International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Review Paper Volume-10, Issue-4 ,2023 Regular Edition E-ISSN: 2349-7084

Comparative Study on Techniques Used for Anomaly Detection in IoT Data

Bezawada Manasa[1], P Venkata Krishna[2]

[1]*Dept. of Computer Science & Engineering, Sri Padmavathi Mahila Visvavidyalayam, Tirupati, Andhra Pradesh, India*
[2]*Dept. of Computer Science, Sri Padmavathi Mahila Visvavidyalayam, Tirupati, Andhra Pradesh, India*

e-mail: manasabezawada04@gmail.com, pvk@spmvv.ac.in
**Corresponding Author:* manasabezawada04@gmail.com,

Available online at: http://www.ijcert.org

https://doi.org/10.22362/ijcert/2023/v10/i04/v10i0406

Received: 15/03/2023, Revised:09/04/2023, Accepted: 18/04/2023 Published:28/04/2023

Abstract:-The Internet of Things (IoT) makes it possible to connect various devices using wireless and cellular technology. As the foundation of Internet of Things (IoT), data from the target regions are gathered by widely dispersed sensing devices and delivered to the processing unit for aggregation and analysis. IoT service quality typically depends on reliability and integrity of data. However, IoT data gathered will be anomalous because of the unfavourable environment or equipment flaws. In order to ensure service quality, an efficient technique of anomaly detection is therefore essential. Finding new or unexpected things in the collected data is called Anomaly detection. The most important developments in recent years that enable automatic feature extraction from raw data are deep learning and machine learning. Role of machine and deep learning techniques to detect anomalies in sensor data is reviewed in this article. Finally, we provide a summary of the difficulties encountered currently in the anomaly detection field to identifying potential future research prospects.

Keywords: IoT, Anomaly, Machine Learning, Deep Learning, Time Series, Data Stream

1. Introduction

By 2025, the number of devices, which are connected in IoT networks is predicted to reach nearly 75 billion, which is three times of world's population [1]. Internet of Things (IoT) is network of disparate items that are linked to the Internet using a variety of technologies, including smartphones, laptops, intelligent devices, and sensors. The Internet of Things (IoT) allows direct user-free communication between various sensors and devices. .IoT is used in many applications like Agriculture, Industry, Smart Cities, Healthcare, and Home Automation. From the past few years, IoT has become the largest data sources. Because of flaws in sensors or environment an unexpected event may be observed in IoT data. Unexpected events in IoT data are called Anomaly. Analysing the collected data and identifying the anomalies is called Anomaly detection. Most of the IoT data is time series data. Time series data may be univariate time series data (data collected from a sensor at different time intervals) or multi variate time series data (data collected from multiple sensors at different time intervals).

Types of Anomalies

a) Point Anomalies

The one kind of anomalies are point anomalies. The time series' return to its prior normal state within a very brief time frame of only a few observations is a key feature of these anomaly types. These point anomalies might be statistical noise, they might be caused by malfunctioning sensing apparatus, or they might be an important short-term event that the system's operators are interested in.

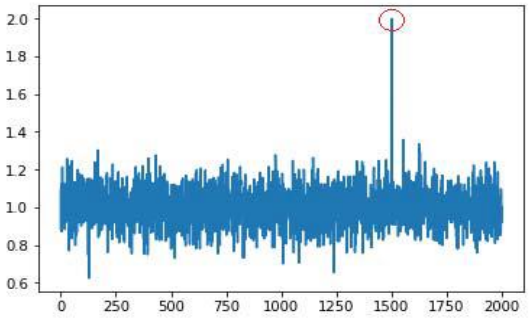
b) **Contextual Anomalies:**

These are sequences are observations that deviate from the time series' anticipated patterns but, when considered

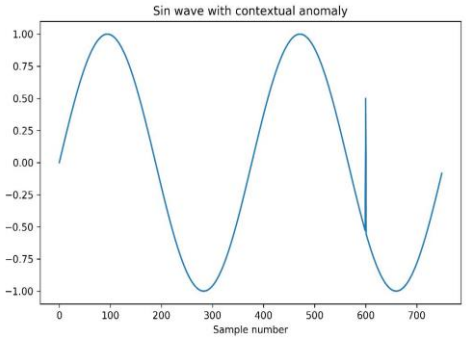
separately, may fall within the acceptable range for that signal. An anomaly that deviates from the standard when viewed in the light of the nearby observations is called a contextual anomaly.

c) **Collective Anomalies:**

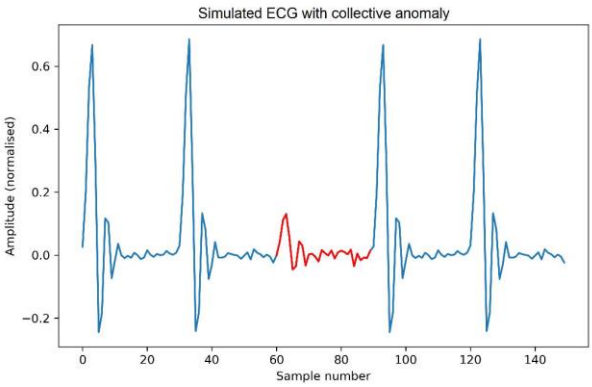
A group of data that are anomalous in comparison to remainder of the data is referred to as a collective anomaly. A collective anomaly's individual observations may or may not be anomalous; it is only when they show together that they raise questions.



Point Anomaly



Contextual Anomaly



Collective Anomaly

Fig: Anomalies Types [2]

Rest of the paper is organized as follows, Section I contains the introduction of IoT data and Anomalies, Section II contain the related work of Anomaly Detection Techniques Section III contain the architecture and essential steps of Anomaly Detection, section IV explain the Performance Evolution andd Section V concludes research work with future directions.

2. Related Work

The techniques which are used for detecting the anomalies are using the classification method. Techniques try to classify the data whether it is normal data or anomaly. Anomaly detection can be performed by using Conventional Techniques, machine learning methods or deep learning methods. This paper presents a few machine learning methods and deep learning methods used to identify anomalies in IoT data.

2.1 Anomaly detection using machine learning techniques

Numerous tasks, such as classification, intrusion detection, regression, recommendation systems and computer vision have made extensive use of machine learning methods. This paper focus on classification techniques which are used for anomaly detection in IoT data.

a. DBSCAN

DBSCAN is one of the clustering technique in Machine Learning. The methodology followed in DBSCAN is the data with high dense neighbours was identified as normal data points and data with low dense that is less number of neighbours was assumed as anomaly points. To manage multivariate time series data DBSCAN considers time window as data point and to separate data and cluster anomaly score is used. The anomaly score technique used by DBSCAN is aggregate score.

b. Isolation Forest:

The base for the Isolation Forest (IF) method is the decision Trees. For every time window IF [4] selects one anomaly score value. To generate Anomaly score algorithm takes one sample and threshold value and try to isolate the sample iteratively. Number of iterations to be considered depends on the threshold value. A sample is considered as anomaly if isolation was performed with less number of iterations otherwise sample is normal.

c. OC-SVM (One-Class Support Vector Machine)

The OC-SVM [5] considers hyper sphere while detecting anomalies in the data. The methodology followed in this technique is all the internal data points of hyper sphere is considered as normal point and external data points of hyper sphere are considered as anomalous points. This method as takes signed distance for identifying anomaly point. The data point is considered as normal if distance between hyper sphere and data point is positive. The distance is negative; data point is anomaly. For multivariate time series data OC-SVM uses time window. J. Ma and S. Perkins [6] developed combining output of OC-SVM to different time windows

d. K-Nearest Neighbor

K-NN , supervised machine learning technique is used to perform regression and classification. Classification is performed by considering the close proximity of the data points. In the manufacturing industry to identify the cyber-attacks Wu et al [7] proposed the machine learning method K-NN. For Anomaly detection [8] proposed K-NN method

e. Decision Tree

Regression or classification methods are built using a tree structure like technique called Decision Tree. This method divides the dataset into manageable categories and also gradually building tree by adding features to the decision tree in the industrial system. In IoT networks a decision tree is used to detect anomalies and attacks [9].

2.2 Anomaly detection by using deep learning techniques

In recent days deep learning-based methods improve anomaly detection in multi -dimensional IoT data. These methods can effectively capture temporal correlation and can model complex, highly nonlinear interactions between numerous sensors. This section presents different deep learning methods used for anomaly detection.

a. Recurrent Networks

RNN is used recently in many DNN-based prediction models. Regression concept is used to predict the succeeding values depending on the past values and prediction error is calculated. The calculated prediction error is used to check whether the data point is anomaly or normal. RNN is FF (Feed forward) neural network with internal memory. Architecture of RNN consist of input layer, hidden layer and output layer. RNN is used for the analysis of the time series data, text data and videos

b. Generative Networks

GAN ("Generative Adversarial Networks") is neural network of unsupervised, consist of two different neural networks. One neural network is named as G (Generator) and another is D (discriminator). In two players' min-max game [10] the two nets G and De trained parallel. The task of discriminator is to detect fake data from the actual data and generator is used to create synthetic data. In anomaly detection for training purpose normal data is considered as input and discriminator identifies anomalies by observing the deviation from learned data. TAnoGAN ("Time Series Anomaly Detection with Generative Adversarial Networks") [11] and MAD-GAN ("Multivariate Anomaly Detection with Generative Adversarial Networks") [12] are GAN based techniques for anomaly detection.

c. Graph based method

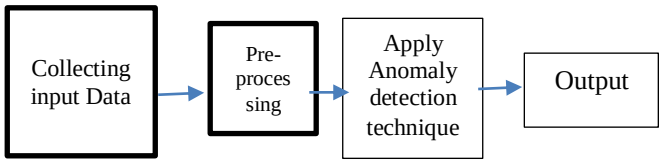
In recent days, Graphs are used to detect anomaly data in multi variate time series IoT data. With the help of graphs temporal dependencies among the sensor data was identified. In graph, nodes represent data and edges represent correlation between data. GDN (Graph Deviation Network), which is a type of GNN uses forecasting based on graph attention [13]. MTAD-GAN ("Multivariate Time-series Anomaly Detection via Graph Attention Network") [14] learns the multivariate time series data intricate dependencies with the help of two graph attention layers. Combination of forecasting-based and reconstruction based methods improves the performance of MTAD-GAN.

d. Transformer

Transformers are used to process sequential data in Neural Networks environment. Modelling of very long-term temporal context information or temporal relationships is difficult in some of the deep learning techniques like LSTM, RNN and GRU. To overcome the drawback of these technique, transformers concept was introduced. The architecture of Transformer consists of encoder and decoder. TranAD with adversial training method proposed in [15]. For anomaly detection Reconstruction-based transformer was proposed [16]. Combination of graphs and transformers concept was prosed [17] for Anomaly detection.

3. Methodology

Anomaly detection in IoT data consists of different steps like collecting input data, pre-processing, Anomaly detection techniques, Output.



a. Collecting Input Data:

In IoT environment Input data was collected from sensors. Sensor data was given as input to the pre-processing step. The sensor data which was collected may be univariate or multivariate time series data.

b. Pre-Processing:

IoT data is noisy and multi-dimensional data so before applying anomaly detection technique we need to perform some Pre-Processing. The missing and noisy data was handles in this step. IoT data is multi-dimensional data instead of taking all the features of data, some important features were extracted by using feature extraction methods.

c. Apply Anomaly Detection Technique:

The output of pre-processing step will be given as input to the anomaly detection technique. In this step analysis of IoT data was performed to check whether the data is normal or anomaly.

d. Output:

At output phase the data was classified as anomaly data or normal data.

4. Performance Evolution

The "accuracy" metric was employed to gauge how well anomaly detection methods performed. However, the stated accuracy will not provide a true picture of the method's effectiveness in the case of unbalanced datasets. Metrics like F1 scores, recall, precision, True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) are used to evaluate performance more precisely. The number of correctly detected positive cases is disclosed by TP. The number of negative instances that are correctly classified as negative cases are disclosed by TN. FP presents the details pertaining to the cases that were incorrectly classified as positive. Similar to this, FN provides details about data that are actually positive in nature but were mistakenly classified like negative. The proportion is defined as, correctly classified as class members over all instances classified as class members is known as precision. Recall, defined as proportion of properly classified class members over all class members. To create a high-quality method to detect anomaly, the requirement of precision and recall must be high.

TPR (True Positive Rate), also known as recall or sensitivity, is defined as the ratio of true positives to the total number of actual positives:

$$TPR = TP / (TP + FN)$$

where TP is the number of true positives (correctly classified positive examples) and FN is the number of false negatives (incorrectly classified negative examples).
FPR (False Positive Rate) is defined as the ratio of false positives to the total number of actual negatives:

$$FPR = FP / (FP + TN)$$

where FP is the number of false positives (incorrectly classified positive examples) and TN is the number of true negatives (correctly classified negative examples).
Precision is defined as the ratio of true positives to the total number of predicted positives:

$$Precision = TP / (TP + FP)$$

Recall is another name for TPR, as defined above.
F1-score is a harmonic mean of precision and recall, which gives equal importance to both measures. It is defined as:

$$F1 - score = 2 * (Precision * Recall) / (Precision + Recall)$$

Efficiency is often used to refer to the overall performance of a classification model, and can be calculated using various measures, such as accuracy, F1-score, or area under the ROC curve. It is generally used to compare the performance of different models or algorithms.

5. Conclusion and Future Scope

The main problem with most AD machine learning approaches is that anomalies are uncommon occurrences that are not recorded in actual life. The difficulty comes in defining a methodology that can distinguish rare events from the other nominal events. Here is a process flow that demonstrates distinct anomaly detection methods such as classification, clustering, nearest neighbour, statistical distribution, information theory, spectral analysis, and graph analysis. Each category offers an algorithm and its AD applications in various domains, along with the assumption "when giving a dataset, what instances are considered as anomalies and what are nominal." An optimal ADT is unaffected by labels or unsupervised learning, independent of distribution, quick, flexible with respect to data types, and high dimensional. Deep learning techniques have the ability of automatic feature extraction, which gives better performance compare to machine learning techniques. For some complex IoT data single DL technique itself is not gives best performance. Hybrid approach that is using combination of more than one deep learning technique can efficiently detect anomalies in uni-variant as well as multi-variant IoT time series data.

References

[1] Ratasich, Denise and Khalid, Faiq and Geissler, Florian and Grosu, Radu and Shafique, Muhammad and Bartocci, Ezio. "A Roadmap Toward the Resilient Internet of Things for Cyber-Physical Systems", IEEE Access, 2019.
[2] A. A. Cook, G. Mısırlı and Z. Fan, "Anomaly Detection for IoT Time-Series Data: A Survey," in IEEE Internet of Things Journal, vol. 7, no. 7, pp. 6481-6494, July 2020, doi: 10.1109/JIOT.2019.2958185.

[3] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discoverng clusters in large spatial databases with noise, in: Proceedings of the Second ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, p. 226231 , 1996.
[4] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation forest, in: Proceedings of the 2008 IEEE International Conference on Data Mining (ICDM), pp. 413–422 , 2008.
[5] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, R. Williamson, Estimating sup-port of a high-dimensional distribution, Neural Comput. 13, 1443–1471 , 2001.
[6] J. Ma, S. Perkins, Time-series novelty detection using one-class support vec-tor machines, in: Proceedings of the International Joint Conference on Neural Networks, pp. 1741–1745 , 2003.
[7] Wu, M.; Song, Z.; Moon, Y.B. Detecting cyber-physical attacks in CyberManufacturing systems with machine learning methods. J.Intell. Manuf. , 30, 1111–1123, 2019.
[8] Gunupudi, R.K.; Nimmala, M.; Gugulothu, N.; Gali, S.R. CLAPP: A self-constructing feature clustering approach for anomaly detection. Future Gener. Comput. Syst. , 74, 417–429, 2017.
[9] Hasan, M.; Islam, M.; Zarif, I.I.; Hashem, M. Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. Internet Things , 7, 100059, 2019.
[10] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.;Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. ACM, 63, 139–144, 2020.

[11] Bashar, M.A.; Nayak, R. TAnoGAN: Time Series Anomaly Detection with Generative Adversarial Networks. In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence, SSCI, Canberra, ACT, Australia, 1–4 pp. 1778–1785, December 2020.
[12] Li, D.; Chen, D.; Jin, B.; Shi, L.; Goh, J.; Ng, S.K. MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks. Int. Conf. Artif. Neural Netw., 11730, 703–716, 2019.
[13] Deng, A.; Hooi, B. Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. Proc. AAAI Conf. Artif. Intell. , 35, 4027–4035, 2021.
[14] Zhao, H.;Wang, Y.; Duan, J.; Huang, C.; Cao, D.; Tong, Y.; Xu, B.; Bai, J.; Tong, J.; Zhang, Q. Multivariate time-series anomaly detection via graph attention network. In Proceedings of the IEEE International Conference on Data Mining, ICDM, Sorrento, Italy, 17–20 , pp. 841–850, 2020.
[15] Meng, H.; Zhang, Y.; Li, Y.; Zhao, H. Spacecraft Anomaly Detection via Transformer Reconstruction Error. Lect. Notes Electr. Eng. , 622, 351–362, 2020.
[16] Tuli, S.; Casale, G.; Jennings, N.R. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. Proc. VLDB Endow. , 15, 1201–1214, 2022.
[17] Chen, Z.; Chen, D.; Zhang, X.; Yuan, Z.; Cheng, X. Learning Graph StructuresWith Transformer for Multivariate Time-Series Anomaly Detection in IoT. IEEE Internet Things J. , 9, 9179–9189, 2022.

