam¨aki, Ismo K¨arkk¨ainen and Pasi Fr¨anti," Outlier Detection Using k-Nearest Neighbour Graph"
[11] Eiman Elnahrawy and Badri Nath,"Context Aware Sensors".
[12] Janakiram, Adi Mallikarjuna Reddy V and A V U Phani Kumar, Outlier Detection in Wireless Sensor Networks using Bayesian Belief Networks.
[13] Sutharshan Rajasegarar, Christopher Leckie, Marimuthu Palaniswami, James C. Bezdek,"Distributed Anomaly Detection In Wireless Sensor Networks", ARC Special Research Center for Ultra-Broadband Information Networks, The University of Melbourne.
[14] X. Li, N. Mitton, I. Simplot-Ryl, and D. Simplot-Ryl, "Dynamic beacon mobility scheduling for sensor localization", IEEE Transactions on Parallel and Distributed Systems, vol. 23, no. 8, (2012), pp. 1439- 1452
[15] B. Hofmann-Wellenhof, H. Lichtenegger and J. Collins, "Global Positioning System Theory and Practice", 4th edn, Springer, New York, (1997).
[16] W. Du, L. Fang and P. Ning, "LAD: Localization anomaly detection for wireless sensor networks", Journal of Parallel and Distributed Computing, vol. 66, no.7, (2006), pp. 874-886.
[17] G. Mao, B. Fidan and B. D. O. Anderson, "Wireless sensor network localization techniques", Computer Networks, vol. 51, no. 10, (2007), pp. 2529-2553.
[18] X. Li, N. Mitton, I. Simplot-Ryl, and D. Simplot-Ryl, "Dynamic beacon mobility scheduling for sensor localization", IEEE Transactions on Parallel and Distributed Systems, vol. 23, no. 8, (2012), pp. 1439- 1452.
[19] J. Jiang, G. Han, C. Zhu, Y. Dong and N. Zhang, "Secure localization in wireless sensor networks: A survey", Journal of Communications, vol. 6, no. 6, (2011).
[20] R. Nagpal, H. Shrobe and J. Bachrach, "Organizing a global coordinate system from local information on an ad hoc sensor network", IPSN'03, (2003).
[21] Q. Xiao, K. Bu, Z. Wang and B. Xiao, "Robust localization against outliers in wireless sensor networks", ACM Transactions on Sensor Networks, vol. 9, no. 2, (2013).
[22] Z. Yang, L. Jian, C. Wu and Y. Liu, "Beyond triangle inequality: Sifting noisy and outlier distance measurements for localization", ACM Transactions on Sensor Networks (TOSN), vol. 9, no. 2, (2013).

 ISSN – 1999-4974