Comparative Study of Data Mining Techniques to Enhance Intrusion Detection

Mitchell D'silva, Deepali Vora
M.E. Information Technology, Vidyalankar Institute of Technology, Mumbai, India.
Asst. Professor Information Technology Department, Vidyalankar Institute of Technology, Mumbai, India.

Abstract

Today, Intrusion Detection Systems have been employed by majority of the organizations to safeguard the security of information systems. Firewalls that are used for intrusion detection possess certain drawbacks which are overcome by various data mining approaches. Data mining techniques play a vital role in intrusion detection by analyzing the large volumes of network data and classifying it as normal or anomalous. Several data mining techniques like Classification, Clustering and Association rules are widely used to enhance intrusion detection. Among them clustering is preferred over classification since it does not require manual labeling of the training data and the system need not be aware of the new attacks. This paper discusses three different clustering algorithms namely K-Means Clustering, Y-Means Clustering and Fuzzy C-Means Clustering. K-Means clustering results in degeneracy and is not suitable for large databases. Y-Means is an improvement over K-means that eliminates empty clusters. Fuzzy C-Means is based on the fuzzy logic that allows an item to belong to more than one cluster and concentrates on the minimization of the objective function that examines the quality of partitioning. The performance and efficiency of Fuzzy C- Means clustering is the best in terms of intrusion detection than the other two techniques. Finally, the paper discusses the comparison of the three clustering algorithms.

Keywords— Intrusion detection, K-Means Clustering, Y-Means Clustering, Fuzzy C-Means Clustering, Data Mining, Firewall.

I. INTRODUCTION

Internet is widely spread in each corner of the world, computers all over are exposed to diverse intrusions from the World Wide Web. To protect the computers from these unauthorized attacks, effective intrusion detection systems (IDS's) need to be employed. Traditional instance based learning methods for Intrusion Detection can only detect known intrusions since these methods classify instances based on what they have learned. They hardly detect the intrusions that they have not learned before. Intrusion detection techniques are of two types namely; Misuse detection and Anomaly detection. Firewalls are used for intrusion detection but they often fail in detecting attacks that take place from within the organization. To overcome this drawback of firewalls, different data mining techniques are used that handle intrusions occurring from within the organization. Data mining techniques have been successfully used for intrusion detection in different application areas like bioinformatics, stock market, web analysis etc. These methods extract previous unknown significant relationships and patterns from large databases. The extracted patterns are then used as a basis to identify new attacks. Data Mining based IDS's require less expert knowledge yet provides good performance and security. These systems are capable of detecting known as well as unknown attacks from the network. Different data mining techniques like classification, clustering and association rule mining can be used for analyzing the network traffic and thereby detecting intrusions. Among the above, clustering algorithms are widely used for intrusion detection as they do not require manual classification of training data. This paper discusses the three clustering algorithms namely K-Means, Y-Means and Fuzzy C-Means describing how each technique overcomes the drawbacks of the previous one. A comparison between the three clustering techniques is made that shows which technique is the most efficient in terms of intrusion detection.

II. INTRUSION DETECTION SYSTEM

Intrusion detection system plays an important role in detecting malicious activities in computer systems. The following discusses the various terms related to intrusion detection.

A. Intrusion

Intrusion is a type of malicious activity that tries to deny the security aspects of a computer system. It is defined as any set of actions that attempts to compromise the integrity, confidentiality or availability of any resource [1].

i) Data integrity: It ensures that the data being transmitted by the sender is not altered during its transmission until it reaches the intended receiver. It maintains and assures the accuracy and consistency of the data from its transmission to reception.

ii) Data confidentiality: It ensures that the data being transmitted through the network is accessible to only those receivers who are authorized to receive the respective data. It assures that the data has not been read by unauthorized users.

iii) Data availability: The network or a system resource ensures that the required data is accessible and usable by the authorized system users upon demand or whenever they need it.

B. Intrusion Detection

Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect malicious activities taking place through the network. ID is an area growing in significance as more and more sensitive data are stored and processed in networked systems.

C. Intrusion Detection System

Intrusion Detection system is a combination of hardware and software that detects intrusions in the network. IDS monitor all the events taking place in the network by gathering and analyzing information from various areas within the network. It identifies possible security breaches, which include attacks from within and outside the organization and hence can detect the signs of intrusions. The main objective of IDS is to alarm the system administrator whenever any suspicious activity is detected in the network.

In general, IDS makes two assumptions about the data set used as input for intrusion detection as follows:

i) The amount of normal data exceeds the abnormal or attack data quantitatively.
ii) The attack data differs from the normal data qualitatively.

III. MAJOR TYPES OF INTRUSION ATTACKS

Most intrusions occur via network by using the network protocols to attack their target systems. These kinds of connections are labeled as abnormal connections and the remaining connections as normal connections. Generally, there are four categories of attacks as follows:

A. DoS – Denial of Service

Attacker tries to prevent legitimate users from accessing the service in the target machine. For example: ping-of-death, SYN flood etc [2].

B. Probe – Surveillance and probing

Attacker examines a network to discover well-known vulnerabilities of the target machine. These network investigations are reasonably valuable for an attacker who is planning an attack in future. For example: port-scan, ping- sweep, etc [2].

C. R2L – Remote to Local

Unauthorized attackers gain local access of the target machine from a remote machine and then exploit the target machine‘s vulnerabilities. For example: guessing password etc [2].

D. U2R – User to Root

Target machine is already attacked, but the attacker attempts to gain access with super-user privileges. For example: buffer overflow attacks etc [2].

IV. TECHNIQUES FOR INTRUSION DETECTION

Each malicious activity or attack has a specific pattern. The patterns of only some of the attacks are known whereas the other attacks only show some deviation from the normal patterns. Therefore, the techniques used for detecting intrusions are based on whether the patterns of the attacks are known or unknown. The two main techniques used are:

A. Anomaly Detection

It is based on the assumption that intrusions always reflect some deviations from normal patterns. The normal state of the network, traffic load, breakdown, protocol and packet size are defined by the system administrator in advance. Thus, anomaly detector compares the current state of the network to the normal behavior and looks for malicious behavior. It can detect both known and unknown attacks.

B. Misuse Detection

It is based on the knowledge of known patterns of previous attacks and system vulnerabilities. Misuse detection continuously compares current activity to known intrusion patterns to ensure that any attacker is not attempting to exploit known vulnerabilities. To accomplish this task, it is required to describe each intrusion pattern in detail. It cannot detect unknown attacks.

V. NEED OF DATA MINING IN

INTRUSION DETECTION

Data Mining refers to the process of extracting hidden, previously unknown and useful information from large databases. It extracts patterns and concentrates on issues relating to their It is a convenient way of extracting patterns and focuses on issues relating to their feasibility, utility, efficiency and scalability. Thus data mining techniques help to detect patterns in the data set and use these patterns to detect future intrusions in similar data. The following are a few specific things that make the use of data mining important in an intrusion detection system:

i) Manage firewall rules for anomaly detection.
ii) Analyze large volumes of network data.
iii)Same data mining tool can be applied to different data sources.
iv) Performs data summarization and visualization.
v) Differentiates data that can be used for deviation analysis.
vi)Clusters the data into groups such that it possess high intra-class similarity and low inter-class similarity.

VI. DATA MINING TECHNIQUES FOR INTRUSION DETECTION SYSTEMS

Data mining techniques play an important role in intrusion detection systems. Different data mining techniques like classification, clustering, association rule mining are used frequently to acquire information about intrusions by observing and analyzing the network data. The following describes the different data mining techniques:

A. Classification

It is a supervised learning technique. A classification based IDS will classify all the network traffic into either normal or malicious. Classification technique is mostly used for anomaly detection. The classification process is as follows:

i) It accepts collection of items as input.
ii) Maps the items into predefined groups or classes defined by some attributes.
iii) After mapping, it outputs a classifier that can accurately predict the class to which a new item belongs.

B. Association Rule

This technique searches a frequently occurring item set from a large dataset. Association rule mining determines association rules and/or correlation relationships among large set of data items. The mining process of association rule can be divided into two steps as follows:

i) Frequent Item set Generation
Generates all set of items whose support is greater than the specified threshold called as minsupport.

ii) Association Rule Generation
From the previously generated frequent item sets, it generates the association rules in the form of "if then" statements that have confidence greater than the specified threshold called as minconfidence.

The basic steps for incorporating association rule for intrusion detection are as follows [3]:

i) The network data is arranged into a database table where each row represents an audit record and each column is a field of the audit records.
ii) The intrusions and user activities shows frequent correlations among the network data. Consistent behaviors in the network data can be captured in association rules.
iii) Rules based on network data can continuously merge the rules from a new run to aggregate rule set of all previous runs.
iv) Thus with the association rule, we get the capability to capture behavior for correctly detecting intrusions and hence lowering the false alarm rate.

C. Clustering

It is an unsupervised machine learning mechanism for discovering patterns in unlabeled data. It is used to label data and assign it into clusters where each cluster consists of members that are quite similar. Members from different clusters are different from each other. Hence clustering methods can be useful for classifying network data for detecting intrusions. Clustering can be applied on both Anomaly detection and Misuse detection. The basic steps involved in identifying intrusion are follows [3]:
i) Find the largest cluster, which consists of maximum number of instances and label it as normal.
ii) Sort the remaining clusters in an ascending order of their distances to the largest cluster.
iii) Select the first K1 clusters so that the number of data instances in these clusters sum up to ¼`N and label them as normal, where ` is the percentage of normal instances.
iv) Label all other clusters as malicious.
v)After clustering, heuristics are used to automatically label each cluster as either normal or malicious. The self-labeled clusters are then used to detect attacks in a separate test dataset.

From the three data mining techniques discussed above clustering is widely used for intrusion detection because of the following advantages over the other techniques:
i) Does not require the use of a labeled data set for

training.
ii) No manual classification of training data needs to be done.
iii) Need not have to be aware of new types of intrusions in order for the system to be able to detect them.

VII. CLUSTERING TECHNIQUES USED IN IDS

Several clustering algorithms have been used for intrusion detection. All these algorithms reduce the false positive rate and increase the detection rate of the intrusions. The detection rate is defined as the number of intrusion instances detected by the system divided by the total number of intrusion instances present in the data set. The false positive rate is defined as total number of normal instances that were incorrectly classified as intrusions defined by the total number of normal instances. Some of the clustering techniques such as K-Means Clustering, Y-Means Clustering and Fuzzy C-Means Clustering are discussed below.

A. K-Means Clustering

K-Means algorithm is a hard partitioned clustering algorithm widely used due to its simplicity and speed. It uses Euclidean distance as the similarity measure. Hard clustering means that an item in a data set can belong to one and only one cluster at a time. It is a clustering analysis algorithm that groups items based on their feature values into K disjoint clusters such that the items in the same cluster have similar attributes and those in different clusters have different attributes. The Euclidean distance function used to compute the distance (i.e. similarity) between two items is given as follows:

$$d\ (p,q) = \sqrt{(p1-q1)^2+(p2-q2)^2} \quad (1)$$

where, p = (p1, p2,…, pm) and q = (q1, q2,…,qm) are the two input vectors with m quantitative attributes. The algorithm is applied to training datasets which may contain normal and abnormal traffic without being labeled previously. The main idea of this approach is based on the assumption that normal and abnormal traffic form different clusters. The data may also contain outliers, which are the data items that are very different from the other items in the cluster and hence do not belong to any cluster. An outlier is found by comparing the radiuses of the data items; that is, if the radius of a data item is greater than a given threshold, it is considered as an outlier. But this does not disturb the K-means clustering process as long as the number of outliers is small.

K-Means Clustering Algorithm is as follows:
i) Define the number of clusters K. For example, if K=2, we assume that normal and abnormal traffic in the training data form two different clusters.
ii) Initialize the K cluster centroids. This can be done by randomly selecting K data items from the data set.
iii) Compute the distance from each item to the centroids of all the clusters by using the Euclidean distance metric which is used to find the similarity between the items in data set.
iv) Assign each item to the cluster with the nearest centroid. In this way all the items will be assigned to different clusters such that each cluster will have items with similar attributes.
v) After all the items have been assigned to different clusters re-calculate the means of modified clusters. The newly calculated mean is assigned as the new centroid.
vi) Repeat step (iii) until the cluster centroids do not change.
vii) Label the clusters as normal and abnormal depending on the number of data items in each cluster.

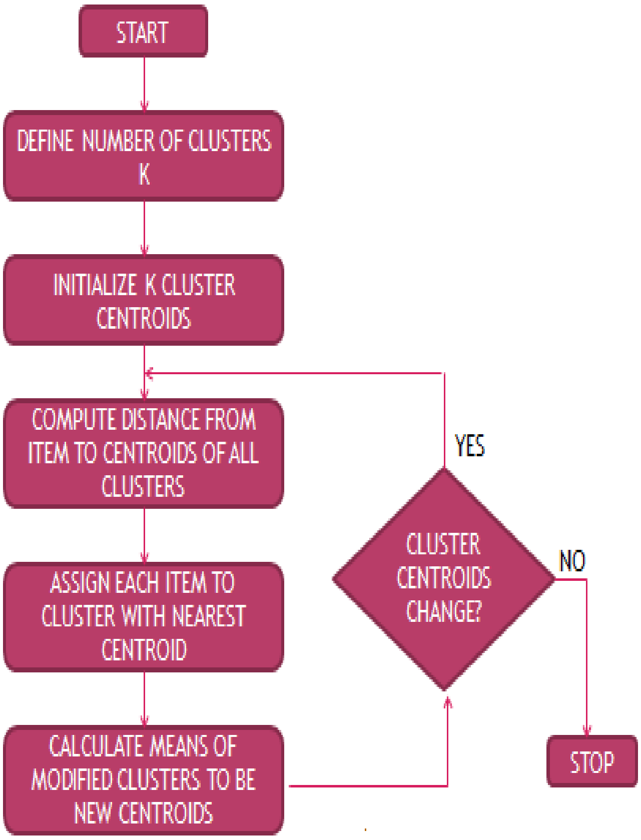
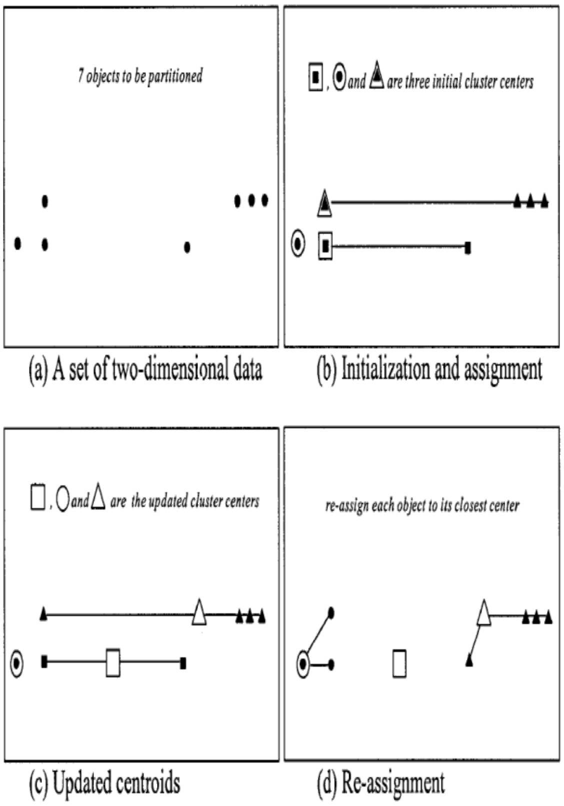

START
DEFINE NUMBER OF CLUSTERS K
INITIALIZE K CLUSTER CENTROIDS
COMPUTE DISTANCE FROM ITEM TO CENTROIDS OF ALL CLUSTERS
ASSIGN EACH ITEM TO CLUSTER WITH NEAREST CENTROID
CALCULATE MEANS OF MODIFIED CLUSTERS TO BE NEW CENTROIDS
CLUSTER CENTROIDS CHANGE?
YES
NO
STOP

Figure 1: Flowchart of K-Means Clustering

An essential problem of the K-means clustering method is to determine an initial partition and the appropriate number of clusters K. It also sometimes leads to degeneracy which means that the clustering process may end up with some empty clusters. The figure 2 shown below represents the generation of empty clusters.


7 objects to be partitioned
□, ○ and △ are three initial cluster centers
□, ○ and △ are the updated cluster centers
re-assign each object to its closest center

(a) A set of two-dimensional data (b) Initialization and assignment

(c) Updated centroids (d) Re-assignment

Figure 2: Generation of Degeneracy [4]

B. Y-Means Clustering

Y-means is another clustering algorithm used for intrusion detection. This technique automatically partitions a data set into a reasonable number of clusters so as to classify the data items into normal and abnormal clusters. The main advantage of Y-Means clustering algorithm is that it overcomes the three shortcomings of K-means algorithm namely dependency on the initial centroids, dependency on the number of clusters and degeneracy. Y-means clustering eliminates the drawback of empty clusters. The main difference between Y-Means and K- Means is that the number of clusters (K) in Y-Means is a self- defined variable instead of a user-defined constant. If the value of K is too small, Y-Means increases the number of clusters by splitting clusters. On the other hand, if value of K is too large, it decreases the number of clusters by merging nearby clusters. Y-Means determines an appropriate value of K by splitting and linking clusters even without any knowledge of item distribution. This makes Y-means an efficient clustering technique for intrusion detection since the network log data is randomly distributed and the value of K is difficult to obtain manually. Y-means uses Euclidean distance to evaluate the similarity between two items in the data set. Y- Means clustering has 3 main steps:

i) Assigning items to K clusters:
Depending on the value of K specified by the user, the items in a data set are assigned to the nearest clusters depending on the distance between the item and the centroid of each cluster

ii) Splitting clusters:
According to the Cumulative Standardized Normal Distribution Function, 99% of the instances of the cluster lie within the circle of radius 2.32 σ where σ is the standard deviation of the data. Therefore the threshold t = 2.32 σ is chosen. The area within the circle is called as the Confident Area of the cluster. Thus, all the points in the cluster that lie outside the Confident Area are considered as outliers [5]. These outliers are then removed from their current clusters and assigned as new centroids. The newly formed centroids then may attract some items from its adjacent clusters and thereby form new clusters. The splitting of clusters will continue until no outlier exists. Splitting turns clusters into finer grains thereby increasing the number of clusters and making the items within the same cluster more similar to each other.

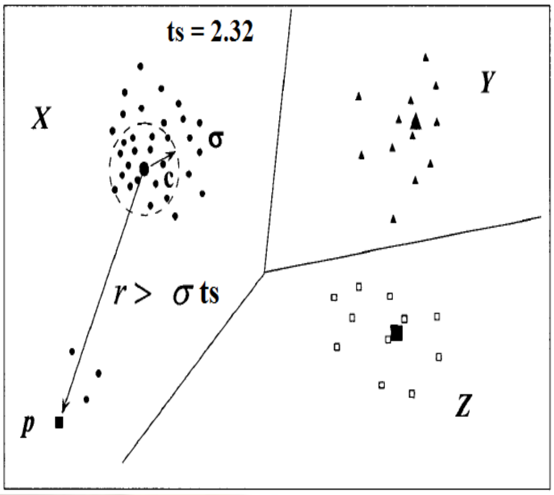


Figure 3: Before splitting the clusters [4]

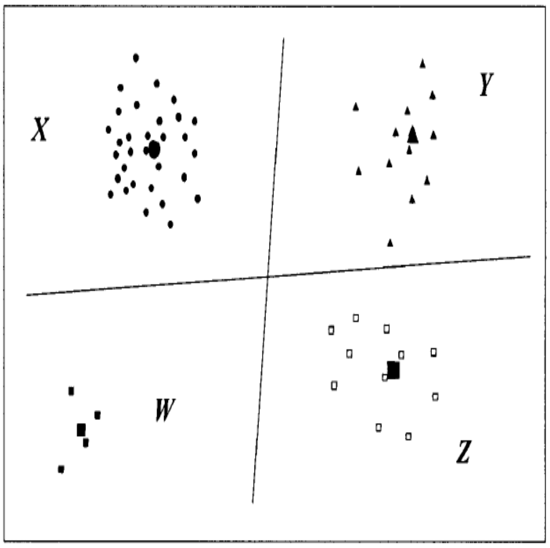


Figure 4: After splitting the clusters [4]

iii) Linking clusters:
When two adjacent clusters have an overlap, they can be merged into a larger cluster. The merging threshold is set to 2.32 σ i.e. whenever there are any data items in one cluster‘s Confident Area that also lie in another cluster‘s Confident Area, the two clusters can be merged [5]. When two clusters are merged their centroids are kept intact and no new centroid is created i.e. the new cluster has two centroids. The advantage is that the clusters can be of arbitrary shapes such as chain-like shapes.

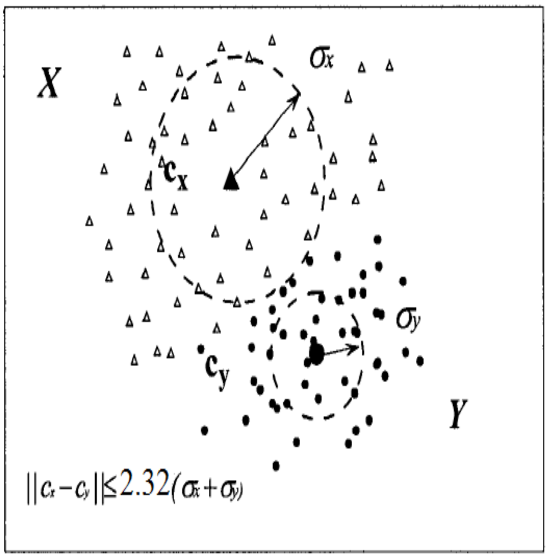


Figure 5: Linking clusters [5]

Y-Means Clustering Algorithm is as follows:
i) Define the number of clusters K.
ii) Initialize the K cluster centroids. This is done by randomly selecting K items from the data set.
iii) Compute the distance from each item to the centroids of all the clusters by using the Euclidean distance metric which is used to find the similarity between items in data set.
iv) Assign each item to the cluster with the nearest centroid. In this way all the items will be assigned to different clusters such that each cluster has all items with similar attributes.
v) After all the items have been assigned to different clusters re-calculate the means of modified clusters. The newly calculated mean is assigned as the new centroid.
vi) Check if there is degeneracy. If yes then remove the empty clusters and goto step (vii).
vii) Find the outliers, if any, of the cluster and then split the clusters.
viii) Check if there is degeneracy. If yes then remove the empty clusters and goto step (ix).
ix) Link the overlapping clusters if any.
x) Label the clusters as normal and abnormal.

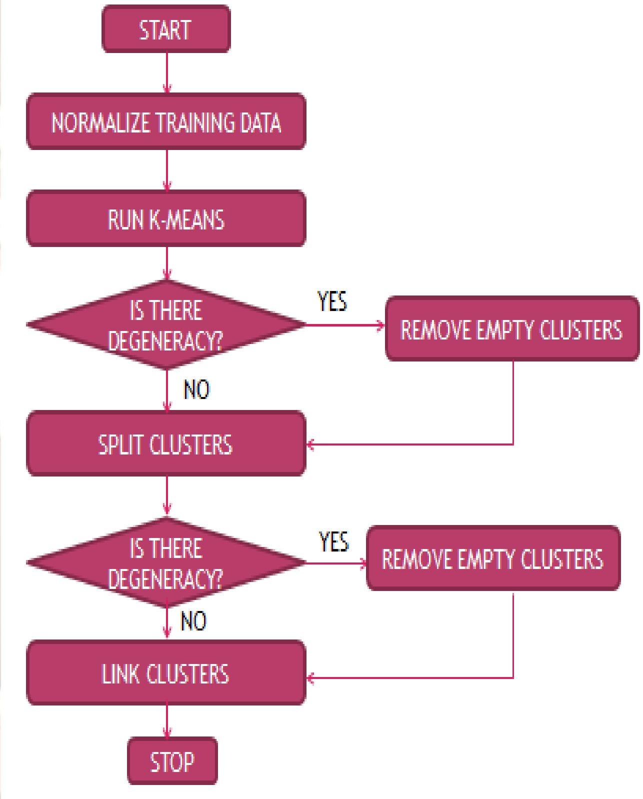


Figure 6: Flowchart of Y-Means Clustering

C. Fuzzy C-Means Clustering

Fuzzy C-Mean (FCM) is an unsupervised clustering algorithm based on fuzzy set theory that allows an element to belong to more than one cluster. The degree of membership of each data item to the cluster is calculated which decides the cluster to which that data item is supposed to belong. For each item, we have a coefficient that specifies the membership degree (u_{ij}) of being in the kth cluster as follows:

$$u_{ij} = \sum_{k=1}^{n} (d_{ij} / d_{ik})^{(2/m-1)} \quad (2)$$

where,

d_{ij} – distance of i^{th} item from j^{th} cluster

d_{ik} – distance of i^{th} item from k^{th} cluster and

m – fuzzification factor

The existence of a data item in more than one cluster depends on the fuzzification value (m) defined by the user in the range of [0, 1] which determines the degree of fuzziness in the cluster. Thus, the items on the edge of a cluster may be in the cluster to a lesser degree than the items in the center of the cluster. When m reaches the value of 1 the algorithm works like a crisp partitioning algorithm and for larger values of m the overlapping of clusters tends to be more [2]. The main objective of fuzzy clustering algorithm is to partition the data into clusters so that the similarity of data items within each cluster is maximized and the similarity of data items in different clusters is minimized. Moreover, it measures the quality of partitioning that divides a dataset into C clusters. FCM algorithm focuses on minimizing the value of the following objective function [6]:

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (u_{ij})^{m} \| x_i - v_j \|^2 \quad (3)$$

where,

m is any real number greater than 1,

u_{ij} is the degree of membership of x_i in the cluster j

x_i is the i^{th} of d-dimensional measured data

v_j is the d-dimensional center of the cluster

$||*||^2$ is any norm expressing the similarity between any measured data and the center.

Fuzzy C-Means Clustering Algorithm is as follows:

i) Randomly select –c‖ cluster centers.

ii) Initialize the fuzzy membership matrix u_{ij} using:

$$u_{ij} = \sum_{k=1}^{n} (d_{ij} / d_{ik})^{(2/m-1)} \quad (4)$$

iii) Calculate the fuzzy centers v_j' using:

$$v_j = \left(\sum_{i=1}^{n} (u_{ij})^m x_i \right) / \left(\sum_{i=1}^{n} (u_{ij})^m \right), \forall j = 1,2,\ldots c \quad (5)$$

iv) Update the membership matrix i.e goto step (ii)

iv) If ‖ $u(k+1)$ - $u(k)$ ‖ < threshold then stop else goto step (iii), where k is the iteration step.

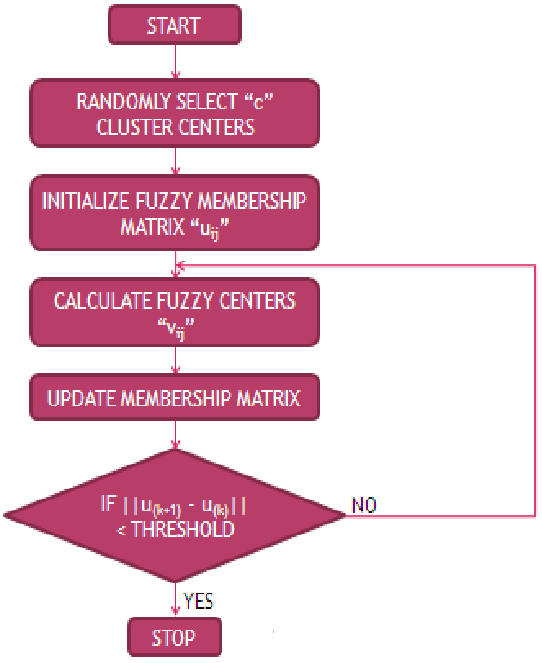

START
RANDOMLY SELECT "c" CLUSTER CENTERS
INITIALIZE FUZZY MEMBERSHIP MATRIX "u_{ij}"
CALCULATE FUZZY CENTERS "v_{ij}"
UPDATE MEMBERSHIP MATRIX
IF $||u_{(k+1)} - u_{(k)}||$ < THRESHOLD
NO
YES
STOP

Figure 7: Flowchart of Fuzzy C-Means Clustering

VIII. COMPARISON OF K-MEANS, Y-MEANS AND FUZZY C-MEANS CLUSTERING

This section presents the comparison of the three clustering techniques namely K-Means, Y-Means and Fuzzy C-Means clustering. The comparison is made by taking into account various criteria like the performance, efficiency, detection rate, false positive rate, purity etc. Each technique has some good features to overcome the drawbacks of the other technique.

Table 1: Comparison of K-Means, Y-Means and Fuzzy C- Means Clustering Techniques

Criteria	**K-Means**	**Y-Means**	**Fuzzy C- Means**
Input	Number of clusters K such that K<m Set of data items (x1,x2,..,xm)	Number of clusters K such that K<m Set of data items (x1,x2,..,xm)	Number of clusters c such that c<m Set of data items (x1, x2,..,xm) Set of cluster centers (v1,v2,..,vc)
Output	Set of K clusters where each cluster has similar items	Set of non-empty clusters where each cluster has similar items	Set of c clusters where each cluster has more similar items
Membership value	Does not exist	Does not exist	Has a membership value as "u_{ij}"
Computation Time	Simple and straight-forward so requires less time	Involves splitting and linking of clusters so requires more time	Involves the calculation of several formulas so requires more time
Purity of cluster	Low	High	High
Empty cluster generation	May or may not generate	No	No
Efficiency	Works well for small data sets	Works well for small as well as large data sets	Works well for small as well as large data sets
Number of clusters an item belongs	One	One	One or more than one cluster
Overall Performance	Depends on the initial number of clusters "K"	Does not depend on the initial number of clusters "K"	Depends on the initial number of clusters "c"
Shape of cluster	Works well for compact and globular clusters	Works well for both globular and nonglobular clusters	Works well for both globular and nonglobular clusters
Detection Rate	Highest	Higher	High

Criteria	**K-Means**	**Y-Means**	**Fuzzy C- Means**
False Positive Rate	Lowest	Low	Lower
Advantages	Simple and fast Works well for compact and globular Clusters	No Degeneracy Does not depend on the initial number of clusters K Works well with globular and nonglobular clusters	Measures the quality of partitioning Items can belong to more than one cluster Works well for globular and nonglobular clusters
Dis-advantages	Need to determine an appropriate number of clusters Degeneracy Does not work well with non- globular clusters Does not measure the quality of clusters	Does not measure the quality of clusters Time consuming	Performance depends on the initial number of clusters Time consuming

IX. CONCLUSION

Data mining techniques are widely used because of their capability to drastically improve the performance and usability of intrusion detection systems. Different data mining techniques like classification, clustering and association rule mining are very helpful in analyzing the network data. Since large amount of network traffic needs to be collected for intrusion detection, clustering is more suitable than classification in the domain of intrusion detection as it does not require labeled data set thereby reducing manual efforts. Data mining techniques can detect known as well as unknown attacks. Data mining technology helps to understand normal behavior inside the data and use this knowledge for detecting unknown intrusions. Three clustering algorithms namely K-means, Y-means and Fuzzy C-means have been discussed. Each of these has both advantages and disadvantages and are an improvement over the other. Among these Fuzzy C-Means clustering can be considered as an efficient algorithm for intrusion detection since it allows an item to belong to more than one cluster and also measures the quality of partitioning. The technique can be used for large data sets as well as data sets that have overlapping items. Moreover it does not generate any empty clusters and has the highest purity thereby creating clusters that consists of highly similar items. The main advantage of Fuzzy C-Means clustering for intrusion detection is the high detection rate and lower false positive rate that it offers. Although Fuzzy C-Means is an efficient technique it is time consuming. The performance of intrusion detection systems can be still improved by combining the features of Fuzzy C- Means clustering technique with some other technique so that it reduces the time required by Fuzzy C-Means for the clustering process and also increases the detection rate and decreases the false positive rate thereby making the intrusion detection system more accurate and effective.

REFERENCES

[1] Ye Qing, Wu Xiaoping and Huang Gaofeng, "An Intrusion Detection Approach based on Data Mining", 2nd International Conference on Future Computer and Communication, pp. no. 695 – 698, 2010 IEEE.

[2] Disha Sharma, "Fuzzy Clustering as an Intrusion Detection Technique", International Journal of Computer Science & Communication Networks, pp. no. 69 – 75, Vol 1 (1), September-

October 2011.

[3] Deepthy K Denatious and Anita John, “Survey on Data Mining Techniques to Enhance Intrusion Detection”, International Conference on Computer Communication and Informatics, Jan 2012.

[4] Yu Guan, “Y-Means: A Dynamic Clustering Algorithm’’.

[5] Sivanadiyan Sabari Kannan**,** “Y-means Clustering vs N-CP Clustering with canopies for Intrusion Detection’’.

[6] Ming-Chuan Hung and Don-Lin Yang, “An Efficient Fuzzy C-Means Clustering Algorithm”, pg. no. 225-232, 2001 IEEE.

[7] Yu Guan and Ali A. Ghorbani, Nabil Belacel, “Y-Means: A Clustering Method for Intrusion Detection’’, CCECE 2003 CCGEI 2003, Montreal May 2003 IEEE.

[8] E. Kesavulu Reddy, V. Naveen Reddy and P. Govinda Rajulu, “A Study of Intrusion Detection in Data Mining’’, Proceedings of the World Congress on Engineering, July 2011, Vol III, London, UK.

[9] Reema Patel, Amit Thakkar and Amit Ganatra, “A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems”, International Journal of Soft Computing and Engineering (IJSCE), pp. no. 265 – 271, ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[10] “Intrusion Detection Systems: Definition, Need and Challenges”.

[11] Harley Kozushko, “Intrusion Detection: Host-Based and Network-Based Intrusion Detection Systems”, Independent Study, September 11, 2003.

[12] Manish Joshi, “Classification, Clustering and Intrusion Detection System”, International Journal of Engineering Research and Applications (IJERA), pp.961-964, ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2, Mar-Apr 2012.

[13] Gerhard Munz and Sa Li, Georg Carle, “Traffic Anomaly Detection Using K-Means Clustering’’**.**

[14] Raymond Chi-Wing Wong and Ada Wai-Chee Fu, “Association Rule Mining and its Application to MPIS”.