# A Survey on Anomaly Detection using Unsupervised Learning Techniques

[1]Barani Priyanga R,[2]Dr.K.Anitha Kumari, [3]Dharani D
[1]PG Scholar, [2]Assistant Professor(Sr.Gr), [3]Assistant Professor
[1,2,3] Department of Information Technology,
[1,2,3]PSG College of Technology, Coimbatore, India

***Abstract-***Anomaly detection method is used to find the abnormal behaviour of the data that do not conform to the normal behaviour called outliers. Many anomaly detection techniques have been discussed related to unsupervised learning. In this survey, benefits and drawbacks of various unsupervised learning techniques are discussed and some suggestions are made on choosing the suitable techniques to detect anomalies. Unsupervised learning is a type of learning algorithm used to draw inferences from datasets consisting of input data without labelled responses. These learning methods are used to find the hidden patterns or grouping in data. Various techniques has grouped from different categories to differentiate between normal and anomalous behaviour of the data used by various application to detect the fraudulent activities. The anomaly detection also provides better threat intelligence and optimize the accuracy of alerts.

***Index terms-*** **Anomaly detection, unsupervised Learning.**

## 1. INTRODUCTION

In recent days industry has evolved into a highly-supervised industry, where operational security and safety are present and act as foundational values. Each and every information has to be monitored to avoid undesired events and to prevent possible attacks and abnormal behaviour. As external threats are increasing most of the system suffers from anomalies which leads to false information. Constant monitoring has to be done to avoid data loss. Anomaly detection is very important, where the nature of the data can be observed constantly. Anomaly detection provides better threat intelligence and optimize the accuracy of alerts. Safety and security of data is augmented. Anomaly detection has to be done with greater scale and speed so that the safety and security of the data can be augmented. Machine Learning has four normal classes of utilizations: grouping, foreseeing next esteem, peculiarity recognition, and finding structure. Among them, Anomaly discovery identifies information focuses in information that does not fit well with whatever remains of the information. It has an extensive variety of uses, for example, misrepresentation location, observation, finding, information cleanup, and prescient support. In most situations, the data is created by one or more generating processes, which are not only representing activities in the system but also observations on entity collections. When the generating process behaves unusually, it creates anomalies or outliers. Thus, an anomaly often contains valuable information about abnormal characteristics of the systems and elements that impacts the generation process.

### 1.1 Aspects of anomaly detection

#### 1.1.1 Nature of Input Data

A primary part of any anomaly detection is the nature of the input data. The input data can be seen as a set of attributes. The attributes can be of different kinds such as categorical, binary or continuous. Each data might have merely one attribute or multiple attributes. Furthermore, the attributes of each data instances may be the same or different types. The nature of attributes determines the applicability of anomaly detection. For example, most statistical models have to be used for continuous and categorical data and for nearness neighbor based models, the nature of attributes would determine the distance measurements.

#### 1.1.2 Output of Anomaly detection

Among all these applications, the data has a "normal" model, and anomalies are recognized as deviations from this normal model. The output of anomalies can be spliced into two types, 1. Anomaly Scores- many anomaly detection algorithms output a score qualifying the level of "outliers" of each data point. This kind of output can contain variety of parameters related to the data point. 2. Binary labels- binary label indicates whether a data point is an anomaly or not. Despite the fact that some anomaly detection algorithms

return binary labels directly, outlier scores can be converted into binary labels. A binary label contains less information than a scoring system. However, it is the final result that is usually needed for decision making.

### 1.2 Unsupervised learning

Unsupervised learning is a kind of machine learning calculation used to draw inferences from datasets comprising of information without marked reactions. The most widely recognized unsupervised learning strategy is cluster investigation, which is utilized for exploratory information examination to discover concealed examples or grouping information. The clusters are displayed utilizing a measure of similitude which is characterized upon measurements, for example, Euclidean or probabilistic separation.

### 1.3 Types of Anomalies

Anomalies or outliers are broadly categorized as follows:

- Point anomalies- A solitary example of information is odd if it's too far-removed from the rest. Nearly all unsupervised anomaly detection algorithm are based on this type.
- Contextual Anomalies- An information is irregular in a particular setting, yet not generally oddity if happen at certain time or certain area.
- Collective anomalies- A collection of related information is anomalous with respect to the entire data.

### 1.4 Normalization

At the point when a dataset was preprocessed with the end goal that it speaks to a point inconsistency discovery issue, the last advance before the unsupervised anomaly detection algorithm is normalization. With the distinctive kinds of information, standardization should be performed by considering foundation learning into account. Min-max normalization is a typical normalization method, where every feature is normalized into a common interval [0,1] and standardizing, where each feature is transformed such that its mean is zero and its standard deviation is one In handy application min-max normalization is regularly utilized, now and then straight-forward normalization can likewise be contra-profitable. Let's assume we have a categorical binary feature converted to [0, 1] and a numerical value measuring a length normalized to [0, 1]. Since the categorical binary feature brings about separations being it is possible that one or zero, it impacts the outcome as the numerical value. This is the reason, why foundation data is likewise essential during standardization to maintain a strategic distance from these errors in the standardization procedure.

## 2. RELATED WORK

From the base of our paper, we mainly present unsupervised learning techniques for anomaly detection. Various categories of anomaly detection techniques are discussed. As we all know there is no ideal algorithm, each one have their own benefits and drawback. In [14] the authors summarize a structured and comprehensive overview of the research on anomaly detection. In [1] the unsupervised learning techniques are discussed and experimented with the results. In [2] Principal Component Analysis (PCA) and Fuzzy Adaptive Resonance Theory (Fuzzy ART) are applied to reduce the high dimensional data vectors and distance between a vector and its projection. In the LOF algorithm [3][4], it solves the problem of mining the local outliers. The COF algorithm [6] improves the effectiveness of an existing local outlier factor (LOF) scheme when a pattern itself has similar neighborhood density as an outlier. In [7] anomaly is detected using k means clustering technique, and experimented with different k values. K-Means [8] clustering method is first applied to the normal training instances to partition it into 'k' clusters using Euclidean distance similarity. Using this technique anomaly scores are extracted, these anomaly scores are used find the final anomaly scores. This paper [9] illustrates how to use apriori algorithm in intrusion detection systems to create an automatic firewall rules generator to detect novel anomaly attack.

## 3. UNSUPERVISED ANOMALY DETECTION ALGORITHMS

Unsupervised anomaly detection algorithms are roughly categorized into the following main groups as illustrated in fig.1. (1) Nearest-neighbor based techniques, (2) Clustering-based methods and (3) Statistical algorithms. Recently, also a new group is emerging based on (4) Subspace techniques. In this work nearest-neighbor based techniques and cluster based methods are discussed, which are the frequently used techniques for anomaly detection. In this paper various unsupervised learning algorithms are discussed to detect anomalies in the unlabeled data.

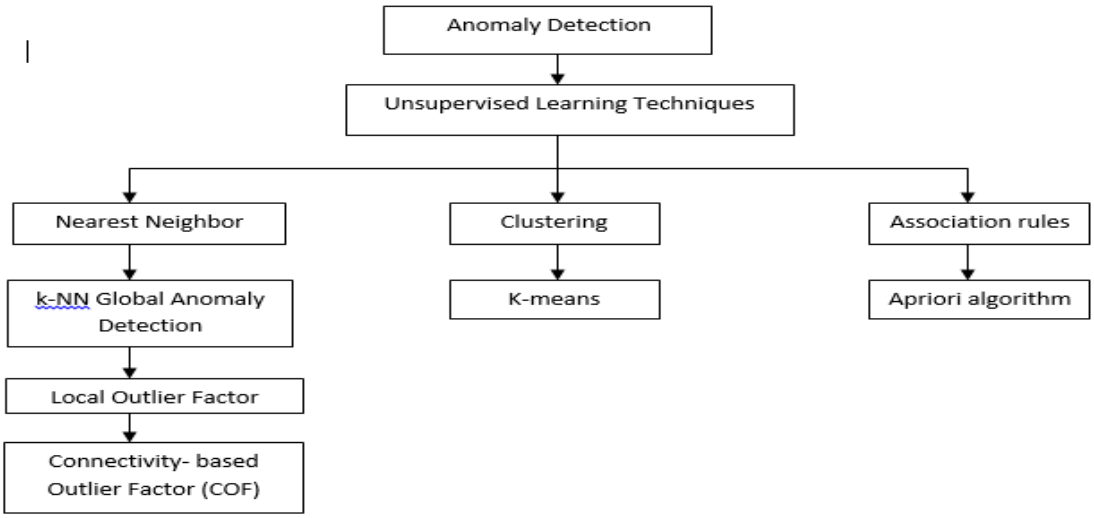


**Fig 1. System Design of Unsupervised Anomaly Detection**

### 3.1 k-NN Global Anomaly Detection

The k-nearest neighbors (k-NN) global anomaly detection algorithm is a is a straight forward method for detecting anomalies, not to be mistaken with k-nearest neighbor classification. As the name already says it concentrates on the global anomalies and not able to detect the local anomalies. k-nearest neighbor has to be find for each record in the dataset. Anomaly score is calculated using the distance of the $k^{th}$- nearest neighbor or the average distance to all the k- nearest neighbors. The first method is referred as $k^{th}$-NN and the latter method as k-NN. K-NN is the preferable method in most of the applications. The absolute value of the score is depend on the dataset itself, the no of dimensions and normalization. The decision of the parameter k is obviously vital for the outcomes. If the *K* value picked is too low, the thickness estimation for the records may be not reliable. Then again, on the off chance that it is too large, thickness estimation might be excessively coarse. As a dependable guideline, *k* should be in the range $10 < k < 50$. In classification *k* is possible to determine using cross validation, where in unsupervised anomaly detection there is no such technique to determine due to missing labels. For that reason, many different values for *k* and average in order to get a fair evaluation when comparing algorithms.

### 3.2 Local Outlier Factor (LOF)

The local outlier factor is the most well-known local anomaly detection algorithm and pioneered the idea of local anomalies. The idea of LOF is carried out by the nearest neighbor based algorithm. The LOF is calculated using the following steps.

1. The *k*-nearest-neighbors must be found for each record x. If there should arise an occurrence of separation tie of the kth neighbor, more than k neighbors are utilized.

2. Utilizing this k-nearest neighbors $N_k$, the nearby thickness for a record is evaluated by calculating the Local Reachability Density (LRD)

3. At long last, the LOF score is computed by comparing the LRD of a record and the LRDs of its k neighbors.

The LOF score is therefore fundamentally a ratio of local densities. This results in the nice property of LOF, that common place events, which densities are as vast as the densities of their neighbors, get a score of around 1.0. Inconsistencies, which have a low nearby thickness, will bring about greater scores. It is also cleared that why this algorithm is local. Obviously, global anomalies also can be recognized since they additionally have a low LRD when contrasting and their neighbors. It is critical to take note of that in abnormality discovery assignments, where local outliers are not of intrigue, this calculation will produce a great deal of false alarms as we discovered amid our assessment. The setting of k is indispensable for this algorithm. Other than the author of this algorithm propose an ensemble procedure for figuring LOF. Scores for various *k*'s up to an upper bound are registered and afterward, the greatest of these scores is taken. Other than registering the LOF score for a solitary k, we additionally consider this methodology in our assessment, alluding to it as LOF-UB (upper bound). For comparison reasons, strategy has been used by setting different upper bounds and averages the obtained results.

### 3.3 Connectivity- Based Outlier Factor

The connectivity based anomaly factor is like LOF, however the density estimation for the records is performed in an unexpected way. In LOF, the k-nearest neighbors are chosen in light of the Euclidean distance. This in a roundabout way expect, that the information is disseminated circularly around the example. On the off chance that this supposition is disregarded, for instance if highlights have a direct straight connection, the density estimation is erroneous. COF needs to remunerate this inadequacy and appraisals the nearby density of the area utilizing a briefest way approach, called the binding separation. Numerically, this anchoring separation is the base of the entirety of all separations associating all k neighbors and the occurrence. For basic cases, where highlights are clearly connected, this density estimation approach performs significantly more precise. The round thickness estimation of LOF can't distinguish the special case, however COF succeeded by interfacing the common records with each other for surveying the local density.

### 3.4 K- means Clustering

K-means clustering is a type of unsupervised learning, when it is unlabelled data. This algorithm is to discover groups in the information, with the quantity of groups represented by the variable *K*. The calculation works iteratively to assign out every data point to one of *K* clusters in view of the highlights that are given. Data points having similar features are clustered. The K-means clustering computation uses iterative refinement to convey a last result. The computation inputs are the amount of clusters *K* and the dataset record. The dataset collection is a gathering of highlights for every data point. The calculations begins with introductory evaluations for the *K* centroids, which can either be randomly produced or randomly chosen from the dataset collection. Every centroid characterizes one of the groups. In this progression, every data point is appointed to its closest centroid, in view of the squared Euclidean separation. All the state formally, if $c_j$ is the accumulation of centroids in set *C*, at that point every data point $x_i$ is assigned to a cluster.

$$\text{Distance}=\text{argmins } D(x_i, c_j) \qquad (3.1)$$

Where,

$x_i$ is Distance between instance and

$c_j$. is cluster center

The iteration is stopped when none of the cluster assignment changes. Average of all the points is taken as centroid. The clusters are populated by the following steps,

i. For each point in the cluster, the nearest point is the current centroid.
ii. Assign the points and update the locations of the centroid.
iii. The points are reassigned to their closest centroid.

Steps ii and iii are repeated until convergence, convergence points do not move between clusters and centroid stabilize. The *K* is selected by trying different *K*, looking at the change in the average distance to centroid, as *K* increases. Various procedures exists for validating *K*, including cross-validation, information criteria, the information theoretic jump strategy, the silhouette technique, and the G-means algorithm. Also, observing the appropriation of information focuses crosswise over gatherings gives knowledge on splitting the data for every *K*.

### 3.5 Association Rules- Apriori Algorithm

Association rule finds all frequent itemsets and generate strong association rules from the frequent itemsets. Apriori algorithm is mining frequent itemsets for Boolean association rules. The Apriori algorithm is an unsupervised algorithm used to find frequent itemsets. Apriori utilizes a "bottom up" approach, where visit subsets are broadened one thing at any given moment. Apriori algorithm works on its two basic principles, first that if an itemset occurs often then all subset of itemset happens habitually and other is that if an itemset happens occasionally then all superset has rarely events.

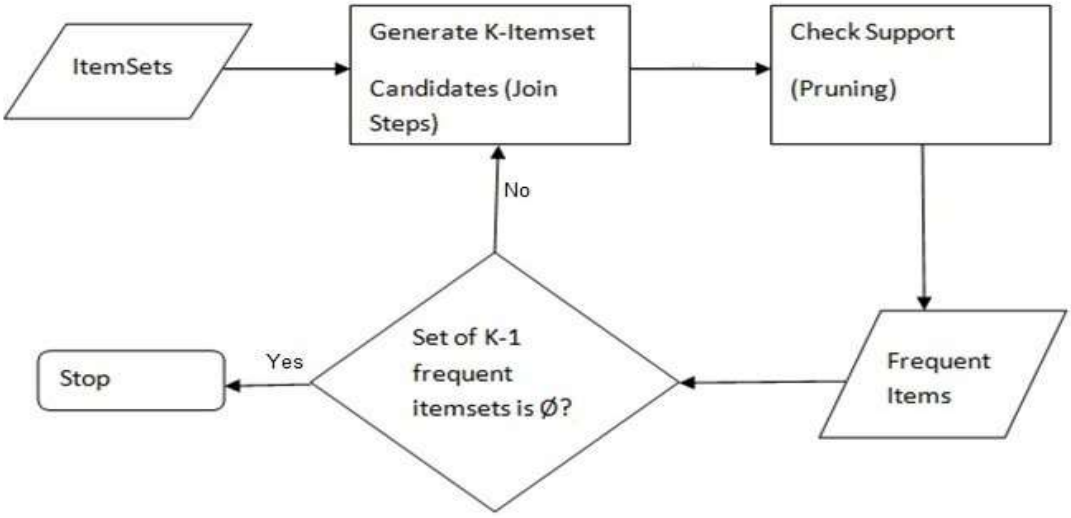


**Fig.2 Overall process of Apriori algorithm**

As mentioned in the fig.2, it is level-wise search method. k- itemsets (itemsets with k items) are utilized to investigate (k+1)- itemsets from value-based databases for Boolean association rules. To start with, the set of frequent 1-itemsets is found (denoted L1). L1 is utilized to discover L2, the set of frequent 2-itemsets. L2 is used to find L3, and so on, until no frequent k-itemsets can be found, ν generate strong association rules from the frequent itemsets. The name of the algorithm depends on the way that the algorithm utilizes earlier information of frequent items, utilizes an iterative approach known as level wise search, where k-items are utilized to explore k+1 items.

Apriori property:

- Apriori property is used to reduce the search space.
- All nonempty subset of frequent items must be also frequent. Anti-monotone in the sense that if a set cannot pass a test, all its supper sets will fail the same test as well.
- Reducing the search space to avoid finding of each Lk requires one full scan of the database (Lk set of frequent k-itemsets)
- If an itemset I does not satisfy the minimum support threshold min_sup, the I is not frequent, P(I) <min_sup
- If an item A is added to the itemset I, then the resulting itemset cannot occur more frequent than I, therefore I ∪ A is not frequent, P(I ∪ A) <min_sup.Apriori algorithm works well for the large dataset.

**Table 1. Performance Measures**

| Techniques | Global Detection | Speed | Dataset |
|---|---|---|---|
| k-NN global anomaly detection | Good | Medium | Medium |
| Local Outlier Factor(LOF) | Bad | Medium | Small |
| Connectivity-based Outlier Factor(COF) | Bad | Medium | Small |
| K-means Clustering | Good | Medium | Large |
| Apriori Algorithm | Good | Medium | Large |

## CONCLUSION

This survey focuses on various unsupervised anomaly detection techniques. As per the table 1. the techniques pertaining to nearest neighbor approach predicts both the local and global anomalies in the data. The high variance in clustering-based algorithms is very likely due to the non-deterministic nature of the underlying *k*-means clustering algorithm. For larger dataset clustering based algorithms are the better choice, where computation is faster. Nearest neighbor works well for most of the scenario. As a nutshell, this survey suggests that the choice of the algorithm must be solely based on the dataset considered for the application.

## REFERENCES

[1] Weiwei Chen and Fangang Kong. 2017. A Novel Unsupervised Anomaly Detection Approach for Intrusion Detection System. IEEE 3rd International Conference on Big Data Security on Cloud.
[2] Abdul Samad bin Haji Ismail, Abdul Hanan Abdullah, Kamalrulnizam bin Abu Bak, MdAsri bin Ngadi, DahliyusmantoDahlan and WitchaChimphlee. 2008. A Novel Method for Unsupervised Anomaly Detection using Unlabelled Data. International Conference on Computational Sciences and Its Applications ICCSA.
[3] Peiguo Fu and Xiaohui Hu. 2016. Biased-Sampling of Density-based Local Outlier Detection Algorithm. 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD).
[4] Tao Wang, Wenbo Zhang, Jun Wei and Hua Zhong. 2012. Workload-Aware Online Anomaly Detection in Enterprise Applications with Local Outlier Factor. IEEE 36th International Conference on Computer Software and Applications.
[5] Tian Huang, Yan Zhu, Qiannan Zhang, Yongxin Zhu, DongyangWang, MeikangQiu and Lei Liu. 2013. An LOF-based Adaptive Anomaly Detection Scheme for Cloud Computing. IEEE 37th Annual Computer Software and Applications Conference Workshops.
[6]Jian Tang, Zhixiang Chen, Ada Wai-chee Fu and David W. Cheung. 2002.Enhancing Effectiveness of Outlier Detections for Low Density Patterns.Springer-Verlag Berlin Heidelberg.
[7] R. Kumari, Sheetanshu, M.K. Singh, R. Jha and N.K. Singh. 2016. Anomaly Detection in Network Traffic using K-mean clustering. 3rd Int'l Conf. on Recent Advances in Information Technology.RAIT.
[8] VasserYasami, SiavashKhorsandi, Saadat Pour Mozaffari and ArashJalalian. 2018 An Unsupervised Network Anomaly Detection Approach by K-MeansClustering& ID3 Algorithms. IEEE.
[9] EhsanSaboori, ShafighParsazad and YasamanSanatkhani,. 2010. Automatic Firewall rules generator for Anomaly Detection Systems with Apriori Algorithm. 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE).
[10] Tian-rui Li and Wu-ming Pan. 2005. Intrusion Detection System Based on New Association Rule Mining Model*. IEEE.
[11] FaridFathnia, FrooghFathnia and Mohammad HosseinJavidi D. B. 2017. Detection of Anomalies in Smart Meter Data: A Density-Based Approach. Smart Grid Conference (SGC).
[12] GoverdhanReddy Jidigaand P. Sammulal. 2014. Anomaly Detection using Machine Learning with a Case Study. IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT).
[13]Yi Zhang, Weiwei Chen, and Jason Black. 2010. Anomaly Detection in Premise Energy Consumption Data. IEEE, 978-1-4577-1002-5/11.
[14] VarunChandola, Arindam Banerjee and Vipin Kumar. 2009. Anomaly Detection : A Survey To Appear in ACM Computing Surveys.09.