TECH-CHRONICLE
AN INTERNATIONAL E-JOURNAL ON
EMERGING TRENDS IN SCIENCE, TECHNOLOGY AND MANAGEMENT

Outliers in Wireless Sensor Network: A Survey

Tripti Sharma
IT Department
Maharaja Surajmal Institute of Technology
New Delhi-110058

Abstract - In today's world, a system is referred to be a smart system if it optimizes the use of energy and bandwidth to send the data. Defining outliers of a WSN is to detect the nodes, sending unusual data. An outlier is defined as the measurement that deviates from the normal behaviour of the sensed data. Therefore, defining and detecting the outliers can lead to differentiate between the useful information and extra information. The aim of the paper is to provide a structured and comprehensive overview of the various factors causing outliers. Also the various techniques for outlier detection in WSN are discussed in detail. In addition, this paper draws a fine line of comparison of the existing techniques.
.

Keyword- *Wireless Sensor Network, Causes,*

I. INTRODUCTION

Wireless Sensor Network is a system consisting of wide range of wireless sensors deployed in a region to sense various types of physical information from the environment. The data sensed by these sensors are sent to the base station (BS) for further analysis. WSNs are used for multiple purposes like habitat monitoring, military surveillances and natural calamities in the remote areas such as forests, deserts, etc. Due to the small size and wireless data transmission property of the nodes, they can be deployed easily, even in remote areas and hilly terrains. However, the number of such nodes is considerably high and monitoring these nodes is quite difficult,especially in cases when the nodes are distributed in the regions far away from the city or town. The network once established, keeps on sensing the data and the energy of the nodes keep on dissipating whenever, they receive some information and send it further to other nodes or BS.The term outlier, also known as anomaly, originally stems from the field of statistics. The two classical definitions of outliers are: Hawkins): “an outlier is an observation, which deviatesso much from other observations as to arouse suspicions that it was generated by a different mechanism”. Barnett and Lewis: “an outlier is an observation (orsubset of observations) which appears to be inconsistent with the remainder of that set of data”.

Now here comes the role of the outlier as the data collected by an outlier is not useful and not needed to be transmitted. As the transmission of data consumes the energy of the receiver to receive the data and process it. So it is better to stop the data collected by an outlier from being

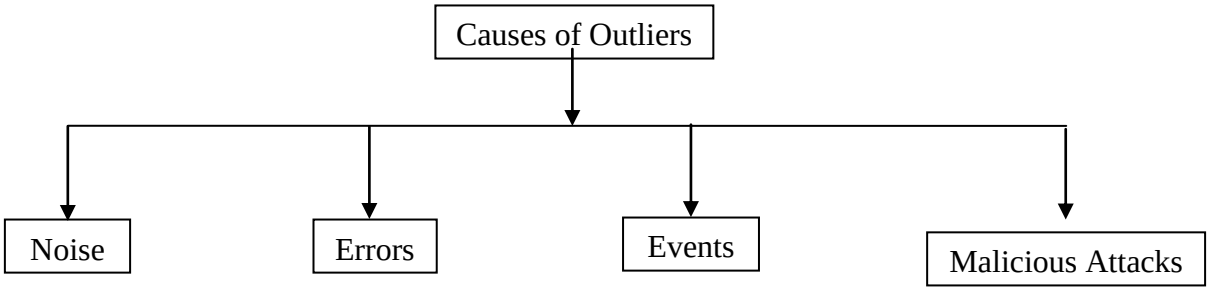
transmitted. A number of methods have been proposed to detect the outliers and to make the network more energy efficient.

The rest of the paper is organized as follows: Section 2 will describe the causes of outliers. Section 3 depicts the review of existing methods to detect an outlier. Finally, section 4 illustrates the conclusion and the future works for further improvement.

II. Causes of outliers

A lot of study has been done to find the causes of outliers in WSN because by detecting them we can make a system more efficient in terms of energy saving, bandwidth and memory consumption.

Causes of outliers are:



A. NOISE

Noise is random unwanted signals that enters the communication system via communication medium and interferes with the transmitted message. It leads to error in message signals. It is one of the basic factors which can occur during any transmission. Noise is of different types, such as:

1. Thermal noise - Thermal noise is approximately white, meaning that its power spectral density is nearly equal throughout the frequency spectrum. The amplitude of the signal has very nearly a Gaussian probability density function. A communication system affected by thermal noise is often modelled as an additive white Gaussian noise (AWGN) channel.

2. Shot Noise - If electrons flow across a barrier, then they have discrete arrival times. Those discrete arrivals exhibit shot noise. The output of a shot noise generator is easily set by the current. Typically, the barrier in a diode is used. Shot noise in electronic devices results from unavoidable random statistical fluctuations of the electric current when the charge carriers (such as electrons) traverse a gap. The current is a flow of discrete charges, and the fluctuation in the arrivals of those charges creates shot noise.

3. Flicker Noise - Flicker noise, also known as 1/f noise, is a signal or process with a frequency spectrum that falls off steadily into the higher frequencies, with a pink spectrum. It occurs in almost all electronic devices, and results from a variety of effects, though always related to a direct current.

4. Burst noise - It consists of sudden step-like transitions between two or more levels (non-Gaussian), as high as several hundred microvolts, at random and

unpredictable times. Each shift in offset voltage or current lasts for several milliseconds, and the intervals between pulses tend to be in the audio range (less than 100 Hz), leading to the term popcorn noise for the popping or crackling sounds it produces in audio circuits.

5.Transit-time noise - If the time taken by the electrons from travelling from emitter to collector becomes comparable to the period of the signal being amplified, that is, at frequencies above VHF and beyond, so-called transit-time effect takes place and noise input admittance of the transistor increases. From the frequency at which this effect becomes significant it goes on increasing with frequency and quickly dominates over other terms.

6.Atmospheric noise (static noise**)** - This noise is also called static noise and it is the natural source of disturbance caused by lightning discharge in thunderstorm and the natural (electrical) disturbances occurring in nature.

7. Industrial noise - Sources such as automobiles, aircraft, ignition electric motors and switching gear, High voltage wires and fluorescent lamps cause industrial noise. These noises are produced by the discharge present in all these operations.

8.Extra-terrestrial noise - Noise from outside the Earth includes:

a) Solar noise - Noise that originates from the Sun is called solar noise. Under normal conditions there is constant radiation from the Sun due to its high temperature. Electrical disturbances such as corona discharges, as well as sunspots can produce additional noise.

b) Cosmic noise - Distant stars generate noise called cosmic noise. While these stars are too far away to individually affect terrestrial communications systems, their large number leads to appreciable collective effects. Cosmic noise has been observed in a range from 8 MHz to 1.43 GHz.

B. ERRORS

Errors in the information taken by the node can cause the node to be the outlier. This means the information sensed by the node itself is not reliable and is of no use. So in this case the information sensed by the node can be discarded at the end of the node itself so as to reduce the power consumption of the receiver, also the requirement of the receiver is also reduced. Errors are due to many reasons; some of them are as follows:

1. Imperfect sensor - This situation occurs when the sensing element in the node is damaged due to some circumstances and is not able to sense accurate data from the surroundings. This can be due to the extreme environmental conditions or some natural calamity.

2. Low Battery - Battery is necessary for the sensors to sense the accurate data. Once the battery becomes low from the minimum requirement of the sensors, the data sensed by the sensor does not remain reliable.

C. Events

Some events may occur in the surrounding of the node due to which it becomes an

outlier. As mostly the WSN is employed inthe areas like forests, deserts, mountains and other area in which normal access and surveillance are a bit difficult, so to overcome the need of the presence of humans in such area Events like fire, floods, etc. leads to the sensing of unusual data and due to which the data deviates from its normal pattern. This occurs due to damaging of the sensing material.

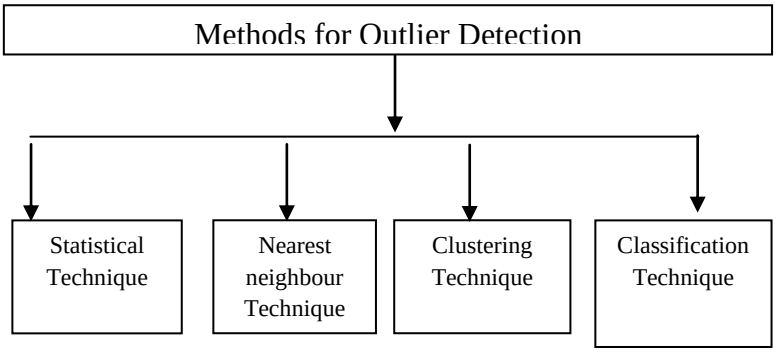
D.MALACIOUS ATTACKS

WSNs have a wide variety of applications especially in the areas of wars and borders which include forests and marine borders so there WSN plays an important role in detecting anything unusual in the surroundings. Nodes of WSN can be equipped with different types of sensors to take different types of data as per our requirements. As WSN are to be used for surveillance purposes by the security forces they are prone to be attacked by the hackers to transmit the data they wanted to reach to the army base.

These malicious attacks and manipulation done by the people who are not linked with the organization which is responsible to take care of the WSN leads to the outlier. To survive through these manipulations grouping method is applied to detect the outliers. In which a group of nodes in the same range of area are selected and the data through all of them is processed, before considering any node as an outlier.

III. Review of Existing Method of Outlier Detection

A lot of study has been done to detect the outliers in a WSN and many techniques have been developed. These techniques can be classified as - Statistical technique, Nearest Neighbour technique, clustering method and classification based techniques.Statisticaltechnique are further categorized into parametric and non-parametric approaches. Gaussian and non-Gaussian techniques further belong to parametric approach, whereas kerneland histogram techniques belong to non-parametric approach. Classification based techniques are categorized as Bayesian network based and support vector machine based techniques.



A. Statistical Techniques

Statistical Techniques are the most ancient techniques to overcome the problems related to outliers. This technique is model based technique. They assume the probability distribution of the data and then check the new information, that whether it fits in the data model formed or not. If the data does not match significantly with the existing model then that particular node is treated as an outlier. Statistical Techniques can further be classifies as - 1.Parametric Approach 2. Non-Parametric Approach.[1]

1. Parametric Approach - This approach assumes the presence of the data in the derived model, this implies that the new data sensed is a part of the known distribution. It then finally examines all the

parameters to detect whether the assumption stands true, this verification is based on the two methods, Gaussian Model and Non - Gaussian Model.

a) Gaussian Model - According to Wu Et Al.[2]. There are two local techniques for defining outlier in a WSN. These techniques are based on the geographical correlation of the readings with the neighbour nodes and the event boundary in sensor networks. In this type of model, reading of each node is processed and difference between its past reading and between the reading of the neighbour nodes. Then an existing node is declared as an outlier if its value deviation is greater than the predefined threshold value. The technique of the event boundary detection considers the previous results of outlying sensor identification and determines a node as an event node if the absolute value of the node's deviation degree in one geographical region is much larger thanthat in another region. Certainty of these outlier detection techniques is not relatively high due to the fact that these techniques ignore the physical correlation of the node readings.Certainty of these outlier detection techniques is not relatively high due to the fact that these techniques ignore the physical correlation of the node readings.

According to Bettencourt Et Al. [3], a local outlier detection technique is to identify the errors and reveal the events in ecological application of WSNs. This technique can differentiate between normal measurements and events by using the geographical-temporal correlations of sensor data. Each node is indicated by the statistical distribution of the difference between its own readings and each of its neighbournodes, as well as between its present and previous measurements.A measurement is identified as malicious if its reading varies significantly when compared to the threshold value. The detected deviated reading is considered as an event if it deviates from its normal pattern, temporarily but should also remain geographically correlated.The limitation of this technique is that it relies on the choice of the appropriate values of the threshold.

Hida et al. [4]evolved a technique to make simpleaggregation operations, such as MAXIMUM or AVERAGE, nodes with better reliability under the presence of faulty sensor readings and failed nodes. This technique is based on the geographic-temporal correlations of sensor data and uses two statistical examinations to detect outliers. All the new values coming to the station are compared with the present and past value of all the sensors in the near neighbourhood. If the new value sets the two examinations true, then it is considered to be eligible to be processed further; otherwise if the new value deviated from the existing limits through a certain threshold then it is declared as an outlier and is discarded from the analysis. The limitation of this technique is that it only deals with one-dimensional outlier data and much more memory is required for a node to store historical values of all its neighbouring nodes.

b) Non- Gaussian Model - Jun et al.[5]evolved a statistical technique which is based on the symmetric α- stable distributionto detect the outliers. The technique utilizes the geographical-temporal correlations of sensor data to detect outliers. Each node in the cluster first sensed the data and updates the

temporal by comparing the assumed data and the sensed data. Then the cluster-head gathers the corrected data and further examine it for outliers by comparing the data to find the deviated observations. This curbs the communication cost as the cluster head does the manipulation and sends only the limited data. The SαS distribution may not be suitable for real sensor data and the cluster-based structure may be susceptible to dynamic changes of network topology.

2) Non-Parametric Approaches -This approach does not assume the availability of data distribution. It defines the distance between a new sensed value and the statistical model, and then compares it with a threshold distance. This verification is based on the two methods, Histograms and Kernel Density Estimator.

Histogram method is the technique based on counting the frequency of occurrence of various information sensed and examining it on the basis of the difference between the newly sensed information and the values of the histogram, to check whether it belongs to one of them. A kernel density estimator is the method based on the kernel functions, to estimate the probability distribution function (pdf) for the normal instances. When a new piece of information is sensed, it is compared from the pdf and if it lies in the probability area it is considered to be an outlier.

• Histogram method- Sheng et al.[6]evolved a histogram method to identify global outliers in data collection applications of sensor networks. This methodtries to curb the communication expenditure by seeking histogram information rather than collecting raw data for centralized manipulation. The central node usesthe histogram information to extract the data distribution from the network and differentiates between outliers and non-outliers. The outliers can be determined by comparing it with a pre-defined threshold distance. The limitation of this technique include the fact that re-collecting more histogram information from the whole network will cause too much communication overhead and the technique only considers one-dimensional data.

• Kernel density estimators- Palpanas et al.[7] evolved a kernel method for online identification of outliers instreaming sensor data. This method does not require any kind of pre-defined sets of data to compare the present value of the node. It uses kernel density estimator to estimate the underlying distribution of sensor data. Therefore, all nodes are fully equipped to identify locally, the outliers in WSN, if the newly sensed information varies significantly from the estimated value. A node is considered to be an outlier when the difference between the value is greater than the pre-defined threshold.The main problem of this technique is its high dependency on the defined threshold, while choice of an appropriate threshold is quite difficult and a single threshold may also not be suitable for outlier detection in multi-dimensional data. Furthermore, the technique does not consider maintaining the model while sensor data is frequently updated

B. Nearest Neighbour Technique

Nearest neighbour techniques are the most commonlyused techniquesto examine thenewly sensed data with respect to its nearest neighbours. It uses various well-structured process to calculate and

determine the distance between the two instances[8], [9]. A node is declared to be outlier if its data sample is away from its neighbours. Euclidean distance is the most common choice in practice for univariate and multivariate continuous attributes.

Branch et al. [10]evolved a technique which used distance as its basis, similarto identify global outliers in sensor networks. This technique tries to curb the communication overhead by a combination of representative data exchanges among neighbouring nodes. Each node considers the distance of the sensed information from the expected value to detect the outlier and then transmits it to all the neighbouring nodes in the WSN to confirm the decision. The neighbouring nodes go through the same process till all the sensor are covered and agree with the decision. This method is flexible in reference to thevarious existing distance-based outlier detection techniques.However, the technique does not adopt any network structure so that every node uses broadcast to communicate with other nodes in the network, which will cause too much communication overhead. Consequently, it does not scale well to the large-scale networks

Zhuang et al.[11]evolved two in-network outlier cleaning methods for data gathering applications of sensor networks. One method is based on wavelet analysis usually used for noises or occasionally,errors which leads to outlier. The other method is based on dynamic time warping (DTW) which finds the distance similarity and compares the values, specifically for outliers that contain errors for more than a particular period of time. In this method, each node changes sensed information to the wavelet time-frequency domain and detectsthe high-frequency data measurements and consider them as outliers and manipulates it to correct it using proper wavelet coefficients.A limitation of this method, however, is its dependency of a suitable pre-defined threshold that is not obvious to define

C. Clustering Technique

Clustering techniques are generally used in the datamining community to group similar data into clusters with similar behaviour. Data isconsidered as outliers if they do not fall in the range of clusters or if the size of the cluster is much smaller than other clusters. Euclidean distance is often used as the dissimilarity measure between two data instances.

Rajasegarar et al.[12] propose a global outlier detection technique based on clustering technique to identify anomalous measurements. This technique minimizes the communication overhead by clustering the sensor measurements and merging clusters before communicating with other nodes.

D. Classification Technique

In existing outlier detection techniques for WSNs, classification-based approaches are categorized into support vector machines (SVM)-based and Bayesian network-based approaches based on type of classification model they use.

1)Support Vector Machine-Based Approaches: In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis.

Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier

2) Bayesian Network-Based Approaches: A Bayesian network, Bayes network, belief network, Bayesian model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).

In[13], the author proposes a Bayesian model-based technique todiscover local outliers and detect faulty sensors. In [14], Janakiram proposes a technique which uses BBN to capture not only the spatio-temporal correlations that exist among the observations of sensor nodes but also conditional dependence among the observations of sensor attributes.

IV.Conclusion

In this paper, we have surveyed various techniques used for the detection of Outliers in a WSN. We have also provided a structured and comprehensive overview of various factors causing outliers. We have compiled the latest research, which will help in determining the trade-off between various outlier detection techniques in a WSN.

V.REFERENCES

[1] N. Zhang, Meratnia and P. Havinga, "Outlier Detection Techniques for Wireless Sensor Networks: A Survey," *IEEE Commun. Surv. Tutorials*, vol. 12, no. 2, pp. 159–170, 2010.

[2] W. Wu, X. Cheng, M. Ding, K. Xing, F. Liu, and P. Deng, "Localized Outlying and Boundary Data Detection in Sensor Networks," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 8, pp. 1145–1157, 2007.

[3] S. M. A. Bettencourt, "Separating the Wheat from the Chaff: Practical Anomaly Detection Schemes in Ecological Applications of Distributed Sensor Networks," *Distrib. Comput. Sens. Syst.*, vol. 4549, pp. 223–239, 2007.

[4] R. Of, "H l u a," *ASA Refresher Courses in Anesthesiology*. .

[5] M. C. Jun, H. Jeong, and C.-C. J. Kuo, "<title>Distributed spatio-temporal outlier detection in sensor networks</title>," vol. 5819, pp. 273–284, 2005.

[6] B. Sheng, Q. Li, W. Mao, and W. Jin, "Outlier detection in sensor networks," *Proc. 8th ACM Int. Symp. Mob. ad hoc Netw. Comput. - MobiHoc '07*, p. 219, 2007.

[7] T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Distributed deviation detection in sensor networks," *ACM SIGMOD Rec.*, vol. 32, no. 4, pp. 77–82, 2003.

[8] E. M. Knorr and R. T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Datasets," *24th Int. Conf. Very Large Data Bases*, pp. 392–403, 1998.

[9] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *ACM SIGMOD Rec.*, pp. 427–438, 2000.

[10] J. Branch, B. Szymanski, C. Giannella, R. W. R. Wolff, and H. Kargupta, "In-Network Outlier Detection in Wireless Sensor Networks," *26th IEEE Int. Conf. Distrib. Comput. Syst.*, pp. 0–7, 2006.

[11] Y. Zhuang and L. Chen, "In-network Outlier Cleaning for Data Collection in Sensor Networks," *Work. VLDB*, 2006.

[12] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. Bezdek, "Distributed Anomaly Detection in Wireless Sensor Networks," *2006 10th IEEE Singapore Int. Conf. Commun. Syst.*, pp. 1–5, 2006.

[13] E. Elnahrawy and B. Nath, "Context-Aware Sensors," Springer, Berlin, Heidelberg, 2004, pp. 77–93.

[14] D. Janakiram, V. A. Reddy, and A. V. U. P. Kumar, "Outlier Detection in Wireless Sensor Networks using Bayesian Belief Networks," in *2006 1st International Conference on Communication Systems Software & Middleware*, pp. 1–6.