Outlier Detection in Sensor Data using Machine Learning Techniques for IoT Framework and Wireless Sensor Networks: A Brief Study

Nimisha Ghosh, Krishanu Maity
Department of Computer Science and Information Technology,
Institute of Technical Education and Research,
Siksha 'O' Anusandhan (Deemed to be University),
Bhubaneswar 751030, Odisha, India
Email: nimishaghosh@soa.ac.in, krishanumaity@soa.ac.in

Rourab Paul, Satyabrata Maity
Department of Computer Science and Engineering,
Institute of Technical Education and Research,
Siksha 'O' Anusandhan (Deemed to be University),
Bhubaneswar 751030, Odisha, India
Email: rourabpaul@soa.ac.in, satyabratamaity@soa.ac.in

Abstract—**Outlier or anomaly detection in the sensed data for Internet of Things framework and Wireless Sensor Networks is a growing trend among researchers. Wireless Sensor Networks form the basis for Internet of Things framework in which the sensors sense a huge amount of data based on which certain actions or decisions or taken. So, the quality of data must be thoroughly checked as any kind of outlier may degrade the quality of the data and hence affect the final decision. Thus, it becomes imperative to maintain the quality of the data. In this work, some machine learning approaches have been discussed which have proved their mettle in outlier detection.**

Index Terms—**Machine Learning, Outlier Detection, Internet of Things (IoT), Wireless Sensor Networks (WSNs)**

I. Introduction

For any Internet of Things (IoT) and Wireless Sensor Networks (WSN) framework, sensors are the fundamental devices for generating data. They are responsible for sensing, processing and storing data. These data are important for many decision making policies [1]. Thus, the reliability of any data is of utmost importance in any application. The sensors have limited resources and capability, thus making the data generated by the sensors often unreliable and inaccurate. These sensors are battery-operated and when these batteries get replenished, the possibility of getting erroneous data increases by many folds. Environmental effects also play a major role in the operation of the sensor nodes. Be it IoT or WSN, the main aim is the data communication between devices without any human intervention. Thus, the sensors are also susceptible to malicious attacks in which the data are exploited by adversaries. All these factors contribute to the unreliability of the sensor data which ultimately influence the final decision making process. Thus, outliers can be deemed to a very important factor which affect the data quality. Any field which is concerned with data collection is majorly affected by outliers like weather monitoring [2], fraud and intrusion detection [3], sensor faults in heating, ventilation and air-conditioning (HVAC) systems [4], traffic anomaly detection [5] etc. In recent times, machine learning has proven to be a very powerful method to detect outliers in sensor data. In this brief survey, some of the works which have addressed the issue of outlier detection using machine learning techniques are outlined.

The rest of the paper is organised as follows: Section II presents the basics of outliers in IoT and WSN. Section III gives a summary of the different machine learning techniques used for detecting outliers in IoT and WSN followed by Section IV which provides the important reserach areas that need to be focused on. Finally, Section V concludes the paper.

II. Outlier Detection using Machine Learning in IoT and WSN

A. Preliminaries

Before delving deep into the survey, some preliminaries need to be outlined. One of the most popular definition of outlier is as follows:

Definition 1. *Outliers [6]: "an outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data".*

In the context of IoT and WSN, an outlier is any data measurement which deviates significantly from the normal set of sensed data. Based on the type, there are three kinds of outliers that can be seen in IoT based networks and WSN [7]; *Errors*, *Events* and *Malicious Attacks*. The detection techniques for these different faults are depicted in Fig. 1.

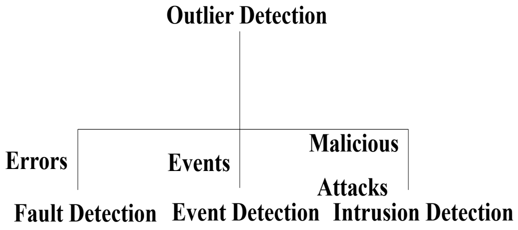

Fig. 1: Different types of outlier and their detection methods

978-1-7281-3908-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICAML48257.2019.00043

B. Type of Outliers

The three kinds of outliers as discussed before are now described in details:

- **Errors**: Error refers to the data coming from a faulty sensor data measurement. Outliers caused by errors are much more frequent than those of events [8]. It occurs due to random change in data and is very different from the other data. These errors affect the data quality and need to be detected so that the data can be used efficiently after removing them. But when the outliers are too erroneous, they are disposed of to conserve energy in the resource-constrained sensors which form the crux of any IoT and WSN framework. Most of the works discussed so far works with error detection in IoT and WSN.
- **Events**: It is considered to be a phenomenon which changes the state of the real-world, like forest-fire, air pollution etc. This kind of outlier lasts for a longer time than errors and tend to change the pattern of the data. But one this to notice is that faulty sensors may also generate this kind of sustainable errors, thus making it quite difficult to distinguish between the two. Thus, spatial correlation forms an important part in this regard, as erroneous sensor data is spatially unrelated but event measurements are spatially related [9].
- **Malicious Attacks**: Malicious attacks are well-known attacks in the context of IoT and WSN. In these type of attacks, when a node gets compromised it masquerades as a normal node to mislead the other nodes [10]. So this can be deemed to be a very serious threat to the network security. IoT devices usually use wireless communication for transferring data, making them vulnerable to security threats. Intrusion detection forms a very important part of detecting malicious attacks. Thus, intrusion detection systems can be implemented to detect these types of outliers in any sensor networks.

  Fig. 2 gives the generic architecture for outlier detection in IoT and WSN.

III. Summary of Machine Learning Techniques in IoT and WSN

There are broadly two outlier detection techniques using machine learning methods which can be broadly classified as given below [7]:

A. Statistical based Methods

These are the model based approaches which assume some kind of probabilistic distribution of the data. Based on how well the data fit the model, the data instances are evaluated. A data instance is considered to be an outlier if the possibility of this data instance to be produced by the model is quite low. The statistical based methods can be largely categorised into parametric and non-parametric approaches.

- ***Parametric methods*** : Parametric approaches assume that the generated data has a known distribution. The distribution parameters are then estimated from the given data set. In [11], the authors have used logistic regression which is a popular parametric method for detecting anomaly or outlier in sensed data.
- ***Non-parametric methods*** : On the other hand, non-parametric methods do not make any prior assumptions about data distribution or the mapping function. Rather they try to best fit the training data to construct the mapping function. thus, they have the ability to fit many functional forms. Nesa et. al [12] used a non-parametric method based on Grey Systems Theory to detect outliers in an IoT framework.

Though parametric methods are simple and they have the ability of learning fast from the data, they suffer from the drawback of a-priori assumption of the data distribution. Contrary to this, non-parametric methods as discussed earlier do not assume anything about the data distribution making it more pliable for any kind of data. But they require a lot of training data to create a mapping function.

B. Classification based Methods

Classification based methods can be considered to be a subset of the statistical learning methods. They work with making the model first learn the intricacies of the training data which encompasses both normal and faulty/outlier data. Then, based on this learning policy, they can classify the unknown test data into either of the two classes. Classification based methods can be divided into supervised and unsupervised learning.

- ***Supervised Learning*** : In supervised learning, a prior knowledge about the data classes are known i.e. both inputs and the corresponding outputs are given. Some popular supervised learning algorithms are:
  1) Support Vector Machine (SVM) : SVM is a classification method which uses the concept of hyperplane to distinguish between two classes. To determine the hyperplane, SVM maximises the margin and tries to distinguish between the classes with minimal errors. Zidi et. al [13] used SVM which is a popular supervised learning method to detect faults in sensor data where the decision is taken at a cluster head to detect the faulty sensor.
  2) k-nearest Neighbour (k-NN): In this method, prediction for a new observation is made by searching through k nearest neighbours based on any distance measures, the most popular being the Euclidean distance. In [14], the authors have used kNN for online anomaly detection in WSNs. In this paper, the authors have proposed a new method based on hypergrid intuition to mitigate the problem of lazy learning of k-NN.
  3) Neural Networks : Neural networks aim to understand the underlying relationship in a data set by mimicking the activities of human brain. This method consists of different layers and is used to solve complex problems. Though there are many works which use neural networks for outlier detection, but the complex nature of the computations provide a hindrance to the

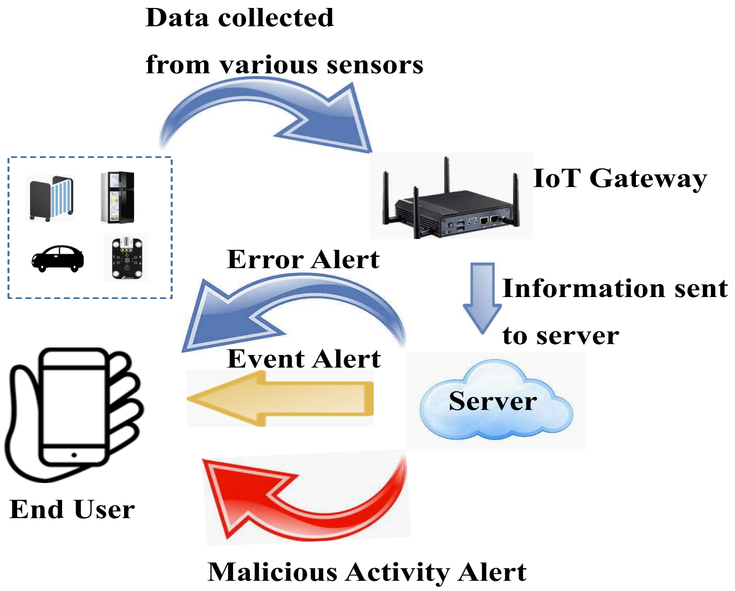

Fig. 2: A generic architecture for different outlier detection in IoT and WSN

application of neural networks in resource-constrained IoT and WSN. Luo et. al [15] have used a modified neural network technique called the autoencoder neural networks to determine anomalies in sensors. They have also implemented their proposed method on a test bed to show the applicability of the said method.

4) Bayesian Learning Method: This method uses probability distribution as its learning parameter. Thus, a prior knowledge about the data set is a must for applying the Bayesian learning methods. Sadreazami et. al [16] have proposed a novel intrusion detection method based on modified Bayesian likelihood ratio test. The results show that the proposed intrusion detection scores over the other existing state-of-the-art methods.

- ***Unsupervised Learning*** : In unsupervised learning, only the input data is given but no corresponding output variables. The goal is to model the distribution in the data to learn more about the data. Based on the inputs, the learning models group them into clusters. Some of the most used unsupervised learning methods are as follows:

1) k-means Clustering: In this method, the aim is to find k number of groups in the data. The algorithm works by iteratively assigning each input data point to one of the k groups based on their underlying feature similarity. In [17], the authors have used genetic k-means algorithm for intrusion detection to assure network security.

2) Principal Component Analysis (PCA) : PCA is a procedure to calculate projection of the original data set into a new dimension of fewer components, thus turning large data sets into smaller ones. PCA is very popular technique for detecting outliers in sensor data. One of the important works in this regard is [18]. In this work, the authors have used recursive PCA which is a variant of the original PCA to aggregate redundant data and detect outliers in sensor data.

Classification based methods mostly provide a very accurate set of outliers by creating a classification model. But, one of the main drawbacks of SVM is its computational complexity and choosing the proper kernel function. Also, for Bayesian models if the number of variables is large, creating an accurate model poses quite a challenge. Table I gives a glimpse into the various types of machine learning methods used to detect outliers in IoT and WSN.

IV. Evaluation and Research Areas

A. Evaluation of outlier detection methods

Mostly, machine learning techniques focus on providing a good accuracy while detecting outliers in sensors. Any outlier detection should have high detection rate and should maintain a low false positive rate. Detection rate refers to the percentage of outliers correctly identified as outliers while false positive rate means the percentage of normal observation incorrectly identified as outliers. Receiver Operating Characteristic (ROC) curve is used to represent the trade-off between these two parameters. The larger the area under the curve, the better is the performance of a machine learning technique.

B. Research Areas

Though there have been some significant works for outlier detection in IoT and WSN, there are several areas in outlier detection which still need to be further explored. Some of these include:

- Use of ensemble methods to detect outliers in IoT and WSN.

TABLE I: Summary of some Machine Learning Methods used for Outlier Detection

Techniques	Type of Outlier	Learning Method	IoT/WSN
Zidi et. al [13]	Error	SVM	WSN
Nesa et. al [12]	Error and Event	Grey Systems Theory	IoT
Nesa et. al [19]	Error and Event	Supervised Learning	IoT
Xie et. al [14]	Malicious Attacks	k-NN	WSN
Ni et. al [20]	Error	Bayesian networks	WSN
Krishnamachari et. al [9]	Event	Bayesian networks	WSN
Saneja et. al [21]	Error	SVM and k-means Clustering	IoT
Diro et. al [22]	Malicious Attacks	Neural Networks	IoT
Pahl et. al [23]	Malicious Attacks	k-means Clustering	IoT
Pajouh et. al [24]	Malicious Attacks	Bayesian Learning Method and kNN	IoT

- Use of machine learning techniques and semantics to differentiate between errors and events.
- Merge offline and online learning techniques for improved outlier detection.

V. Conclusion

In this work, various machine learning techniques which detect outliers in IoT and WSN have been described. In IoT and WSN frameworks, sensors form the crux of generating raw data and are also responsible for detecting environmental changes. Thus, detecting outliers is much needed to analyse error free data generated from sensors. Some works have also been tabulated which are useful in detecting various kinds of outliers in sensor data. It can ne easily concluded from the discussion that classification methods are the most extensively used learning methods for detecting outliers in IoT and WSN. The existing shortcomings in both IoT and WSN require developing more suitable outlier detection techniques for both univariate and multivariate data. Also, while developing new machine learning methods mobility of node and network topology change should also be seriously considered.

References

[1] N. Ghosh and I. Banerjee, "Iot-based freezing of gait detection using grey relational analysis," *Internet of Things*, p. 100068, 2019.

[2] J. M. Shepherd and S. J. Burian, "Detection of urban-induced rainfall anomalies in a major coastal city," *Earth Interactions*, vol. 7, no. 4, pp. 1–17, 2003.

[3] S. Hajiheidari, K. Wakil, M. Badri, and N. J. Navimipour, "Intrusion detection systems in the Internet of things: A comprehensive investigation," *Computer Networks*, 2019.

[4] V. Reppa, P. Papadopoulos, M. M. Polycarpou, and C. G. Panayiotou, "A distributed architecture for hvac sensor fault detection and isolation," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 4, pp. 1323–1337, July 2015.

[5] K. Xie, X. Li, X. Wang, J. Cao, G. Xie, J. Wen, D. Zhang, and Z. Qin, "On-line anomaly detection with high accuracy," *IEEE/ACM Transactions on Networking*, vol. 26, no. 3, pp. 1222–1235, June 2018.

[6] V. Barnett and T. Lewis, *Outliers in statistical data*. Wiley, 1974.

[7] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 12, no. 2, pp. 159–170, Second 2010.

[8] F. Martincic and L. Schwiebert, "Distributed event detection in sensor networks," *2006 International Conference on Systems and Networks Communications (ICSNC'06)*, pp. 43–48, 2006.

[9] B. Krishnamachari and S. Iyengar, "Distributed bayesian algorithms for fault-tolerant event region detection in wireless sensor networks," *IEEE Transactions on Computers*, vol. 53, no. 3, pp. 241–250, March 2004.

[10] M. Alotaibi, "Security to wireless sensor networks against malicious attacks using hamming residue method," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 8, Jan 2019.

[11] M. Hasan, M. M. Islam, M. I. I. Zarif, and M. Hashem, "Attack and anomaly detection in iot sensors in iot sites using machine learning approaches," *Internet of Things*, vol. 7, p. 100059, 2019.

[12] N. Nesa, T. Ghosh, and I. Banerjee, "Non-parametric sequence-based learning approach for outlier detection in iot," *Future Generation Computer Systems*, vol. 82, pp. 412 – 421, 2018.

[13] S. Zidi, T. Moulahi, and B. Alaya, "Fault detection in wireless sensor networks through svm classifier," *IEEE Sensors Journal*, vol. 18, no. 1, pp. 340–347, Jan 2018.

[14] M. Xie, J. Hu, S. Han, and H. Chen, "Scalable hypergrid k-nn-based online anomaly detection in wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 8, pp. 1661–1670, Aug 2013.

[15] T. Luo and S. G. Nagarajan, "Distributed anomaly detection using autoencoder neural networks in wsn for iot," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[16] H. Sadreazami, A. Mohammadi, A. Asif, and K. N. Plataniotis, "Distributed-graph-based statistical approach for intrusion detection in cyber-physical systems," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 1, pp. 137–147, March 2018.

[17] Sandhya G and A. Julian, "Intrusion detection in wireless sensor network using genetic k-means algorithm," in *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, May 2014, pp. 1791–1794.

[18] T. Yu, X. Wang, and A. Shami, "Recursive principal component analysis-based data outlier detection and sensor data aggregation in iot systems," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 2207–2216, Dec 2017.

[19] N. Nesa, T. Ghosh, and I. Banerjee, "Outlier detection in sensed data using statistical learning models for iot," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[20] K. Ni and G. Pottie, "Sensor network data fault detection with maximum a posteriori selection and bayesian modeling," *ACM Trans. Sen. Netw.*, vol. 8, no. 3, pp. 23:1–23:21, Aug. 2012.

[21] B. Saneja and R. Rani, "A hybrid approach for outlier detection in weather sensor data," in *2018 IEEE 8th International Advance Computing Conference (IACC)*, Dec 2018, pp. 321–326.

[22] A. A. Diro and N. Chilamkurti, "Distributed attack detection scheme using deep learning approach for internet of things," *Future Generation Computer Systems*, vol. 82, pp. 761 – 768, 2018.

[23] M. Pahl and F. Aubet, "All eyes on you: Distributed multi-dimensional iot microservice anomaly detection," in *2018 14th International Conference on Network and Service Management (CNSM)*, Nov 2018, pp. 72–80.

[24] H. H. Pajouh, R. Javidan, R. Khayami, A. Dehghantanha, and K. R. Choo, "A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in iot backbone networks," *IEEE Transactions on Emerging Topics in Computing*, vol. 7, no. 2, pp. 314–323, April 2019.