Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

# Network Anomaly Detection Using Unsupervised Machine Learning :Comparative study

**Gheed Tawfeeq Waleed [a], Abeer Tariq Mawlood [b], Abdul Mohssen Jaber [c]**

[a] *Department of Computer science , University of technology , Baghdad, Iraq . Email : gheed94@yahoo.com*

[b] *Department of Computer science , University of technology , Baghdad, Iraq . Email : 110032@uotechnology.edu.iq*

[c] *Department of Computer science , University of technology , Baghdad, Iraq . Email : AbdulMoohse53@yahoo.com*

ARTICLE INFO

*Article history:*
Received: 06 /08/2019
Rrevised form: 21 /08/2019
Accepted : 25 /08/2019
Available online: 25 /09/2019

*Keywords:*

Network Intrusion Detection,
Unsupervised Machine Learning,
Clustering.

ABSTRACT

The enormous growth in computer networks and in Internet usage in recent years, combined with the growth in the amount of data exchanged over networks, have shown an exponential increase in the amount of malicious and mysterious threats to computer networks. Machine Learning (ML) approaches have been implemented in the Network Intrusion Detection Systems (NIDS) to protect computer networks and to overcome network security issues. Anomaly detection has important applications in different domains such as fraud detection, intrusion detection, customer's behavior and employee's performance analysis. In this paper we have taken the Bank credit card dataset for finding Outlier detection. four Clustering methods have been compared and considered BIRCH Algorithm to be the best for finding noise and very effective for large datasets than the other clustering algorithms .

DOI : 10.29304/jqcm.2019.11.4.621

## 1 . Introduction

Network anomaly can be defined as a variation in the standard behavior that is related to a network with the existence of intruder in a network or because of the overload taking place in the network. The standard functionality of network service is impacted by such anomalous events. There are various elements that are considered important in characterizing the network's normal behavior, including the type related to the application that runs on the network, the type related to the network data that will be computed, and the network's traffic volume. Furthermore, the Intrusion Detection System (IDS) can be defined as the procedure that is used to monitor the abnormal actions happening in a computer network or system and compare it with normal events for the purpose of identifying if there are any indications of intrusion. Intrusion indicates a malicious action that has the aim to compromise the availability, integrity and confidentiality of the components of the network in an effort to interrupt the network's security policy. [1]

Corresponding author Gheed Tawfeeq

Email addresses: gheed94@yahoo.com

Communicated by Qusuay Hatim Egaar

The Internet has become a huge part of our daily life and an important tool nowadays. It is helpful for people in a wide range of fields, like entertainment, education, business, an so on. It's been utilized as a significant component of different business models. Thus, it is of a major concern to transmit sensitive information. Intrusion detection can be considered as a main research issue for personal and business networks. Due to the existence of various attacks on the network in the internet, Internet-based attacks could be prevented by the use of some systems, [2] specifically, IDSs which help the network in resisting external attacks. The main aim of Intrusion Detections Systems is providing wall of defense for the purpose of confronting the attacks of computer systems on the Internet. Intrusion Detections Systems could be utilized for detecting various types of malicious network communication and computer system utilization, while traditional firewalls cannot achieve such task. Intrusion detection depends on the suggestion that the behavior of legal users and intruders is different.

Clustering can be defined as an important approach for the anomaly-based un-supervised detection of intrusions. A standard specification related to data mining [3]; clustering is an approach used to group set of objects according to their similarities. The cluster's similarity is more and the cluster's dissimilarity is distinct, also Clustering can be considered as a kind of unsupervised study approach. Such approach could be used on the un-labeled data, it operates by dividing similar data to same class and dividing dissimilar data to distinct classes. Unsupervised anomaly-based detection frequently try to cluster (test data-set) in groups of similar samples that could be either normal or anomaly data.

Generally, IDSs could be divided into 2 groups: misuse detection and anomaly detection according to their detection methods. Anomaly detection have the aim of determining if the variation from recognized normal usage pattern could be labeled as intrusion while misuse detection applies patterns related to weak system spots or familiar attacks for identifying the intrusions. Certain anomaly detection systems have been created according to various ML approaches. For example, certain researches use single learning approaches, including support vector machines (SVMs), genetic algorithms (GA), neural networks (NNs), and so on, while certain systems depends on combining different learning approaches, like hybrid or set of methods. Specifically, such approaches are created as classifiers, that are utilized for

classifying or recognizing if the incoming Internet access is attack or just normal access.
Statistics community examined the problem of detecting or outliers from nineteenth. In the past decades, machine learning gained a lot of importance in anomaly detection. The researchers has developed a lot of anomaly-based intrusion detection methods.[8]

Conventional machine learning approaches limited the capability of processing real data in its raw form. Therefore, to construct feature extractor involved years of considerable work and knowledge for transforming the raw data to its appropriate feature vector and representation from which learner machine can be classifying the patterns in input .

The presented study has been presented in the following way. Section two offers summary related to the Centroid-Based Technique: K-Means approach , Section three describe the hierarchical clustering and BIRCH clustering approach, section four presents the Density based clustering , section five concerned with the performance measurement utilized for experiment, section six presents the data set utilized for the experiments. section seven is concerned with a comparison between the offered approaches. Section eight provides discussion and conclusion and discussion for future work.

## 2. Centroid-Based Technique: The k-Means Method

One of the major conventional clustering algorithms is the K-means algorithm, the data in this algorithm will be divided into k clusters and ensure that data in same cluster are considered to be similar, whereas low similarities will be presented when the data is in various clusters. At first, the K-means algorithm select K data item in a random way to be initial cluster center, after that, the remaining data will be to clusters with uppermost similarity according to its distance from the center of the cluster, then recalculate the center of the clusters of all clusters. Repeating the procedure till all the centers of clusters remain the same with no change. Therefore, the data will be split to K clusters. Regrettably, the K-means clustering could have some sensitivity to the outliers, and the set of objects closer to centroid could be empty, where the case centroids can not be updated .

K-means[10] take input parameter(k) and partition set of n objects to k clusters in order to result in low intercluster similarity and high intracluster similarity. The cluster similarity will be estimated conforming with the object's mean value in the cluster, that could be seen as the center of gravity or centroid of cluster. "In what way k-means algorithm function?" K-means algorithm functions in the following way. Initially, it chooses k of objects in a random way, each of them represent center or mean of cluster. For the objects that remained, object will be specified to cluster to which it has the most similarity, according to distance between cluster mean and the object. After that, it works on calculating new mean for all clusters. Such procedure iterate until no change occurs. Usually, square-error standard will be applied, specified by the following equation:

$$\sum_{i=1}^{k} \sum_{p\in Ci} |p - mi|^{2} \qquad (1)$$

In which E can be considered as summation of square error regarding all the dataset objects, p can be defined as the point which represent certain object; mi is specified as the mean related to the cluster Ci (p and Ci are multi-dimensional). Put differently, for all the objects in all clusters, distance from an object to the center of its cluster will be squared, also the distance will be summed. Such standard have the goal of making the created k clusters as separate and compact as feasible.

Certain benefits of applying K-means algorithm is that the flexibility is more efficient for large datasets, when the time complexity is o(tkn), and t defines the algorithm's iteration times, k can be defined as the number of the clusters, and n is specified as the number of data points in data set.

**Algorithm 2.1**

**Input** : K(Number of clusters) , D(dataset)

**Output :** a set of k clusters

**Method:**

**Step 1:** Arbitrary choose k objects from D as the initial cluster centers .

**Step 2 :** (re)assign each object to the cluster to which the object is the most similar , based on the mean value of the objects in the cluster.

**Step 3 :** Update the cluster means, by calculating the mean value of the objects for each cluster.

**Step 4 :** Repeat steps 2&3 until no change happens.

Disadvantage cluster number must be provided initially, yet such number is typically received post clustering. K-means don't have the ability of handling data in categorical attribute, as it is considered to have high sensitivity to the isolated points. It does not have the ability of discovering clusters of non-ball shape or clusters which are quite dissimilar. It typically trap in a local optimization rather than the global optimization. Furthermore, the results related to the algorithm are not balanced, there will be entirely different clustering results with the same input parameter.

K-means approach could be used just in the case when defining the cluster's mean. This could not be the situation in certain applications, like involving data with categorical attributes. K-means approach is not applicable for detecting non-convex shape clusters or different-size clusters. Furthermore, it has high sensitivity to the outlier data points and to noise due to the limited number of data could substantially impact mean value. Limited modifications of k-means approach exists. These could differ in selecting initial k means, calculating dissimilarity, and plans for the calculation of cluster means. An exciting plan that frequently yield optimum results is applying hierarchical agglomeration algorithm, that specify the number of clusters and identify the preliminary clustering, and after that apply iterative relocation for improving clustering.

## 3. BIRCH: Balanced Iterative Reducing and Clustering Using Hierarchies

The Hierarchical Clustering can be defined as the procedure of creating maximal collection related to the sub-sets of objects (referred to as clusters), with a feature that any 2 clusters are nested or disjoint. Regularly, it could be noticed as creating rooted binary tree that have the objects as its leaves, clusters corresponds to leaves related to sub-trees. Hierarchical clustering create clusters' hierarchy, that could signify tree structure that is referred to as dendo-gram. The tree root contain single cluster which includes all observations, the leaves consistent with distinct observations. Generally, the algorithms utilized for hierarchical clustering could be agglomerative, through which one begins at the leave and continually merge the clusters, or divisive, where it starts at the root and recurrently splits those clusters. Any valid metric could be applied for measuring the similarity between pair of observations.

Hierarchical clustering split the data to nested group of partitions and could be beneficial for determining data taxonomies. Hierarchical clusters are created via agglomerative algorithms through bottom-up method, in which all the examples are distinct cluster, also the clusters will be iteratively merged with neighbors. There are 2 major agglomerative clustering methods; complete link and single link. The two approaches are considered to be graph-based: each all the examples are a vertex; the edges will be added depending on distance between pairs of vertices. The level of hierarchical cluster will be determined via distance threshold: edge will be added to graph in the case when 2 examples are separated through a distance that is < threshold.

Connected and fully connected elements may be respectively defined as clusters [12] for single and complete link approaches. Hierarchy cluster procedure is creating via repeatedly increasing threshold for producing larger clusters. As the complete and single link clustering estimate the distance between the dataset's pairs of objects they have greater time complexity compared to the partitional algorithms (K-means), yet, they generate the optimum solutions .[12]

One of the major contributions related to BIRCH is the formulation regarding a clustering task in a suitable way for extremely large data sets through making memory and time constraints explicit. Another major contribution of BIRCH is that it does exploit the remark that the space is typically not occupied in uniform way, and thus not all the data points are equally significant the clustering. BIRCH handles the points' dense region (or sub clusters) communally through storing compact summarization. Therefore, BIRCH will reduce the clustering problem of original data points in single clustering problem the group of summaries that is extremely smaller when compared to original data set. BIRCH is considered to be incremental, in the way that the decisions of clustering are taken with no scan for all the data points or all the present clusters. [5]

BIRCH aims to take advantage of the current memory for the purpose of deriving the most efficient sub-clusters (for ensuring the precision) and reducing the costs of I/O (for ensuring the effectiveness) through organizing clustering and decreasing the procedure through the use of in-memory balanced tree structure related to the bounded size. This approach presents 2 notions, clustering feature and clustering feature tree (CF tree), that are applied for summarizing the representations of the clusters that helps clustering approach to have decent speed and scalability in huge data-bases and for making it efficient dynamic and incremental clustering regarding the incoming objects.

```
Algorithm 3.1 . BIRCH
Input : Dataset ,Threshold T, Maximum radius of cluster R, the branching factor B.
Output : Computation of cf points ( cf is the number of points in a cluster N, linear sum of the points in the cluster LS , the square sum of N data SS).
Step 1 : Load the dataset into the memory
Step 2  : Initial in- memory cf-tree is constructed with one scan of the dataset
```

Step 3 : Rebuild the cf-tree with a larger T.

Step 4 : Use the clustering Algorithim on Cf leaves.

Step 5 Do additional passes over the dataset and reassign data points to the closest centroid point (from step 4)

CF-tree can be[9] defined as a height-balanced tree that have 2 parameters: threshold T and branching factor (L for the leaf nodes and B for the non-leaf nodes). All the non-leaf nodes consist of maximum B entries regarding form in which i=1,2,..B, “child” considered to be a pointer to the i$^{th}$ child node, also CFi is considered as the clustering feature entry of sub cluster specified via the child. Leaf node consist of L entries, all entries are CF. Additionally, all leaf nodes have 2 pointers (prev) and (next), that are utilized for grouping all the leaf nodes for effective scans. Leaf node define sub-cluster created of all sub-clusters defined via its entries. However, each entity in the leaf node should meet a threshold, regarding threshold value T, diameter (or radius) of all leaf entries must be < T. Tree size can be considered as a function related to T.

The tree will keep getting smaller as T is getting larger. We specify node to fit in page with size P, in which P is defined as BIRCH parameter. As soon as determining the dimension d related to data space, the size related to no-leaf and leaf entries are recognized, then L and B will be defined through P. Thus, P have the ability of varying the performance tuning. It is applied for guiding new insertion the appropriate sub-cluster for clustering like the B+-tree is applied for guiding new insertion in the suitable sorting position. Yet, clustering feature-tree is an extremely compact representation regarding the data-set since all entries in leaf node is no single data point yet sub-cluster (that absorb as many data points as particular threshold value allow).[6]

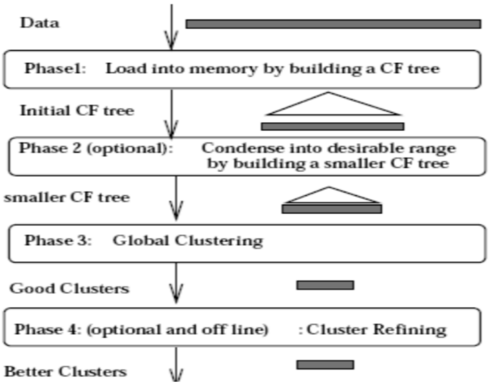

Data
Phase1: Load into memory by building a CF tree
Initial CF tree
Phase 2 (optional): Condense into desirable range by building a smaller CF tree
smaller CF tree
Phase 3: Global Clustering
Good Clusters
Phase 4: (optional and off line) : Cluster Refining
Better Clusters

**Figure1. BIRCH overview.**

Figure1 presents BIRCH overview ,which involves 4 stages, which are: (i)Loading, (ii)Optional Condensing, (iii)Global Clustering, and (iv) Optional Refining. The major aim of the first stage is scanning all data and building initial in-memory clustering feature -tree through the use of certain amount of memory and to recycle the disk space. Clustering feature -tree have the aim of reflecting clustering information regarding the data set in as many details as feasible depending on the memory limits. As the crowded data points are collected in sub- clusters, also the sparse data points have been eliminated as the outliers, this stage create in-memory data summary.

## 4. Density-based Methods

The density-based approaches depends on simple supposition: the clusters are thought out to be a dense areas in data space which have been divided via areas of a lower density. The major concept is keep growing the cluster on condition that certain threshold is exceeded via the density in neighborhood. This indicates that for all data points in particular cluster, neighborhood related to particular radius should consist of no less than minimum number of points. Such approaches are efficient at filtering out the outliers and detecting arbitrary-shape clusters.

An example of the density based methods is Density based spatial clustering of applications with noise (DBSCAN). DBSCAN can be defined as first-density-detecting system that is based on 2 concepts (minpts and epsilon). Epsilon defines the radius regarding a search circle, while minpts defines the number of minimal neighbors in search circle. Such concepts are used for examining e-neighbors contained in the objects. Through using such expansion, DBSCAN could precisely identify arbitrary patterns and clusters of different size, and filters noise.[7]

## 6. Dataset

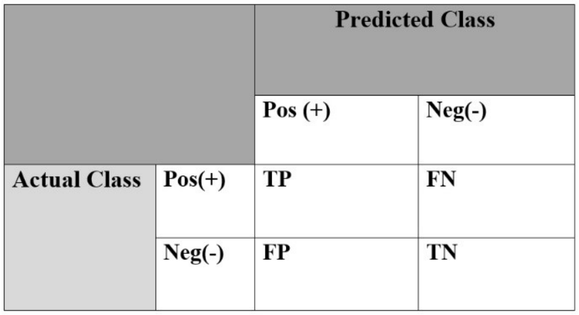
I this study we used the German Credit Risk dataset [13] which conatins 1000 data objects or attributes with 20 categorical attributes prepared by Prof.Hofmann . where each entry represents a person who takes a credit by a bank .Each person is classified as good or bad credit risks according to the set of attributes . We choose this dataset because it has enough information for this experiment and has a mix of both continuous or categorical variables .

The expirement is done on the following nine attributes: Age , Sex, Job , Housing, Saving accounts, checking account , credit amount , Duration , purpose . and ignore the rest of the columns because their description are obscure .

## 5. Performance measures

Estimating the performance regarding the clustering model is done according to counts the test records properly and wrongly predicted via the model. Such counts will be arranged in a table that is referred to as the confusion matrix, this table summarize some instances that have been predicted incorrectly or correctly through classification model and it is provided in Table1.

**Table(1) . Binary classification Confusion matrix**



| | | Predicted Class | |
|---|---|---|---|
| | | Pos (+) | Neg(-) |
| Actual Class | Pos(+) | TP | FN |
| | Neg(-) | FP | TN |

The next terminologies are frequently applied in the case of indicating counts in confusion matrix. [11]

- True positive (TP), that indicates the number of the positive samples that has been predicted properly via classification model.
- True negative (TN), that indicates the number of the negative samples that has been predicted properly via classification model.
- False negative(FN), that indicates the number of the positive samples that has been predicted incorrectly by classification model.
- False positive (FP), that indicates the number of the negative samples that has been predicted incorrectly via classification model.

Counts in the confusion matrix could be defined with regard to percentages. True positive rate (TPR) that is known as (Recall) measure is specified as fraction of properly predicted positive observations to every observation in the actual class, i.e.,

**Recall = TP/(TP + FN).** **(2)**

Likewise, true negative rate (TNR) referred to the as (Precision) could be specified as fraction of the Precision is the ratio of properly projected positive observations to total projected positive observations. i.e.,

**Precision = TN/(TN + FP).** **(3)**

The last measurement that we used in our experiment is (f1 score ) that is an estimation of the accuracy of the test. It take into account precision and recall of the test for computing the score

***f*1 = 2(recall * precision / recall+ precision)** **(4)**

## 6. Experimental results

We applied each of the mentioned unsupervised machine learning methods (k-means , Hierarchical clustering , BIRCH, DBSCAN) On the given German dataset risk , and cluster the dataset into 2 classes the first one is the normal class which contains the data that did not exposed to any type of frauds , and the other is the fraud or anomaly class . And compare the results By calculating the previously mentioned measures of performance (Recall, Precision, f1 score) . To evaluate which system gives the best results in clustering the dataset.

Table(2): results of K-means

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Normal | 0.54 | 0.57 | 0.55 | 361 |
| Anomaly | 0.50 | 0.46 | 0.48 | 329 |
| Total | 0.52 | 0.52 | 0.52 | 690 |

Table(3): results of Hierarchical clustering

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Normal | 0.59 | 0.62 | 0.60 | 369 |
| Anomaly | 0.53 | 0.60 | 0.56 | 321 |
| total | 0.56 | 0.61 | 0.58 | 690 |

Table(4): results of BIRCH

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Normal | 0.66 | 0.73 | 0.71 | 380 |
| Anomaly | 0.64 | 0.71 | 0.69 | 310 |
| total | 0.65 | 0.72 | 0.70 | 690 |

Table(5): results of DBSCAN

| | precision | recall | f1-score | Support |
|---|---|---|---|---|
| Normal | 0.0 | 0.0 | / | 0 |
| Anomaly | 0.10 | 0.0 | / | 0 |
| total | 0.05 | 0.0 | / | 0 |

## 7. Conclusion

ML methods have gained significant consideration among intrusion detection studies for addressing the drawbacks of knowledge base detection approaches.

In the presented study, we offered clustering based machine learning approaches and utilized these methods in unsupervised anomaly based network intrusion detection. We examined anomaly detection approach via applying it to German credit card risk .

## References

[1] Leung, K. and Leckie, C. (2005) Unsupervised anomaly detection in network intrusion detection using clusters. Proc. of 28 Australasian conference on Computer Science - Volume 38, Newcastle, NSW, Australia, January/February, pp. 333–342. Australian Computer Society, Inc. Darlinghurst.

[2] Leon, E., Nasraoui, O., and Gomez, J. (2004) Anomaly detection based on unsupervised niche clustering with application to network intrusion detection. IEEE Congres on Evolutionary Computation, 1, 502–508.

[3] Chimphlee, W., Abdullah, A. H., Sap, M. N. M.,Chimphlee, S., and Srinoy, S. (2005) Unsupervised clustering methods for identifying rare events in anomaly detection. Proc. of World Academy of Science, Engineering and Technology, October.

[4] Zhong, S., Khoshgoftaar, T., and Seliya, N. Clustering based network intrusion detection. Int'nl J of Reliability, Quality and Safety Engineering,14.

[5] George Kollios,Dimitrios,Gwopulos,Nick Koudas,Stefan Berchtold "Efficient Biased Sampling for Approximate Clustering and Outlier Detection in Large Datasets" IEEE Transactions on Knowledge and Data Engineering.2003.

[6] Ng, Raymond T. and Han, Jiawei, Efficient and Effective Clustering Methods for Spatial Data Mining, Proc. Of VLDB, 1994.

[7] Tran Manh Thang, Juntae Kim. "The anomaly detection by using DBSCAN clustering with multiple parameters "2011 International conference on information science and applications, 2011.

[8] Ritika Wason. "deep Learning : Evolution and expansion", cognitive systems Research , 2018.

[9] Truong, Cao Duy, and Duong Tuan Anh. "An efficient method for motif and anomaly detection in time series based on clustering" , International Journal of Business Intelligence and Data Mining, 2015.

[10] Masih, Shraddha, Shruti Dubey, Dharmendra Pathak, and Neha Rahatekar. "Data mining of WHO data warehouse with PASW modeler" , 7 1% Exclude quotes Of f Exclude bibliography On Exclude matches < 1% 2011 3rd International Conference on Electronics Computer Technology, 2011.

[11] Wenxiu Ding, Xuyang Jing, Zheng Yan, Laurence T. Yang. "A Survey on Data Fusion in Internet of Things: Towards Secure and Privacy-Preserving Fusion" , Information Fusion, 2018.

[12] Alec Pawling. "Anomaly detection in a mobile communication network" , Computational and Mathematical Organization Theory, 10/09/2007.

[13] Dateset is available online on : https://www.kaggle.com/uciml/german-credit/metadata