International Journal of Advanced Research in IT and Engineering

ISSN: 2278-6244

EVALUATING THE EXISTING SOLUTION OF OUTLIER DETECTION IN WSN SYSTEM

Manoj Kumar*

Abstract: *In wireless networks, the dimensions that deviate from the normal behavior of sensed data are taken to be as outliers. The potential sources of outliers can be noise and errors, events, and malicious attacks on the network. This paper gives an overview of existing outlier detection techniques specifically developed for the wireless sensor networks.*

*Associate Professor, Department of Computer Sc. & Engg Green Hills Engineering College, Kumarhatti, Solan

1 WIRELESS SENSOR NETWORKS

A Wireless sensor network is composed of tens to thousands of sensor nodes which are densely deployed in a sensor field and have the capability to collect data and route data back to base station [1]. Wireless sensor network WSN (Wireless Sensor Network) is deployed in the region to monitor a large number of tiny sensor nodes, wireless communication means and form a multi-hop network of self-organizing system [2]. Wireless sensor networks (WSN) are wireless network composed of spatially distributed autonomous devices using sensors to cooperatively monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants, at different locations. Due to the deployment flexibility and maintenance simplicity, WSN applications have been seen in many areas [3] like telecommunication, medical, defence etc. The use of wireless sensor networks (WSNs) has grown enormously in the last decade, pointing out the crucial need for scalable and energy-efficient routing and datan gathering and aggregation protocols in corresponding large-scale environments [4].

2 SENSOR NETWORK ARCHITECTURE

We consider the sensor network architecture depicted in Figure 1.1. In the architecture SNs are grouped into clusters controlled by a single command node. Sensors are only capable of radio-based short-haul communication and are responsible for probing the environment to detect a target/event. Every cluster has a gateway node that manages sensors in the cluster. Clusters can be formed based on many criteria such as communication range, number and type of sensors and geographical location. Sensors receive commands from and send readings to its gateway node, which processes these readings.

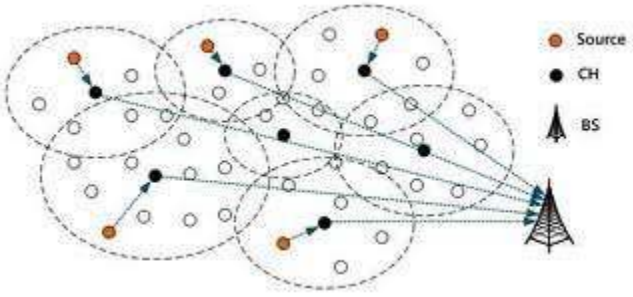

Figure 1.1: Sensor Network Architecture

The gateway node sends to the command node reports generated through fusion of sensor readings, e.g. tracks of detected targets. The command node presents these reports to the

user and performs system-level fusion of the collected reports for overall situation awareness.

3 INTRODUCTION TO OUTLIER DETECTION

Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior. These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains. Outlier detection has been a widely researched problem and finds immense use in a wide variety of application domains such as credit card, insurance, tax fraud detection, intrusion detection for cyber security, fault detection in safety critical systems, military surveillance for enemy activities and many other areas.

The importance of outlier detection is due to the fact that outliers in data translate to significant (and often critical) information in a wide variety of application domains. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination. In public health data, outlier detection techniques are widely used to detect anomalous patterns in patient medical records which could be symptoms of a new disease. Similarly, outliers in credit card transaction data could indicate credit card theft or misuse. Outliers can also translate to critical entities such as in military surveillance, where the presence of an unusual region in a satellite image of enemy area could indicate enemy troop movement. Or anomalous readings from a space craft would signify a fault in some component of the craft. Outlier detection has been found to be directly applicable in a large number of domains.

4 OUTLIER DETECTION IN WSN

In WSNs, outliers can be defined as, “those measurements that significantly deviate from the normal pattern of sensed data” [5]. This definition is based on the fact that in WSN sensor nodes are assigned to monitor the physical world and thus a pattern representing the normal behavior of sensed data may exist. Potential sources of outliers in data collected by WSNs include noise & errors, actual events, and malicious attacks. Noisy data as well as erroneous data should be eliminated or corrected if possible as noise is a random error without any real significance that dramatically aspects the data analysis [6].Outliers caused

by other sources need to be identified as they may contain important information about events that are of great interest to the researchers.

Recently, the topic of outlier detection in WSNs has attracted much attention. According to potential sources of outliers as mentioned earlier, the identification of outliers provides data reliability, event reporting, and secure functioning of the network. Specifically, outlier detection controls the quality of measured data, improves robustness of the data analysis under the presence of noise and faulty sensors so that the communication overhead of erroneous data is reduced and the aggregated results are prevented to be affected. Here, we exemplify the essence of outlier detection in several real-life applications.

- Environmental monitoring
- Habitat monitoring
- Health and medical monitoring
- Industrial monitoring
- Target tracking
- Surveillance monitoring,

It should be noted that several research topics have been developed for identifying sources of outliers occurred in WSNs. As illustrated in Figure 1, these topics include fault detection [7;8], event detection [9;10;11] and intrusion detection [12;13].

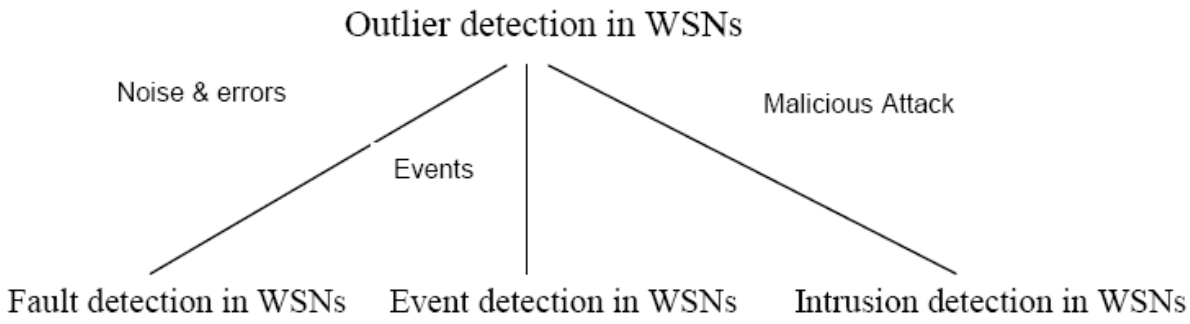

Figure 2.1: Outlier sources in WSNs and their corresponding detection techniques

5 TECHNIQUES DESIGNED FOR WSNS

Recently, many outlier detection techniques specifically developed for WSNs have emerged. In this section, we provide a technique-based taxonomy framework to categorize these techniques.

As illustrated in Figure 2.2, outlier detection techniques for WSNs can be categorized into statistical-based, nearest neighbor-based, clustering-based, classification-based, and spectral decomposition-based approaches. Statistical-based approaches are further categorized into parametric and non-parametric approaches based on how the probability distribution model is built.

5.1 Statistical-Based Approaches

The statistical outlier detection techniques are essentially model-based techniques. They assume or estimate a statistical (probability distribution) model which captures fit the model. A data instance is declared as an outlier if the probability of the data instance to be generated by this model is very low [5].

a. Parametric-Based Approaches.

Parametric techniques assume availability of the knowledge about underlying data distribution, i.e., the data is generated from a known distribution. It then estimates the distribution parameters from the given data.

– **Gaussian-based models**

[14] present two local techniques for identification of outlying sensors as well as identification of event boundary in sensor networks. In the technique for identifying outlying sensors, each node computes the difference between its own reading and the median reading from its neighboring readings. Then it standardizes all differences from its neighborhood. A node is considered as an outlying node if the absolute value of its reading's deviation degree is sufficiently larger than a pre-selected threshold. The technique of event boundary detection is based on 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 than that in another region [15] present a local outlier detection technique to identify errors and detect events in ecological applications of WSNs. Each node learns the statistical distribution of difference between its own measurements and each of its neighboring nodes, as well as between its current and previous measurements. A measurement is identified as anomalous if its value in the statistical significance test is less than a user-specified threshold. The drawback of this technique is that it relies on the choice of the appropriate values of the threshold.

– **Non-Gaussian-based models**.

[16] utilizes the spatio-temporal correlations of sensor data to locally detect outliers. Each node in a cluster first detects and corrects temporal outliers by comparing the predicted data and the sensing data. Then the clusterhead collects the rectified data from all other nodes in the cluster and further detects spatial outliers that deviate remarkably from other normal data.

b. Non-Parametric-Based Approaches.

Non-parametric techniques do not assume availability of data distribution. Two most widely used approaches in this category are histograms and kernel density estimator. Histogramming models involve counting frequency of occurrence of different data instances (thereby estimating the probability of occurrence of a data instance) and compare the test instance with each of the categories in the histogram and test whether it belongs to one of them. Kernel density estimators use kernel functions to estimate the probability distribution function (pdf) for the normal instances. A new instance that lies in the low probability area of this pdf is declared as an outlier [5].

- **Histogramming**.

[17] present a histogram-based technique to identify global outliers in data collection applications of sensor networks. The sink uses histogram information to extract data distribution from the network and filters out the non-outliers. The identification of outliers is achieved by a fixed threshold distance or the rank among all outliers.

- **Kernel functions**.

[18] propose a kernel-based technique for online identification of outliers in streaming sensor data. This technique requires no a priori known data distribution and uses kernel density estimator to approximate the underlying distribution of sensor data. Thus, each node can locally identify outliers if the values deviate significantly from the model of approximated data distribution.

[19] further extend the work of [18] and solve the two previous problems of insufficiency of a single threshold for multi-dimensional data and maintaining the data model built by kernel density estimator.

5.2 Nearest Neighbor-Based Approaches

Nearest neighbor-based approaches are the most commonly used approaches to analyze a data instance with respect to its nearest neighbors in the data mining and machine learning

community. [26] propose a technique based on distance similarity to identify global outliers in sensor networks. This technique attempts to reduce the communication overhead by a set of representative data exchanges among neighboring nodes. Each node uses distance similarity to locally identify outliers and then broadcasts the outliers to neighboring nodes for verification. The neighboring nodes repeat the procedure until the entire sensor nodes in the network eventually agree on the global outliers.

The other technique uses dynamic time warping (DTW) distance-based similarity comparison specifically for outliers that are erroneous and last for a certain time period. In this technique, each node transforms raw data into the wavelet time-frequency domain and identifies the high-frequency data measurements as outliers and corrects them using proper wavelet coefficients.

5.3 Clustering-Based Approaches

Data instances are identified as outliers if they do not belong to clusters or if their clusters are significantly smaller than other clusters.

[22] minimizes the communication overhead by clustering the sensor measurements and merging clusters before communicating with other nodes. Initially, each node clusters the measurements and reports cluster summaries rather than transmitting the raw sensor measurements to its parent. The parent then merges cluster summaries collected from all of its children before sending them to the sink. An anomalous cluster can be determined in the sink if the cluster's average inter-cluster distance is larger than one threshold value of the set of inter-cluster distances.

5.4 Classification-Based Approaches

They learn a classification model using the set of data instances (training) and classify an unseen instance into one of the learned (normal/outlier) class (testing). The unsupervised classification-based techniques require no knowledge of available labeled training data and learn the classification model which fits the majority of the data instance during training. The one-class unsupervised techniques learn the boundary around the normal instances while some anomalous instance may exist and declare any new instance falling outside this boundary as an outlier.

- **Support Vector Machine-Based Approaches.**

SVM techniques separate the data belonging to different classes by fitting a hyperplane between them which maximizes the separation. The data is mapped into a higher dimensional feature space where it can be easily separated by a hyperplane. Furthermore, a kernel function is used to approximate the dot products between the mapped vectors to find the hyperplane [5].

[22] propose that the sensor data that lies outside the quarter sphere is considered as an outlier. Each node communicates only summary information (the radius information of sphere) with its parent for global outlier classification. This technique identifies outliers from the data measurements collected after a long time window and is not performed in real-time.

- **Bayesian Network-Based Approaches.**

Bayesian network-based approaches use a probabilistic graphical model to represent a set of variables and their probabilistic independencies. They aggregate information from different variables and provide an estimate on the expectancy of an event to belong to the learned class [5].

6 CONCLUSION

We have studied the problem of outlier detection in WSNs and technique-based taxonomy framework which categorize the current outlier detection techniques designed for WSNs. Also advantages and disadvantages of some of the techniques are also given in the paper

REFERENCES

[1]. A.A. Abbasi and M. Younis, “A survey on clustering algorithms for wireless sensor networks”, Computer Communications, 30, 2826–2841, 2007.

[2]. Basilis Mamalis, Damianos Gavalas, Charalampos Konstantopoulos, and Grammati Pantziou, “Clustering in Wireless Sensor Networks”, pages 1-32, Zhang/RFID and Sensor Networks AU7777_C012 Page Proof Page 323 2009-6-24.

[3]. Bruce Schneier, “Applied Cryptography -Protocols, algorithms, and source code in C[M]”,Second edition, New York: John Wiley &Sons, 1996.

[4]. Chatterjee, Mainak and Das, Sajal K. and Turgut, Damla, “WCA: A Weighted Clustering Algorithm for Mobile Ad Hoc Networks”, Volume 5, pages 193-204. Springer Netherlands, 2002. 10.1023/ A: 1013941929408.

[5]. Chandola, V., Banerjee, A. and Kumar, V. (2007) 'Outlier detection: a survey', Technical Report, University of Minnesota.

[6]. Tan, P. N. (2006) 'Knowledge Discovery from Sensor Data', Sensors. Tan, P. N., Steinback. M. and Kumar, V. (2006) 'Introduction to data mining', Addison Wesley.

[7]. Chen, J., Kher, S. and Somani, A. (2006) 'Distributed fault detection of wireless sensor networks', Proceedings of the 2006 workshop on dependability issues in wireless ad hoc networks and sensor networks, pp. 65-72.

[8]. Luo, X., Dong, M. and Huang, Y. (2006) 'On distributed fault-tolerant detection in wireless sensor networks', IEEE Transactions on Computers, Vol. 55, No. 1, pp. 58-70.

[9]. Krishnamachari, B. and Iyengar, S. (2004) 'Distributed Bayesian algorithms for fault tolerant event region detection in wireless sensor networks', IEEE Transactions on Computers, Vol. 53, No. 3, pp. 241-250.

[10]. Martincic, F. and Schwiebert, L. (2006) 'Distributed event detection in sensor networks', Proceedings of the International Conference on Systems and Networks Communication, pp. 43-48.

[11]. Ding, M., Chen, D., Xing, K. and Cheng, X. (2005) 'Localized fault-tolerant event boundary de¬tection in sensor networks', Proceedings of IEEE Conference of Computer and Communications Societies, pp. 902-913.

[12]. Silva, A. P. R., Martins, M. H. T., Rocha, B. P. S., Loureiro, A. A. F., Ruiz, L. B. and Wong, H. C. (2005) 'Decentralized intrusion detection in wireless sensor networks', Proceedings of the 1st ACM international workshop on Quality of service & security in wireless and mobile networks, pp. 16-23.

[13]. Bhuse, V. and Gupta, A. (2006) 'Anomaly intrusion detection in wireless sensor networks', Journal of High Speed Networks, Vol. 15, No. 1, pp. 33-51.

[14]. Wu, W., X. Cheng, M. Ding, K. Xing, F. Liu, P. Deng (2007) 'Localized outlying and boundary data detection in sensor networks', IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 8, pp. 1145-1157.

[15]. Bettencourt, L. A., Hagberg, A. and Larkey, L. (2007) 'Separating the wheat from the chaff: practical anomaly detection schemes in ecological applications of

distributed sensor networks', Proceedings of IEEE International Conference on Distributed Computing in Sensor Systems.

[16]. Hida, Y., Huang, P. and Nishtala, R. (2003) 'Aggregation query under uncertainty in sensor networks. http : //www.cs.berkeley.edu/ rajeshn/pubs/cs252project.pdf.

[17]. Sheng, B., Li, Q., Mao, W. and Jin, W. (2007) 'Outlier detection in sensor networks', Proceedings of MobiHoc.

[18]. Palpanas, T., Papadopoulos, D., Kalogeraki, V. and Gunopulos, D. (2003) 'Distributed deviation detection in sensor networks', ACM Special Interest Group on Management of Data, pp. 77-82.

[19]. Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V. and Gunopulos, D. (2006) 'Online outlier detection in sensor data using non-parametric models', Journal of Very Large Data Bases.

[20]. Zhang, K., Shi, S., Gao, H. and Li, J. (2007) 'Unsupervised outlier detection in sensor networks using aggregation tree', Proceedings of ADMA. Zhang, Y., Meratnia, N. and Havinga, P. J. M. (2007) 'A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets', Technical Report, University of Twente.

[21]. Zhuang, Y. and Chen, L. (2006) 'In-Network outlier cleaning for data collection in sensor networks', Proceedings of VLDB.