JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# A Review of Different Machine Learning Techniques

[1] K S L Prasad, [2] Korukonda Balaji

[1]Assistant Professor, Department of MCA, Loyola Academy Degree & PG College, Hyderabad, Telangana, India
[2]Assistant Professor, Department of B.Sc.(CS&CS), Loyola Academy Degree & PG College, Hyderabad, Telangana, India

***Abstract :*** Anything that deviates from what is anticipated, customary, or normal is referred to as an abnormality. In many applications, it is essential to identify anomalies. The most common argument is that unexpected behaviour always results in a loss of some kind, which could be damage to the system or the loss of important data. Many techniques for detecting anomalies have been developed specifically for certain types of data or applications. To help researchers select an effective anomaly detection methodology, we have listed a few machine-learning anomaly detection techniques in this paper.

*Index Terms* - Anomaly, Anomaly Detection, Intrusion Detection, Outlier, Supervised, Unsupervised

## 1. INTRODUCTION

Finding data pieces, points, or uncommon occurrences that exhibit unexpected behaviour or that differ noticeably from other similar data items, points, or events is known as anomaly detection. We refer to these uncommon objects as anomalies, outliers, exceptions, flaws, or contamination. The two most frequently used terms are anomaly and outliers. Widespread applications of anomaly detection include industrial damage detection, textural surface flaws, sensor networks, intrusion detection systems, fraud detection in credit card transactions, and system (hardware and software) health monitoring systems. Anything that deviates from the norm and what is expected is called an anomaly. Thus, data items or points that differ from typical data or have the potential to produce unexpected behaviour are referred to as anomalies in machine learning. The anomalies are divided into three categories: contextual, collective, and point abnormalities. The significance of anomaly detection stems from the possibility that either the abnormalities themselves or the unexpected behaviour in the system is what gives rise to the anomalies. This lack of clarity in behaviour is dangerous. This unexpected behaviour may result in system damage, vital data theft, or the halting of regular services. Therefore, it's critical to spot these anomalies and take appropriate action in response to them.

## 2. BASIC METHODOLOGY OF ANOMALY DETECTION

There are many ways of anomaly detection in machine learning - it is a four-step process as given in Figure 1.

### 2.1 Feature extraction stage and data pre-processing

In this stage the required features are selected for detection and data is stored with the extracted features. The data is also pre-processed to handle missing or unknown data, normalization, scaling, and other processes.

### 2.2 Training

In this stage, the model would be selected and the model would be trained to learn normal (and/or abnormal) data/behaviour of the system.

### 2.3 Detection

Once the model is trained it is deployed to detect any anomaly in the real-world input. If the deviation is found to cross a predefined threshold it will be considered as an anomaly and an alert will be raised.

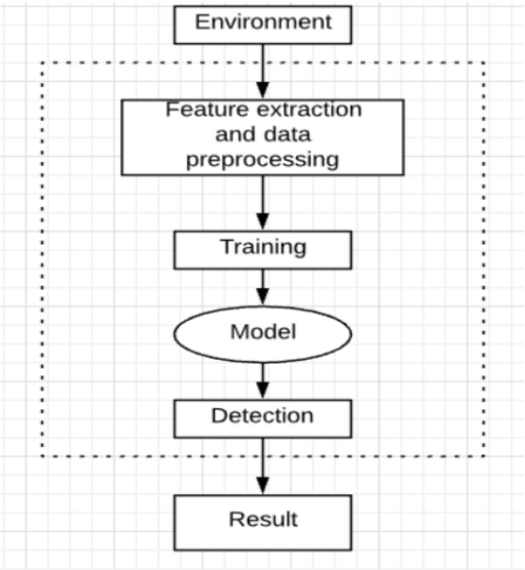

Fig 1. Basic Methodology of Anomaly Detection

## 1. APPLICATION OF ANOMALY DETECTION

Important applications of anomaly detection include:

### 3.1 Intrusion Detection System

Anomalies in computer networks or hosts that are based on computers. By identifying anomalies in network traffic or computer behaviour that may indicate a call to the operating system and result in anomalous computer behaviour, anomaly detection plays a crucial role in intrusion detection systems.

### 3.2 Industrial damage detection

Wear and tear and constant operation of the machines cause damage to industrial equipment and machinery. Therefore, early discovery of these impairments can aid in averting significant losses. Therefore, anomaly detection is crucial to the identification of damages. System health monitoring and operation can be separated into categories of industrial harm. A malfunction in one of the machine's components arises when there is system health monitoring. If by operation, the machine's structural flaw is to blame.

### 3.3 Texture surface defect detection

The texture surface consists of continuously repeated patterns. The defect can be in the texture or the pattern. Anomaly detection helps in the detection of such defects.

### 3.4 Credit card fraud detection

The primary uses of anomaly detection are in the identification of fraudulent credit card applications and credit card activity. High cost, high rate, and purchases of goods that the user has never before purchased are indicators of fraud. etc. Finding the fraud as soon as the transaction is complete is difficult.

## 2. DIFFERENT MACHINE LEARNING TECHNIQUES FOR ANOMALY DETECTION

Many techniques in Machine learning can be used for anomaly detection which can be mainly classified into supervised and unsupervised learning techniques.

### 4.1 Supervised machine learning

One of the primary categories of machine learning in which input and output are accessible is this one. The primary objective of this learning process is to identify a function capable of establishing a relationship between input and output. Put differently, identify a mapping between the input and the output. The most popular machine learning approaches and algorithms for anomaly detection are K closest neighbours, decision trees, support vector machines, and Bayesian networks.

### 4.1.1 Support Vector Machine (SVM):

One well-liked supervised machine learning algorithm is SVM. Although regression problems can also be resolved, classification difficulties are the ones that are typically resolved using it. SVM's primary goal is to identify a hyperplane that can discriminate between different data classes. For example, data from one class will be on one side of the hyperplane, and data from other classes will be on the other. Therefore, SVM must identify several hyperplanes to do multiclass classification. By classifying the training data as normal, abnormal, or anomalous, SVM can be used to find anomalies. By identifying classes within the abnormal data, more precise classes can be obtained.

SVM was employed by the authors of [2] to detect runtime anomalies with Hardware Performance Counters (HPC). They established an SVM model for application anomaly detection. For training, they used the suite of CHStone benchmark apps. SVM was trained using characteristics from HPC to identify anomalies. They defined the hyperplane using a radial basis function, or RBF. The model was able to distinguish between two comparable apps with success. They had a high degree of accuracy in identifying anomalies, and none of them were mistaken for typical data.

SVM was employed by the authors of [3] to detect network anomaly traffic. They distinguished between attack and legitimate traffic. After analysing four distinct attack patterns, the model outperformed incremental LOF and HPStream in terms of accuracy.

**4.1.2 K-nearest neighbor (KNN):** Similar to SVM, KNN is mostly used to solve classification problems, while it can also be applied to regression. It is predicated on the idea that similar objects are closer to one another. This algorithm, also known as lazy learning, stores data instead of using it for training. It computes and sorts a similarity metric between a new data item and the existing data, selecting the first k most similar instances or data and using a majority rule to determine the new data item's class or label. Therefore, choosing the right K value to utilise is a crucial task.

To detect anomalies in ELV DC Pico grids—the state of various appliances within a circuit—the authors of [4] employed KNN. They only took into consideration data from a window block rather than the entire set of data and used KNN with Euclidean distance as similarity metrics. They employed a gadget to transform signals into KNN-useful features. They tested their model in three separate circuits and were able to obtain good true positive and false positive accuracy.

**4.1.3 Bayesian Networks:** As a probabilistic graph model, Bayesian networks are a well-liked supervised machine learning approach. The two learning components of this approach are the parameter and the structure. The structure is a directed acyclic graph (DAG), in which interactions are connected with the edges and random variables are associated with each node. The conditional dependencies and independent variables among the random variables are expressed by the DAG. Finding the conditional probability for each random variable in the network is the goal of parameter learning. Therefore, to construct and apply Bayesian networks, one needs to be aware of the problem's random variables, their relationships, and the conditional probability associated with each variable. Finding conditional probability given a specific class would be the job for this algorithm, which can be applied to anomaly detection in situations when a class attribute is known.

The authors of [5] used Bayesian networks to classify anomalies in network traffic along with Markov Chain Monte Carlo and Maximum Likelihood Estimation for structure and parameter learning respectively. They used UNB ISCX IDS 2012 and UAN W32.Worms 2008 datasets with 8 different features for training and testing. They achieved a high accuracy for the classification of normal traffic and different types of attacks which was more efficient than other methods that used classical datasets

**4.1.4 Decision Tree:** A notable supervised machine learning tool that is more frequently used for classification than regression is the decision tree. Additionally, the learning algorithm is non-parametric. The popular structure of a tree, in which each internal node represents a feature and each leaf node indicates a class, serves as the foundation for this technique. The rule for classifying that class is hence the path from root to leaf node. This algorithm's primary goal is to build a decision tree based on the data gathered at each node, which determines how much information a feature can provide about a class and how judgements can be made. By choosing the right criteria for dividing data into normal and abnormal categories, this technique can be utilised to detect anomalies.

The authors of [10] used decision trees to detect anomalies in real-time servers - Short Message Service Center (SMSC). They wrote a script to collect about 500 attributes of data from the server ran the script for 30 days and collected 8600 records. They divided the recorded data into normal and abnormal classes. They achieved a very high accuracy in the classification of anomalies and normal data.

**4.2 Unsupervised machine learning**

It's a machine learning method where training is limited to input data. This kind of machine learning algorithm's primary goal is to identify relationships and similarities between the data as well as common recurrent patterns in the data. The features of the data are also summarised and explained with the aid of this kind of machine learning technique. A few unsupervised machine learning techniques that are used to discover anomalies are Adaptive Resonance Theory, KMeans, C-means, and Self-Organizing Map.

**4.2.1 Self-Organising Map (SOM):**

SOM is an artificial neural network-based unsupervised machine learning approach that lowers the dimensionality of the discrete input space. Therefore, dimensionality reduction can be achieved with this strategy. SOM's primary goal is to modify the network so that certain segments react to particular kinds of inputs comparably. Weights are randomly chosen when neurons are first established. By calculating the distance between the input and the weight of the neurons, this technique applies competitive learning. The neurons whose weights are closest to the input weights are designated as the best matching unit (BMU). In the SOM network, the BMU neurons that are closest to the input are converted to input weights. The network receives the input for an extended period. As a result, the network adapts by assigning output nodes to associate input with data groupings or patterns. By providing inputs to uncover patterns of how typical data are, SOM may be utilised for anomaly detection. When the network responds differently, we can infer an anomaly.

Authors of [11] used SOM for anomaly detection on warp-knit fabric surfaces was used at 2 levels. On 1st level, it was used for detection of the type of fabric, and at the second level, it was used to detect defects in the texture. Datasets are not readily available of texture images hence were taken directly from industries and are unpredictable. Since the defects of texture surfaces are not predictable, unsupervised learning is favorable. 8 different types of defects needed to be detected and SOM was able to achieve a decent accuracy of 80%.

**4.2.2. K-means:** One well-liked unsupervised machine learning algorithm is K-means. Finding common patterns in the data and grouping comparable data into clusters are the key goals of the analysis. K-means, then, searches for a predetermined number of clusters that K defines. The user determines what K is. When applying fresh data, machine learning can distinguish between normal and abnormal data clusters. Based on this similarity measure, the machine can select one of the clusters.

Authors of [12] used a K-means clustering algorithm to detect anomalies in network traffic. They made modifications to the amount initially determined of clusters. They used the concept of a sliding window, which limits the amount of data transmitted or processed. They used the KDDCup99 dataset for training. They achieved a high accuracy in the classification of normal and abnormal traffic

**4.2.3 Fuzzy C-means (FCM):** Fuzzy clustering is different from hard clustering in the sense that in fuzzy clustering, we can have data points belonging to more than one cluster at a given instance, whereas in hard clustering, data points are assigned only to one cluster. Data points belonging to different clusters are as different as possible and those belonging to the same cluster are as alike as possible.

For each data point lying in any cluster initially, the fuzzy membership value is obtained using fuzzy logic. In FCM, we are first required to find out the centroid of the various data points/items. For each of these items, the distance is calculated from the centroid, each cluster has its centroid. On doing cluster analysis if we find that the data point is "closer" to another cluster, then it is considered a part of that cluster thereafter. This process continues until the clusters so formed reach a constant state.

Authors of [13] presented a way of using fuzzy c means clustering in an unsupervised manner for anomaly detection for intrusion detection and the results were compared with other known methods such as Expectation Maximisation, K-medoids, and it was recorded that FCM obtained fairly better results.

**4.2.4 Adaptive Resonance Theory (ART):** The hypothesis of adaptive resonance pertains to certain features of the brain's information-processing mechanism. It covers some neural network models that tackle issues like pattern recognition and prediction using supervised and unsupervised learning techniques. There are two steps in this process. The learning step comes first. This stage involves categorising a data collection that contains many samples that are recorded under typical circumstances based on how similar they are. A sample is a set of process variables that are obtained at a particular instance. The terms "normal categories" relate to the categories that were learned as a result of the learning stage. The step for detecting outliers is the second. Every sample is obtained, and its classification is established by first assessing the normalcy with the aid of ART. The sample does not exhibit any anomalous or outlier behaviour if the obtained normalcy falls into one of the normal categories. If not, a new category that will be referred to as an anomaly for upcoming samples must be created.

The authors of [14] compared the performances of four anomaly detection systems based on the ART model. They were Single ART model, Multiple ART models, Distributed ART models, and Multiple and distributed ART models. A criterion was developed that further complemented the work done on distributed model systems by evaluating the occurrence of the anomaly as well as the quantity, in other words, the anomaly's level.

## 3. CONCLUSION

This paper lists and briefly describes some machine learning algorithms that can be applied to anomaly detection. This paper aims to educate academics and other stakeholders on the fundamentals of various machine learning algorithms for anomaly identification. Every application experiences different types of abnormalities daily, necessitating the necessity for a strategy to adjust. Therefore, for handling novel types of anomalies, it is preferable to use unsupervised machine learning approaches. On the other hand, supervised learning is a better option and produces better accuracy outcomes in scenarios when it is guaranteed that only a certain number of anomalies will occur.

## REFERENCES

[1]. Shikha Agrawal, Jitendra Agrawal, "Survey on Anomaly Detection using Data Mining Techniques", Published in Procedia Computer Science, Volume 60, 2015, pp 708-713

[2]. Muhamed Fauzi Bin Abbas, Sai Praveen Kadiyala, Alok Prakash, Thambipillai Srikanthan, Yan Lin Aung, "Hardware performance counters-based runtime anomaly detection using SVM", Published in 2017 TRON Symposium (TRONSHOW), DOI: 10.23919/TRONSHOW.2017.8275073

[3]. Gao Yan "Network Anomaly Traffic Detection Method Based on Support Vector Machine", Published in 2016 International Conference on Smart City and Systems Engineering (ICSCSE), DOI : 10.1109/ICSCSE.2016.0011

[4]. Yang Thee Quek, Wai Lok Woo, Logenthiran Thillainathan, "IoT Load Classification and Anomaly Warning in ELV DC Picogrids Using Hierarchical Extended k-Nearest Neighbors " , Published in IEEE Internet of Things Journal ( Volume 7, Issue 2, Feb 2020), pp.863-873

[5]. M. J. Vargas-Muñoz, R. Martínez-Peláez, P. Velarde-Alvarado, E. Moreno-García, D. L. Torres-Roman, J. J. Ceballos-Mejía, "Classification of network anomalies in flow level network traffic using Bayesian networks", Published in 2018 International Conference on Electronics, Communications, and Computers (CONIELECOMP), DOI:10.1109/CONIELECOMP.2018.8327205

[6]. M. Panda, A. Abraham y M. R. Patra, "Hybrid intelligent systems for detecting network intrusions," Security Communication Networks, vol. 8, p. 2741–2749, 2015

[7]. B. Balasingam, P. Mannaru, D. Sidoti, K. Pattipati, P. Willett, "Online Anomaly Detection in Big Data: The First Line of Defense Against Intruders", Studies in Big Data, vol. 24, pp 83-107, 2017

[8]. V. Hodge, J. Austin, "A survey of outlier detection methodologies", Artif. Intell. Rev., 22(2), 2004

[9]. H. Jantan, "Human talent prediction in HRM using C4.5 classification algorithm", IJCSE, vol. 02, no. 08, p. 2529, 2010

[10]. Georges Chaaya, Hoda Maalouf, "Anomaly detection on a real-time server using decision trees step by step procedure", Conference: 2017 8th International Conference on Information Technology (ICIT), DOI:10.1109/ICITECH.2017.8079989

[11]. Dimuthu Wijesingha, Buddhika Jayasekara, "Detection of defects on Warp-knit Fabric surfaces Using Self Organising Map", Published in: 2018 Moratuwa Engineering Research Conference (MERCon), DOI : 10.1109/MERCon.2018.8421944

[12]. I Wayan Oka Krismawan Putra, Yudha Purwanto, Fiky Yosef Suratman, "Modified k-means algorithm using timestamp initialization sliding window to detect anomaly traffic", Published in: 2015 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), DOI : 10.1109/ICCEREC.2015.7337042

[13]. B S Harish, S V Aruna kumar, "Anomaly based intrusion detection using modified Fuzzy clustering", Published in: International Journal of Interactive Multimedia and Artificial Intelligence · January 2017,pp 54-59.