Outlier Detection: A Survey on Techniques of WSNs Involving Event and Error Based Outliers

Deep Shikha Shukla
Dept. of ECE
SRMU, Lucknow, India
shukla_deepshikha@yahoo.com

Avinash Chandra Pandey
Dept. of CSE/IT
JIIT, Noida
avinash.pandey@jiit.ac.in

Ankur Kulhari
Dept. of CSE/IT
JIIT, Noida
ankur.kulhari@jiit.ac.in

***Abstract*— In the recent few years, many wireless sensor networks have been distributed systematically in the real world to collect valuable raw sensed data. However, the crucial point of challenge is to extract high level knowledge from this raw sensed data. In the application of data analysis, a necessary pre-processing step is anomaly detection, also known as deviation detection or data cleansing. Outliers in wireless sensor networks (WSNs) are those measures that deviate from a defined pattern. Outlier detection can be used to remove noisy data, detect faulty nodes and discover interesting events. Numerous small and low cost nodes loaded with capabilities of integrated sensing and computation are involved in a WSN structure.Due to high density WSNs are exposed to faults and nasty attacks causing inaccurate and unreliable sensors reading, making Wireless sensor networks prone to outliers. This survey provides an outline of outlier detection techniques and approaches focusing on event and error based outliers.**

Keywords- outlier; outlier detection; anomaly; wireless sensor networks(WSNs); clustering;

I. INTRODUCTION

Wireless sensor networks (WSNs), currently with its built out in MEMS technology have gained world-wide attention in recent years, which has facilitated the development of smart sensors. In comparison to traditional sensors these are cheap and small in size with limited resource of processing and computing.These sensor nodes are enabled with the capability of sensing data and collecting information from the environment, based on some local decision process and transmit the sensed data to the user. In a WSN number of sensor nodes (few tens to thousands) are incorporated to work together, so that, a particular region can be monitored to collect the environment related data. [1]

The WSNs is not only used to detect time-critical events but also to provide fine-grained real time data about the physical world. WSNs has a vast range of applications in persona field, industrial, business, and military domains, such as environmental monitoring, object and inventory tracking, monitoring health and medical issues, battlefield observation, industrial safety and control. WSNs work on the principle of sensing environment to collect the data from it. Maximum time data collected by WSNs is corrupted by noise or have some missing values, error, inconsistent and duplicated data.

Anomalies identification is a must required step to make sure that data collected by WSNs must provide good data quality, secure monitoring and reliable detection of interesting and critical events. In WSNs, outliers are set of data, diverged from the standard pattern of data instances [2]. Distinguishing between causes of outliers is important as it gives an insight on how to handle the detected outlier [3].The potential source of outlier in WSNs can be because of error or event data. This survey paper focuses on various techniques to detect outliers in WSNs based on error and event data.

The main contribution of this paper is as follows:

- Briefly describes the fundamentals of outlier and outlier detection in WSNs.
- Discuss the mechanism involved in outlier detection.
- Describes various techniques of outlier detection based on both error and event data.

II. LITERATURE SURVEY

Outlier detection is a well known problem in wireless sensor network. Due to their special characteristic of constrained available resources, frequent physical failure, WSNs are likely to produce outliers more in comparison to their other wireless counterparts. In 2006 Subramaniam et al. [4] proposed an outlier detection technique for wireless sensor networks. In 2009 J.Austin et al. [5] defined various techniques for outlier detection based on statistical, nearest-neighbor and clustering based approaches. Rajasegarar et al. [6] proposed a global outlier detection techniques based on clustering-based technique. Branch et al. [7] also proposed a technique based on distance similarity to segregate global outliers from set of sensed data. Jun et al. [8] presented a statistical-based technique, which uses a symmetric alpha stable distribution to model outliers. Sheng et al. [9] presented a histogram based technique to identify global outliers in data collection applications of sensor networks. Wu et al. [10] present two local techniques for identification of outlying sensors as well as identification of event boundary in sensor networks. Bettencourt et al.[11] present a local outlier detection technique to identify errors and detect events in ecological applications of WSNs. Palpanas et al.[12] propose a kernel based technique for online identification of outliers in streaming sensor data. Zhuang et al. [13] present two in- network outlier techniques for data collection applications of sensor networks. Elnahrawy and Nath [14] present a Bayesian model-based technique to discover local outliers and detect faulty sensors. Jankiram et al. [15] given a technique based on Bayesian belief network to

discover local outliers in streaming sensor data. A Fawzy et al. [16] presented a combined cluster based and nearest neighbor based approach to account interesting events for outlier detection.

III. FUNDAMENTALS OF OUTLIER

This section includes definition of outliers, sources of outliers and information needed for outlier detection in WSNs.

A. How to define Outlier

The origins of the term anomaly or outlier lie in the field of statistics [5]. According to Hawkins [17]*"Outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism"*. However Barnett and Lewis [18] defined it as "*Outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data"*.

Except the above defined classical definition of outliers, there is existence of definition that defines outlier with respect to specific type of data set. In WSNs, measurements which diverge from the regular pattern are known as outliers.

B. Sources and Types of Outliers

The major sources of occurrence of outliers are noise and errors, event, and malicious attacks. Since malicious attack resulted outliers is not concerned with the data. Our main focus is on outliers caused by errors and events.

a) *Errors:* Errors are basically those observations which differ from proper measured value. Faulty sensor is accountable for errors and noisy data. Outliers caused by errors may occur repeatedly, while outliers caused by events tend to have extremely smaller probability of occurrence [19].

b) *Events:* Events refer to observations that indicate a change in the state of the environment, compared to the predefined 'usual behavior. An event is defined as a particular phenomenon that changes the real-world state, For example Forest fire, air pollution, etc. this kind of outliers generally lasts for a reasonably long period of time and changes historical pattern of sensor data [20].

Outliers can be classified into three categories. These are point, contextual, and collective outliers and can be defined as follows:

a) *Point outliers:* A point outlier can be defined as an instance of an individual data that is diverged from normal pattern of the data. These types of outliers are simplest in nature and can be recognized easily.

b) *Contextual outliers:* The abnormalities of data instance in a specific context are referred as contextual or conditional outlier.

c) *Collective outliers:* Collective outlier is a group of those linked data instances that differ in their pattern with reference to whole dataset.

Aim of outlier detection for WSNs is to discover outliers and to differentiate between errors and events with a high precision and with low false positive rates (FPRs). Outlier detection also considers the resource constraints of WSNs such that memory capacity, battery power etc because sometimes a highly accurate technique that does not meet the resource constraints would simply be unusable.

IV. OUTLIER DETECTION TECHNIQUES

In this section, the centre of interest is the various techniques that can be used in a WSN to detect outliers. We will analyze these techniques briefly.

A. Statistical-Based Approaches

Statistical based approach is the earliest defined approaches to find a way of the problem of outlier detection. The statistical outlier detection techniques [20] are basically model-based techniques [2]. In this technique a statistical (probability distribution) model capable of capturing the distribution of the data is assumed and then it evaluates the data instances to get how well they fit the model. Any data instance will be declared as an outlier if the probability of the data instance to be generated by this model is very low [20].The statistical based approach can further be defined as parametric and non-parametric based approaches.

a) *Parametric-Based Approaches:* In this technique data is abstracted from a known distribution and this data abstraction is based on the assumption regarding the availability of information about data distribution. On basis of assumed distribution, these techniques are further categorized into *Gaussian-based model* [20] and *non-Gaussian-based model* [20].

b) *Non-Parametric Based Approaches:* Unlike parametric approaches [16], non-parametric techniques [16] do not rely on the assumptions made regarding the availability of data distribution. In this technique, to determine whether the observation is an outlier or not, firstly, a distance measure is defined between a new instance and the statistical model and then some kind of thresholds are applied [20]. Non-parametric techniques to detect outliers make an approach with respect to the probability of occurrence of a data instance and for this purpose they use *histogram models* and *kernel functions* to estimate the probability of occurrence of data instance.

B. Nearest Neighbor-Based Approaches

Nearest neighbor-based approach [16] is widely used in the data mining and machine learning area to analyze any data instance with respect to its nearest neighbors. This technique mainly works on the computation of the distance between two data instances and on basis of this approach, if data instance falls away from its neighbor then it is called outlier. For this purpose of computing distance, nearest neighbor based approaches uses several well defined distance notions. Advantages of nearest neighbor based approach are [16]:

- It is unattended in nature and does not make any presumptions in connection with the underlying distribution of the data.

- Applying nearest neighbor-based techniques to different data type is simple, and primarily requires defining an appropriate distance measure for the given data.

C. Clustering-Based Approaches

Cluster-based techniques, a popular approach in data mining are used to group similar data instances into clusters with similar behavior. Data instances can be called as outlier if they are small in size than other clusters or if they do not belongs to clusters are identified as clusters [20]. Prons of cluster based approach are as follows [16]:

- It is easily adaptable to incremental mode (i.e. after learning the clusters, new points can be inserted into the system and tested for outliers).
- It does not have to be supervised.
- It is suitable for anomaly detection from temporal data [21].
- The testing phase for clustering based techniques is fast since the number of clusters against which every test instance needs to be compared is a small constant [21].

D. Classification-Based Approaches

Classification approaches are based on important logical approaches and is used in the data mining and machine learning area. In classification approaches to predict the outcome, algorithm is trained using training set containing a set of attributes and the respective outcomes, usually called predication attributes. Classification-based approaches provide an exact set of outliers by building a classification model to classify [20].Classification based approaches can further be classified into support vector machines (SVM)-based and Bayesian network-based approaches.

E. Spectral Decomposition-Based Approaches

Spectral decomposition-based approaches make use of principal component analysis (PCA) technique for outlier detection. Where PCA, before outlier detection, reduces the dimensionality and finds a new subset of dimension having the behavior of the data. Specifically, the top few principal components capture the build of variability and any data instance that violates this structure for the smallest components is considered as outlier.

V. Outlier Detection

A. Mechanism Involved

For the detection of outlier in wireless sensor networks, there are two proposed mechanism, centralized mechanism and in-network/distributed mechanism. In the case of centralized mechanism, both the clustering and outlier detection algorithm are performed when all the data from each sensor node is transmitted to the sink node.

While, in case of in-network/distributed mechanism, first the cluster algorithm is performed on the data at the sensor node and later at the gateway node the outlier detection algorithm is performed. Clustering algorithm is moved down to the network level to perform clustering where each sensor node performs the clustering algorithm on its own data to produce the clusters and the parent nodes combine its own clusters with the clusters from its intermediate children. Finally outlier detection algorithm is performed to detect the outlier at the gateway node.

B. Use of Event and Error techniques for Outlier Detection

It has been seen that related work in outlier detection is done under the circle of event detection domain. Although errors and events are semantically dissimilar and different mechanisms cause them, conceptually they both are same i.e. they both follow the basic idea of outlier behavior deviation from normal pattern of rest of the data. In-spite this, the main distinguish fact between them is that errors are local in nature and are node dependent while events are global in nature, so their neighbor nodes will also be affected.

Till now the researches on outlier detection were focused on event detection. Martinic and Schwiebert [19] employ a cell based network architecture to locally detect events based on collaboration among neighboring nodes. Krishnamachari and Iyengar [22] propose a distributed Bayesian protocol to detect event regions in presence of faulty sensor. Ding et al. [23] attempt to identify event boundaries since detection of event boundary may become more important than detection of event region.

In its research A. Fawzy et.al [16] proposed a new clustering based approach combined with nearest neighbor-based approach is proposed to classify outliers, i.e. noisy data or interesting events or erroneous data. The proposed methodology consists of following four steps:

a) *Pre-processing (clustering):* First, the clustering algorithm is applied on all the sensory data to group data into clusters.

b) *Outlier Detection:* Second step involves the process of applying outlier detection algorithm for each produced cluster, in order to label out each cluster as normal or outlier cluster. [21].

c) *Outlier Classification:* Third step involves the classification of the degree of outlier value (error or event) [21].

d) *Measuring Sensor Truthfulness:* The last step is to compute the truthfulness of each sensor node to increase our certainty in trusting a specific node.

VI. Conclusion

In this paper, the main attention is given to the outlier detection problem in wireless sensor networks. Here in this paper various outlier detection techniques suitable for WSNs have been discussed. A new methodology taking into account the error data for outlier detection with all the event based technique is focused in this survey. Our approach is based on both event and error based techniques for outlier detection.

References

[1] J.Yick, B.Mukherjee, D.Ghosal, Wireless sensor network survey, Elsevier, Computer Networks 52(2008) 2292-2330.

[2] Y. Zhang, N. Meratnia and P. Havinga, Ensuring high sensor data quality through use of online outlier detection techniques in International Journal of Sensor Networks, 7 (3). pp. 141-151. ISSN 1748-1279.

[3] M.Bahrepour, Y.Zhang, N.Meratnia, and P.J.M.Havinga, Use of Event Detection Approaches for Outlier Detection in Wireless Sensor Networks, IEEE 2009.

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

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

[6] S.Rajasegarar, C.Leckie, M.Palaniswami, JC.Bezdek, Distributed anomaly detection in wireless sensor networks.UK: IEEE, ICCS; 2006, pp.12-16.

[7] J.Branch, B.Szymanski, C. Giannella, R.Wolff, In network outlier detection in wireless sensor networks. In: Proceedings of IEEE ICDCS; 2006.

[8] M.C. Jun, H. Jeong, and C.C.J. Kuo, Distributed Spatio-Temporal Outlier Detection in Sensor Networks, Proc. SPIE, 2006.

[9] B. Sheng, Q. Li, W. Mao, and W. Jin, Outlier Detection in Sensor Networks, Proc. MobiHoc, 2007.

[10] 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.

[11] L.A. Bettencourt, A. Hagberg, and L. Larkey, Separating the Wheat from the Chaff: Practical Anomaly Detection Schemes in Ecological Applications of Distributed Sensor Networks, Proc. IEEE International Conference on Distributed Computing in Sensor Systems, 2007.

[12] T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, Distributed Deviation Detection in Sensor Networks, ACM Special Interest Group on Management of Data, pp. 77-82, 2003.

[13] Y. Zhuang and L. Chen, In-Network Outlier Cleaning for Data Collection in Sensor Networks, Proc. VLDB, 2006.

[14] E. Elnahrawy and B. Nath, Context-Aware Sensors, Proc. EWSN, 2004.

[15] D. Janakiram, A. Mallikarjuna, V. Reddy, and P. Kumar, Outlier Detection in Wireless Sensor Networks using Bayesian Belief Networks, Proc.IEEE Comsware, 2006.

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

[17] D.M. Hawkins, Identification of Outliers, London: Chapman and Hall, 1980V. Barnett and T. Lewis, Outliers in Statistical Data, New York: John Wiley Sons, 1994.

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

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

[20] Y. Zhang, N. Meratnia, P. Havinga, Outlier Detection Techniques for Wireless Sensor Networks: A Survey, IEEE Communications Survey and Tutorials, Vol. 12, No.2, Second Quarter 2010.

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

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

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

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

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

[26] N. Upasani, H.Om, Outlier Detection: A Survey on Techniques Involving Fuzzy and/or Neural Approaches, IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions, July 2013, IIT Kanpur, India.

[27] N. Shahid, I. H. Naqvi, S. B . Qaisar, One-class support vector machines: analysis of outlier detection for wireless sensor networks in harsh Environments in Springer Science+Business Media Dordrecht 2013.