RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 02 February
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.012

The Enhanced Anomaly Deduction Techniques for Detecting Redundant Data in IoT

S Subha[1], *J G R Sathiaseelan*[1]

[1]*Department of Computer Science, Bishop Heber College (Bharathidasan University), Trichy17, Tamilnadu, India*

Emails: subhaebenezerraja@gmail.com , jgrsathiaseelan@gmail.com

Article History

Received: 4 January 2023
Accepted: 14 February 2023

Keywords:

IoT;
Machine Learning;
Anomaly detection techniques

Abstract

Anomaly detection in Internet of Things is a challenging issue and is being addressed in a wide range of domains, including fraudulent detection, malware protection, information security and diagnosis of diseases. Due to the distributed nature of wireless transmission and the insufficient resources of end nodes, traditional anomaly detection techniques cannot be used in IoT directly. To extract uncommon behaviors or patterns from complex data, nevertheless, is a difficult task. As a result, this paper offers a thorough analysis of ML based methods to identify anomaly in the IoT healthcare data. Further, a detailed comparison of their performance is provided with reference to their benefits and disadvantages.

1. Introduction

The study of anomalies is highly valued in the IoT industry. It alludes to the process of extracting patterns from the data whose actions differ from those anticipated. An object is typically referred to as an outlier or anomaly if its behavior differs noticeably from the rest of the population. Depending on the particular application scenes, the term "anomaly" may also refer to an anomaly, discordant object, exception, aberration, surprise, or strangeness. Due to its ability to uncover uncommon but significant phenomena and identify intriguing or unexpected patterns, anomaly detection is extremely essential in a number of fields, including decision-making, clustering, and pattern categorization (Larriva-Novo et al.). Anomaly detection has recently taken the top spot among data-related concerns (Yu et al.). Fig. 1 depicts the anomaly detection types. The techniques of supervised anomaly detection involve training a classifier relating to a separated data collection into "regular" and "unusual" groups. Therefore, this technique is rarely used in anomaly detection because of the limited amount of labelled data and the classes' inherent unbalance. However, when the unlabelled dataset contains a tiny fraction of aberrant data, this assumption always results in performance reduction.

Unsupervised anomaly detection involves an unlabelled dataset. It assumes the most of the data sets in the unlabelled dataset are "normal" and it searches for information that varies from the "normal" data points. This document's remainder is divided into the following sections: The duties associated with anomaly deduction are discussed in Section 2. The anomaly deduction techniques are explained in Section 3.

2. Review of Literature

Goldstein and Uchida. (Jui, Hoq, and Majumdar) demonstrate a comparison of various detection of anomaly approach for several types of data sets. The paper largely focuses on multivariant tabular data and excludes newer DL based approaches, although

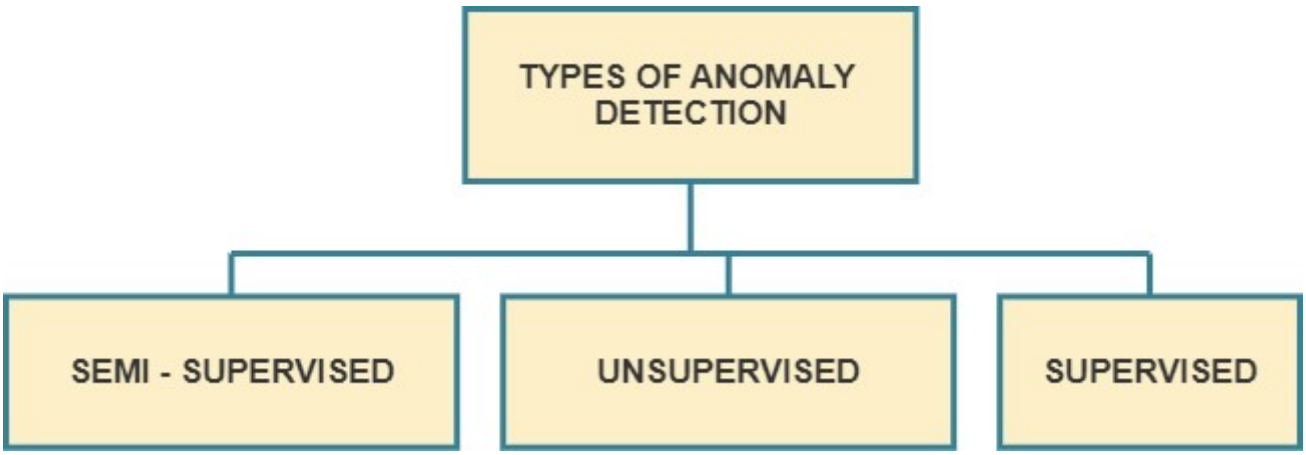

FIGURE 1. Types of Anomaly Detection

comparing several approaches based on the same data sets to provide an enhanced comparison.

Gupta et al. (Pavithra, Anandhakumar, and Meenashisundharam) have conducted a thorough investigation on temporal data anomaly detection methods. However, Comparing of DL-based approaches and analysis are absent from their work, which nonetheless offers a thorough overview of many technologies used in numerous temporal data sets.

Kiran et al. (Sonawane) presents a description of techniques based on DL for anomaly identification in videos. Adewumi et al (Alcalde-Barros et al.) gives an extensive a description of techniques based on deep learning used in the investigation of fraud. Hodge et al (Nematzadeh, Ibrahim, and Selamat) comprehensively surveys statistics and early methods for anomaly identification based on machine learning.

Huang et al (Sun et al.) created the RBDA anomaly detection algorithm (Rank Based Detection algorithm). As a measure of how close an object is to its neighbors, it uses their ranks. Tang and L He (Li et al.) developed an approach for anomaly detection that used k nearest neighbors, reversed nearest neighbors.

Chandola et al. (Carrera et al.) provides a complete review of anomaly detection techniques which incorporates not only techniques based on fundamental machine, there are various techniques that use statistical approaches and information theory, including nearest neighbors and clustering. Be aware that the neighbor rankingbased methods are sensitive to the model's parameter k, making it challenging to select the appropriate k for various applications. To cope with this problem,

Ha et al (Nematzadeh, Ibrahim, and Selamat) presented an iterative random sampling process and a heuristic method to choose the

value of k. In each sample, greater inlier Ness scores should be assigned to the picked objects because it is assumed that outlying objects are less likely to be chosen than inlying ones in random sampling.

Huang et al (Sun et al.) pointed out that only subsets of relevant characteristics can provide valuable

information for an item with many attributes; residual attributes are useless for the task and may make the anomaly detection model difficult to separate. Therefore, identifying outliers from relevant subspaces will be a fun and effective task

Aggareal and Yu. (Li et al.) utilized an evolutionary approach to obtain the subspaces; in this manner, the

subspace with the most negative scarcity coefficients was regarded as a space projection. However, the effectiveness of this algorithm greatly depends on the baseline populations.

Jianwu Wang et al. (Carrera et al.) recommended a (SAPIM) framework to precisely detect system anomalies from the sensor data to prevent further harm and save production maintenance expenses.

3. Anomaly Detection Techniques

The following major categories can be used to classify anomaly detection techniques are classification based, clustering based, nearest neighbour based and statistical based. The primary assumption of classification- based techniques is that the differentiation among abnormal and typical occurrences can

also be described for a specific feature space. Nearest neighbour-based methods make the assumption that irregularities are located far from their nearest neighbour in sparse neighbourhoods. They are mainly unsupervised algorithm. Clustering based techniques work by grouping of similar objects into clusters and consider that anomalous must be either unrelated to each and every group, far from the middle of its group, or part of tiny, sparsely clusters.

3.1. Statistical based approach

Statistical approaches label objects as anomalies if it is different from the expected model.

The term ”nonparametric statistics” describes a statistical methodology where it is not assumed that the data came from models and where the data are not indicated as coming from models that have been defined and are depending on a small number of factors; Parametric tests are those that make assumptions about the characteristics of the sampled data.

Gaussian Mixture and Regression Method

This model assumes each data point belongs to Gaussian distribution. Here identifying anomaly is the process of detecting unusual data points and it is identifying low density regions. To achieve high accuracy regression model is used in anomaly detection.

Histogram

Histogram is a best method to check univariate data that contains single variable. A histogram divides the range of values into various clusters and shows the frequency.

3.2. Spectral analysis

Spectral analysis is a clustering method that divides data into subspaces for normal, abnormality and noise based on the properties the information.

3.3. Anomaly Detection Approaches Based Machine Learning

Machine learning approaches can be used to automate and enhance anomaly detection, particularly when large amounts of data are involved. The techniques for detecting anomaly are categorized based on machine learning algorithms used.

1. Clustering based approaches
2. Classification based approaches
3. Nearest neighbour-based approaches

Neural-Network

Anomaly detection methods based on neural networks (NNs) first train the NN on a sample set of normal data before feeding the NN with the observed data. Outlier status is assigned to the data point that the NN rejected. Because they may optimize the training NN with only a few parameters and do not require any prior assumptions about the properties of the dataset, NN-based techniques have been widely used in the fields of classification detection and anomaly detection. Recurrent NNs, replicator NNs (RNNs), BP NNs, and other types of NNs have all been used.

Bayesian networks

Bayesian networks are excellent at analysing an event that already happened and determining the chances that any one of various potential known causes was a contributing element. This could relate a probabilistic relationship between diseases and symptoms. The majority of anomaly detection techniques rely on prior knowledge, such as information or data gathered from neighbour nodes, which is inefficient for resource- constrained sensor nodes because it necessitates additional massage exchange and uses a lot of energy.

Support Vector Machine

SVM is applied to both regression and classification problems. Hence the issues of classification is frequently used. This classification algorithm (SVM), In n-dimensional space, each data item is described as a point. Identifying the hyper-plane that effectively separates the two classes will be the next classification step. The SVM model is the boundary that works effectively in dividing the two classes.

Rule-based Classifier

Rule-based systems need an individual to make any changes, whereas machine learning systems

can learn from the past and adapt to new situations on their own. Either rules can be arranged, in which case the final class is the one that corresponds to the rule activated with the highest priority. If not, we can distribute votes for each class in accordance with some of their weights, meaning that the rules remain unordered.

Clustering based methods

Grouping of unlabelled examples are called clustering. An approach of grouping data points into different clusters composed of relevant data points. The potentially similar items are still in a cluster that has little to no resemblance to the next group, in

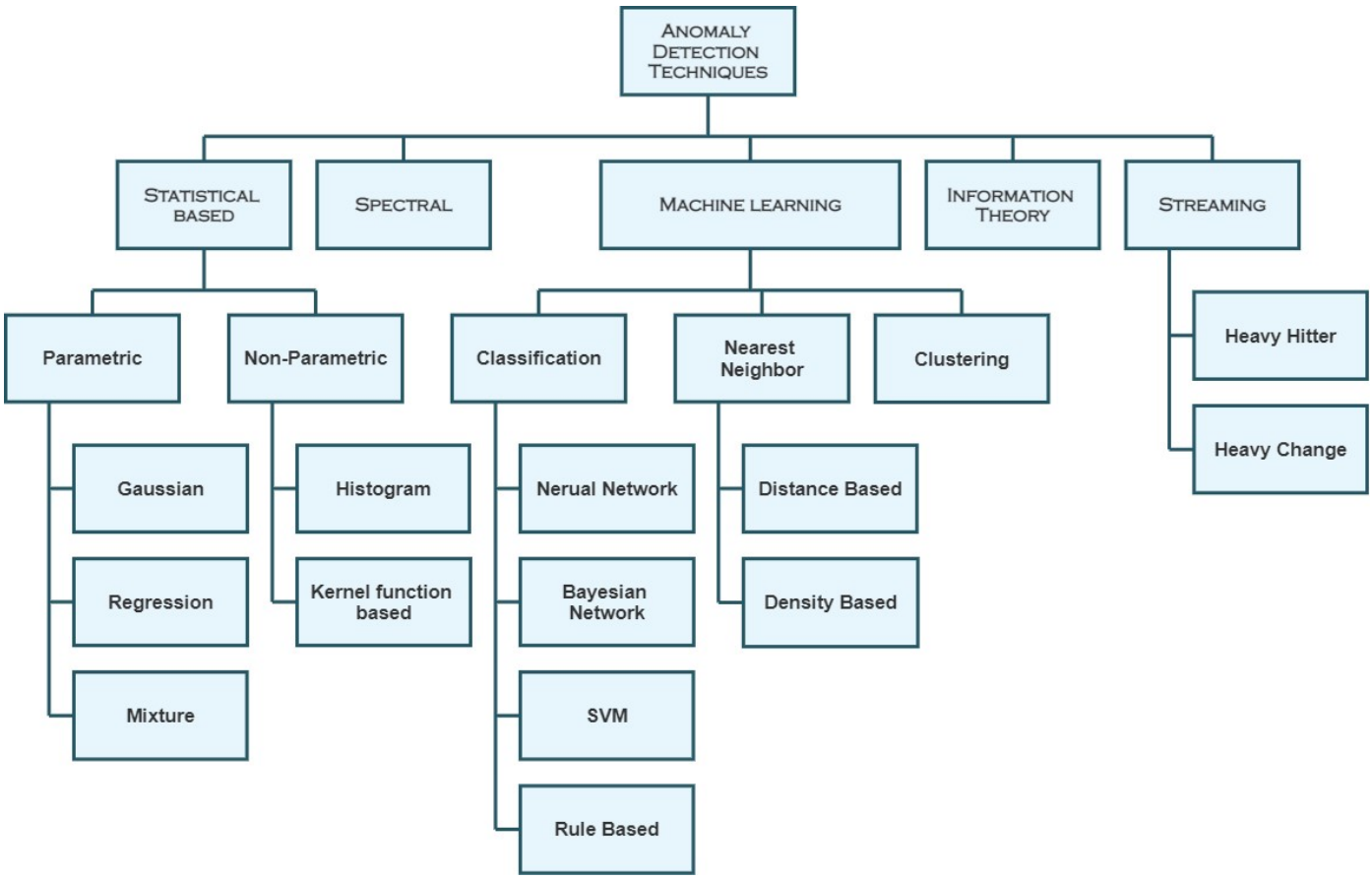

FIGURE 2. **Anomaly Detection Techniques**

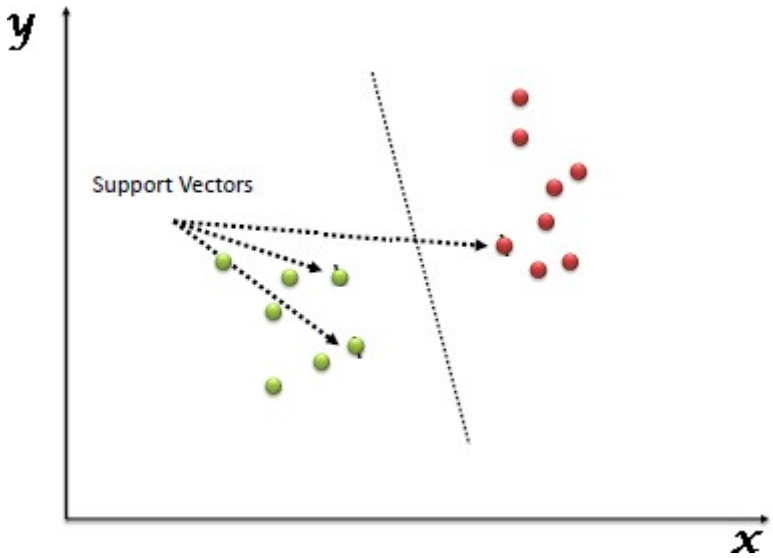

FIGURE 3. **SVM**

order to do so, it looks for similar patterns inside the unlabelled data set, like movement, structure, texture, pigment, and so on, thereafter categorize the data based on whether or not these patterns are present.

Distance based methods in Machine Learning

Machine learning algorithms called distance-based algorithms classify queries by calculating the distances between the queries and various internally stored exemplars. The query's classification is most strongly influenced by examples that are nearest to it. The following list contains four potential distance metrics that are frequently employed in ML. Euclidean distance, Manhattan distance, Minkowski

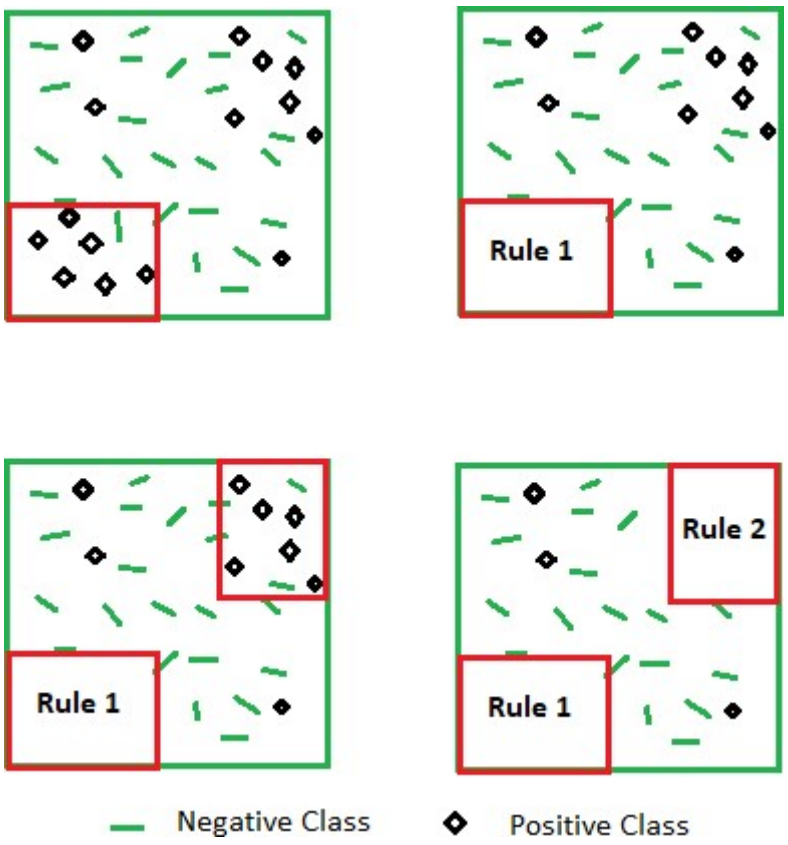

FIGURE 4. Rule based classifier

distance, Hamming distance

Euclidean distance

The Euclidean distance is the distance between two points that is the shortest. Consider A&B are the two points.

Therefore Distance= ((a1- b1)2 + (a2 – b2)2)1/2

Manhattan distance

It represents the total amount of absolute differences between points along all dimensions.

The sum of absolute distance between two points are distance= —$a_1 - b_1$ — + —$a_2 - b_2$ —

Minkowski distance

This method is a generalized form of Euclidean distance and Manhattan distance and it includes a variable called the order or S, it enables the calculation of an alternative distance unit. Euclidean distance= (sum for I to n (abs(v1[i]-v2[i]))^s)$^{1/s}$

Here S is the parameter and S are set to 1, The formula is identical to that for the Manhattan distance. When S is set to 2, The formula is identical to the Euclidean distance.ie. S1= Manhattan distance and S2 =Euclidean distance

Hamming distance

It determines the separation between two bitstrings, often known as binary strings or binary vectors. Assuming the hamming distance d(10101, 11110) is 3 then the minimum distance between all possible pairs in a set of words -d_{min} = 2 d (000, 011) =2 d (000,101) =2 d (000, 110) =2 d (011, 101) =2 d (011, 110) =2

d (101, 110) =2

Density based methods in machine learning

In data space, a cluster or subgroup is considered to be a contiguous zone of high point density,

separated from neighboring clusters by sparse areas. To identify the zones where components are grouped together and the places in which those points are dispersed by empty or sparse regions. The points which are noisy don't correspond to a cluster. Finding convex or spherical clusters can be done using partitioning techniques and hierarchical clustering. This is only suitable for well divided and compact clusters. As a result, it is likewise impacted by the data's noise and anomalies.

3.4. Information Theory

Information theory promotes machine learning, develops on statistics, and is concerned with data reduction and distribution. Information gives us a method to put a number on the novelty factor of an

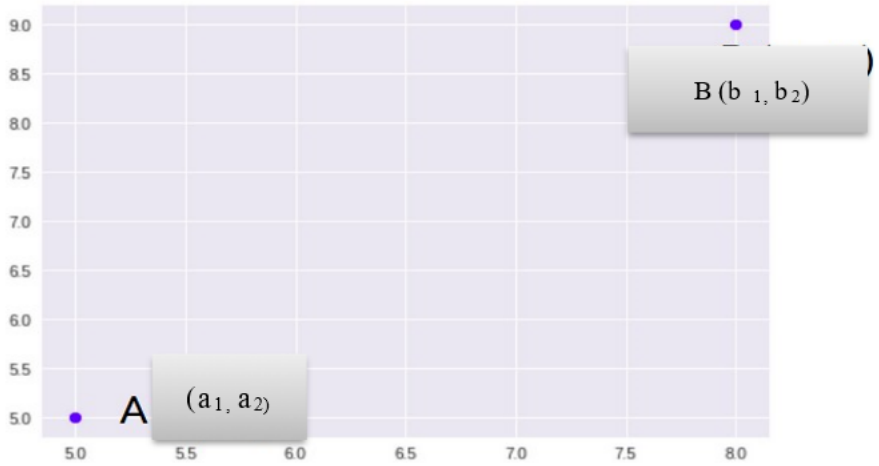

FIGURE 5. Two points A & B

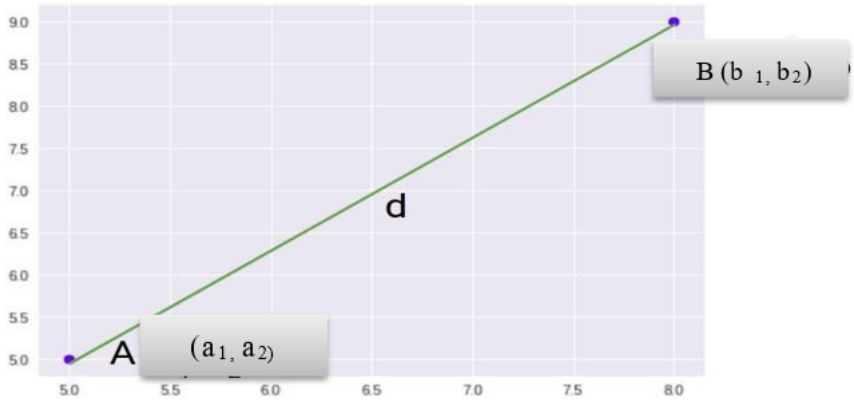

FIGURE 6. Identifying the distance between A & B

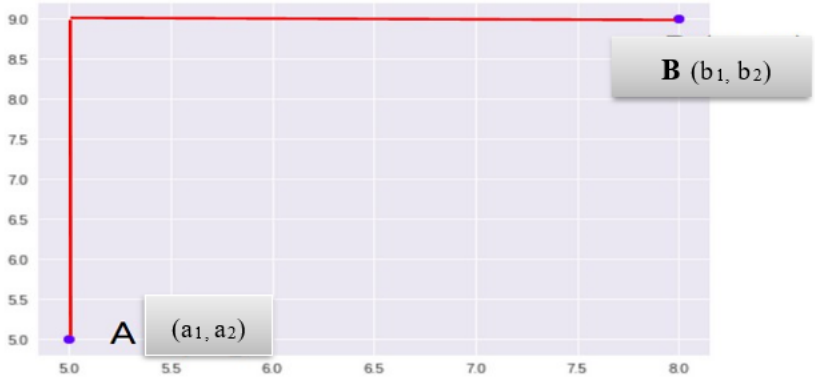

FIGURE 7. Distance between the two points

occurrence that is calculated in bits.

3.5. Streaming

A massive data flow is a process that absorbs and modifies data in real-time segments between an origin and an endpoint. Streaming ML is the applica-

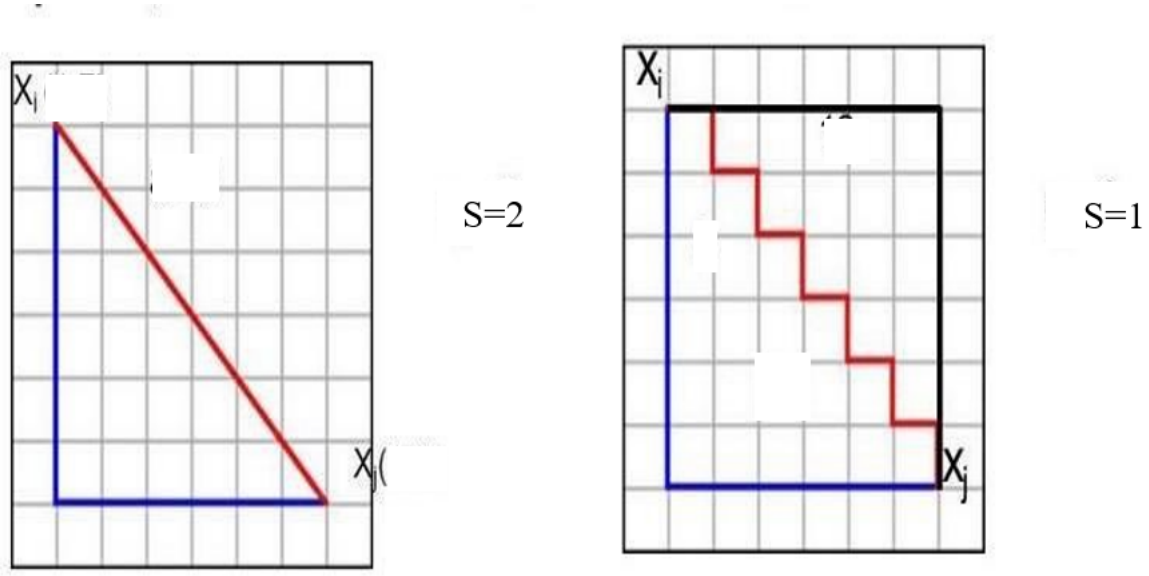

FIGURE 8. Euclidean distance and Manhattan distance

tion of an ML model to this pipeline.

4. Conclusion

Anomaly detection is a big issue in IoT. Numerous anomaly detection methods for IoT have been presented during the past few of years. Traditional data analysis techniques are inadequate due to the exponential growth of IoT data, however machine learning algorithms excel at processing and analyzing large amounts of data. Therefore, in this paper view the state of the technology in anomaly detection approaches based on Machine learning is given. Each method has its distinctive gains and drawbacks. It is difficult to predict which method is best for all IoT applications. Numerous machine learning algorithms are available, and they can be chosen depending on the individual application circumstances.

References

Alcalde-Barros, Alejandro, et al. "DPASF: a flink library for streaming data preprocessing". *Big Data Analytics* 4.1 (2019). 10.1186/s41044-019-0041-8,%202019.

Carrera, Diego, et al. "Online anomaly detection for long-term ECG monitoring using wearable devices". *Pattern Recognition* 88 (2019): 482–492. 10.1016/j.patcog.2018.11.019.

Larriva-Novo, Xavier, et al. "An IoT-Focused Intrusion Detection System Approach Based on Preprocessing Characterization for Cybersecurity Datasets". *Sensors* 21.2 (2021): 656–656. 10.3390/s21020656..

Li, Jinbo, et al. "Clustering-based anomaly detection in multivariate time series data". *Applied Soft Computing* 100 (2021): 106919–106919. 10.1016/j.asoc.2020.106919.

Nematzadeh, Zahra, Roliana Ibrahim, and Ali Selamat. "A hybrid model for class noise detection using k-means and classification filtering algorithms". *SN Applied Sciences* 2.7 (2020). 10.1007/s42452-020-3129-x..

Pavithra, A, C Anandhakumar, and V Nithin Meenashisundharam. "Internet of Things with BIG DATA Analytics - A Survey". *International Journal of Scientific Research in Computer Science Applications and Management Studies IJSRCSAMS* 8.1 (2019).

Sonawane, Sandip. "Survey on Technologies, uses and Challenges of IoT". *International Journal of Engineering Research & Technology (IJERT)* 8 (2019): 2278–0181. 10.17577/ijertv8is120162.

Sun, Hongyu, et al. "Fast Anomaly Detection in Multiple Multi-Dimensional Data Streams". *2019 IEEE International Conference on Big Data (Big Data)* (2019): 1218–1223. 10.1109/bigdata47090.2019.9006354.

Yu, Xiang, et al. "An adaptive method based on contextual anomaly detection in Internet of Things through wireless sensor networks". *International Journal of Distributed Sensor Networks* 16.5 (2020): 155014772092047–155014772092047. 10.1177/1550147720920478.

Embargo period: The article has no embargo period.

To cite this Article: Subha, S, and J G R Sathiaseelan. "The Enhanced Anomaly Deduction Techniques for Detecting Redundant Data in IoT." International Research Journal on Advanced Science Hub 05.02 February (2023): 47–54. http://dx.doi.org/10.47392/irjash.2023.012