International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 3 Issue 1, January-2014, pp: (557-564), Impact Factor: 1.252, Available online at: www.erpublications.com

A Comprehensive Study of Clustering Techniques to Analyze NSL-KDD Dataset and Research Challenges

Shailesh Singh Panwar, Rohit Sharma, Vipin Kumar, Vijay Maheshwari
Department of Computer Science and Engineering, Graphic Era University, Dehradun, India
School of Computer Science and Engineering, Shobhit University, Meerut, India
Department of Computer Science and Engineering, Graphic Era University, Dehradun,India

Abstract: Distinguishing malicious activities over Internet is a major challenge to the research community as well as to corporations. Several data mining techniques have been adopted for this purpose i.e. classification, clustering, association rule mining, regression, visualization etc. Clustering provides a better representation of network traffic in order to identify the type of data flowing through network. Clustering algorithms have been used most widely as an unsupervised classifier to organize and categorize data. In this paper we have analyzed four different clustering algorithms using NSL-KDD dataset. We tried to cluster the dataset in two classes i.e. normal and anomaly using K-means, EM, DB clustering and COBWEB. The main objective of this evaluation is to determine the class labels of different type of data present in intrusion detection dataset and to find out efficient clustering algorithm. The results of the evaluation are compared and challenges faced in these evaluations are than discussed.

Keywords: Clustering, NSL-KDD dataset, intrusion detection, K-means, EM, Density based clustering, COBWEB.

I. INTRODUCTION

Due to the popularization of the Internet and local networks, intrusion events to computer systems are growing. The rapid proliferation of computer networks has changed the prospects of network security. This generated a need of a system that can detect threats to the network instead of simply relying on intrusion prevention systems. Detecting such threats not only provides information on damage assessment, but also helps to prevent future attacks. These attacks are usually detected by tools referred to as Intrusion Detection System. Researchers have developed intrusion detection system for various environments depending upon the security concerns of different networks. The function of Intrusion Detection System is to gather and analyze information from various areas within a computer or a network to determine all possible security breaches. Over the past ten years, intrusion detection and other security technologies such as cryptography, authentication, and firewalls have increasingly gained in importance [1].

Data Clustering is considered an interesting approach for finding similarities in data and putting similar data into groups. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups [2].Clustering algorithms are used extensivelynot only to organize and categorize data, but arealso useful for data compression and model construction. By finding similarities in data, one can represent similar data with fewer symbols for example. Also if we can find groups of data, wecan build a model of the problem based on those groupings.Another reason for clustering is its descriptive nature which can be used to discover relevant knowledge in huge dataset.

In this paper we reviewed four of the clustering algorithms and tried to find out the different clusters of similar types. First a normalization process is done to make this dataset suitable for clustering. Than four of the clustering algorithms were implemented to cluster the instances in two clusters i.e. anomaly and normal. Number of instances in each cluster is identified. Performance of the four algorithms is computed on the basis of time taken and correctly classified instances. We implemented K-means, EM, Density Based Cluster, and COBWEB cluster to analyze the detection rate over NSL-KDD intrusion detection dataset. The result ant labeled clusters are than used to identify future instance belongingness. This process is shown in figure 1 above.

The rest of this paper is organized as follows: Section II includes the literature review about different kinds of work done by the various authors related to clustering in network intrusion detection systems. In section III a brief introduction to ID using clustering is given. The experiments and results are discussed in section IV. In section V we have discussed various research challenges followed by conclusion and future work in section VI.

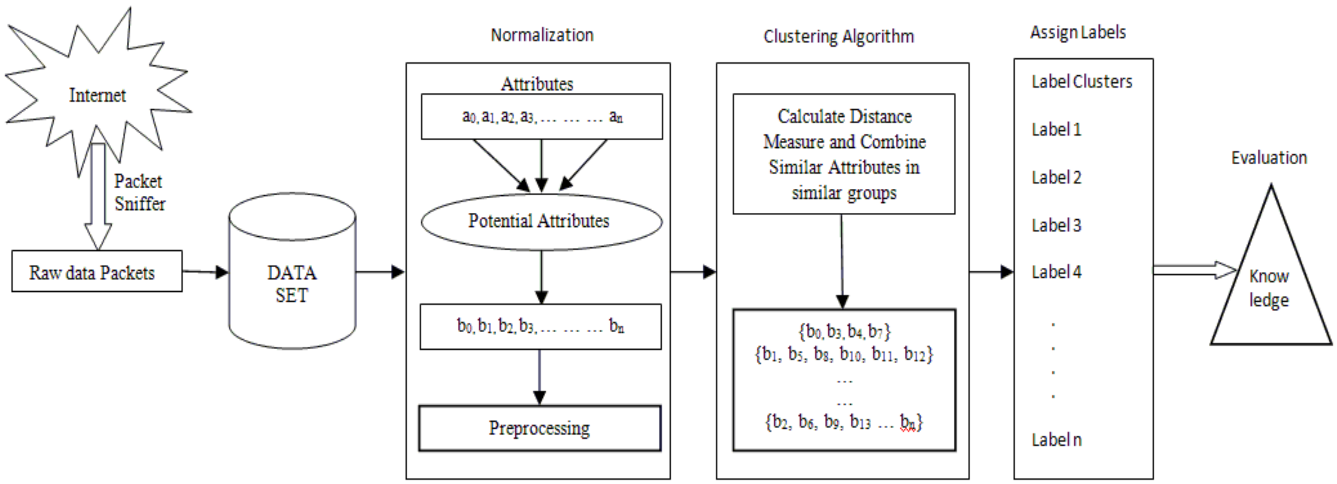

Fig 1: Block diagram of clustering process

II. RELATED WORK

Current anomaly detection is often associated with high false alarm with moderate accuracy and detection rates when it's unable to detect all types of attacks correctly. To overcome this problem, Muda et al.[3] proposed a hybrid learning approach through combination of K-Means clustering and Naïve Bayes classification. They cluster all data into the corresponding group before applying a classifier for classification purpose. An experiment is carried out to evaluate the performance of the proposed approach using KDD Cup'99 dataset. Result shows that the proposed approach performed better in term of accuracy, detection rate with reasonable false alarm rate.

H. Om et al[4] proposed a hybrid intrusion detection system that combines k-Means, and two classifiers: K-nearest neighbor and Naïve Bayes for anomaly detection. It consists of selecting features using an entropy based feature selection algorithm which selects the important attributes and removes the irredundant attributes. This system can detect the intrusions and further classify them into four categories: Denial of Service (DoS), U2R (User to Root), R2L (Remote to Local), and probe. The main goal is to reduce the false alarm rate of IDS.

Existing IDS techniques includes high false positive and false negative rate.Nadiammai et al. [5] implemented some of the clustering algorithms like k means, hierarchical and Fuzzy C Means, to analyze the detection rate over KDD CUP 99 dataset and time complexity of these algorithms. Based on evaluation result, FCM outperforms in terms of both accuracy and computational time.

Y. Qing et al. [6] presented an approach to detect intrusion based on data mining frame work. In the framework, intrusion detection is thought of as clustering. The reduction algorithm is presented to cancel the redundant attribute set and obtain the optimal attribute set to form the input of the FCM. To find the reasonable initial centers easily, the advanced FCM is established, which improves the performance of intrusion detection since the traffic is large and the types of attack are various. In the illustrative example, the number of attributes is reduced greatly and the detection is in a high precision for the attacks of DoS and Probe, a low false positive rate in all types of attacks.

The focus of Haque et al. [7] is mainly on intrusion detection based on data mining. The main part of Intrusion Detection Systems (IDSs) is to produce huge volumes of alarms. The interesting alarms are always mixed with unwanted, non-interesting and duplicate alarms. The aim of data mining is to improve the detection rate and decrease the false alarm rate. So, here we proposed a framework which detect the intrusion and after that, it will show the improvement of k-means clustering algorithm.

Poonam et al.[8] compares the performance of the four algorithms on outlier detection efficiency. The main objective is to detect outliers while simultaneously perform clustering operation. Denatious et al. [9] presents the survey on data mining techniques applied on intrusion detection systems for the effective identification of both known and unknown patterns of attacks, thereby helping the users to develop secure information systems.

III. INTRUSION DETECTION USING CLUSTERING

An Intrusion is method of comprising confidentiality, integrity, scalability and availability of network resources. It monitors and analyzes the user and network traffic, verifies system configurations and vulnerabilities and alerts the administrator through alarms.

In IDS we have two types namely Host IDS and Network IDS [5].

A. Host Based Intrusion Detection Systems (HIDS): Anomaly detection techniques can be applied for both host based and network based intrusion detection systems analyzes the sequential nature of data. But a point anomaly detection technique is not applicable in this domain. In HIDS, the Intrusion Detection is performed for a single host. So challenging issues in Host Based Systems is less compared to Network based Intrusion Detection. But individual IDS must be connected for each host.

B. Network Intrusion Detection Systems (NIDS): This type of systems deals with detecting intrusions over network and data appears in a sequential fashion. The intrusions occur as anomalous patterns. In NIDS, large networks of computers are connected to other networks and also to the Internet. Data can be determined at different levels of granularity. Challenging issues is more in this system, because the nature of anomalies changes over time because the intruder uses one network and tries to attack another network.

In figure 2 raw (binary) audit data is first processed into ASCII network packet information (or host event data), which is in turn summarized into connection records (or host session records) containing a number of basic features, such as service, duration, etc. Data mining programs are then applied to the connection records to compute the frequent patterns (i.e., association rules and frequent episodes), which are in turn analyzed to construct additional features for the connection records. Classification programs are then used to inductively learn the detection models. This process is of course iterative. For example, poor performance of the classification models often indicates that more patter mining and feature construction is needed [10].

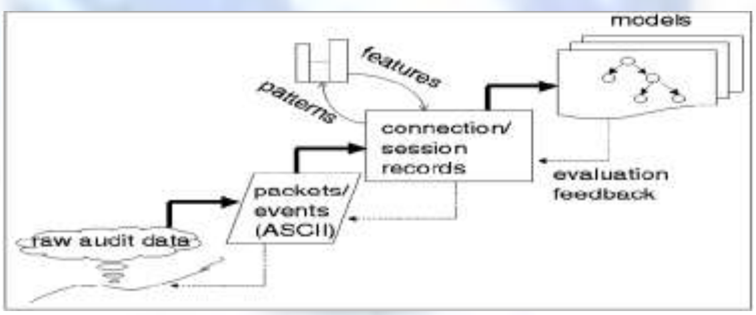

Fig 2: Building ID models using data mining (Lee et al. [10])

In a cluster there may be many groups according to the dataset it differs. But objects in one group are not similar to the objects in other groups and vice versa. Hence the goal of clustering is to include the essential group in a set of unlabeled data. A clustering algorithm tries to find natural groups of components/data based on some similarity. In addition, the clustering algorithm locates the centroid of a group of data-sets [5]. To determine cluster membership, the majority of algorithms evaluate the distance among a point and the cluster centroid. The output from a clustering algorithm is fundamentally a statistical description of the cluster centroid with the number of elements in each cluster. For clustering data points, there should be high intra cluster similarity and low inter cluster similarity. A clustering method which results in such type of clusters is considered as good clustering algorithm.

Clustering methods can be classified as [11]:

a) Hierarchical Clustering: Instead of clustering the whole dataset at once, stepwise procedure is followed for clustering the dataset. For example: division clustering, agglomerative clustering.

b) Partitional Clustering: In this type of clustering, data points are divided into k subparts based upon certain relevance criteria. For example: K-means clustering, Fuzzy c-means clustering and QT clustering.

The four clustering techniques analyzed in this paper are discussed as follows-

1) Simple K-Means: K-Means[12][13]is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through certain number of clusters (assume k clusters) fixed a priori [14]. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point it is necessary to re-calculate k new centroids as bar centers of the clusters resulting from the previous step. After obtaining these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop, one may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

2) EM (Expectation Maximization):In statistics, an expectation–maximization (EM) algorithm[15] is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters instatistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step[16]. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

3) Density based clustering: In density-based (DB) clustering[17], clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. In contrast to many newer methods, it features a well-defined cluster model called "density-reachability". Similar to linkage based clustering; it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects' range. The key drawback of density based clustering is that they expect some kind of density drop to detect cluster borders. Moreover they cannot detect intrinsic cluster structures which are prevalent in the majority of real life data.

Table I. List of Attributes in NSL-KDD Dataset

duration	protocol_type	Service
dst_host_srv_serror_rate	dst_host_rerror_rate	dst_host_srv_rerror_rate
flag	src_bytes	dst_bytes
land	wrong_fragment	urgent
hot	num_fialed_logins	logged_in
num_compromised	root_shell	su_attampted
num_root	nu_file_creations	num_shells
num_access_files	num_outbound_cmds	is_host_login
is_guest_login	count	srv_count
serror_rate	srv_serror_rate	rerror_rate
srv_rerror_rate	same_srv_rate	diff_srv_rate
srv_diff_host_rate	dst_host_count	dst_host_srv_count
dst_host_same_srv_rate	dst_host_diff_srv_rate	dst_host_same_src_port_rate
dst_host_srv_diff_rate	dst_host_serror_rate	class

COBWEB: It is an incremental system for hierarchical conceptual clustering.COBWEB [18]incrementally organizes observations into a classification tree. Each node in a classification tree represents a class (concept) and is labeled by a probabilistic concept that summarizes the attribute-value distributions of objects classified under the node. This classification tree can be used to predict missing attributes or the class of a new object.There are four basic operations COBWEB employs in building the classification tree. Which operation is selected depends on the category utility of the classification achieved by applying it. The operations are: Merging Two Nodes- Merging two nodes means replacing them by a node whose children is the union of the original nodes' sets of children and which summarizes the attribute-value distributions of all objects classified under them. Splitting a node-A node is split by replacing it with its children. Inserting a new node - A node is created corresponding to the object being inserted into the tree. Passing an object down the hierarchy - Effectively calling the COBWEB algorithm on the object and the sub-tree rooted in the node.

Table II. Clustering of Normal and Anomaly dataset

No. of Instances / Algorithms	Normal	Anomaly	Incorrectly Classified
K-Means	15242	9950	2557
Density Based Cluster	14228	10964	3379
EM	18304	6888	4859
Cobweb	15247	9945	4491

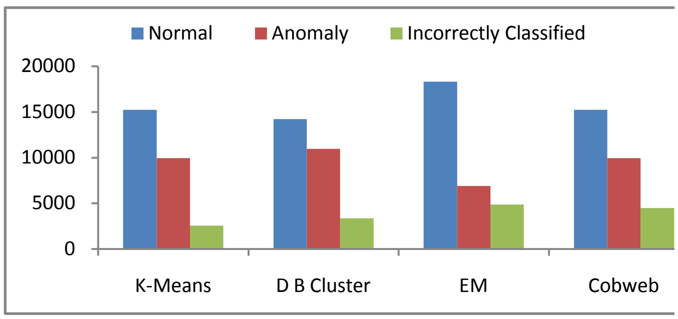

Fig III. No. of Clustered Instances

IV. EXPERIMENT AND RESULTS

This section has been divided in three sections – Setup, Results and Analysis.

A. Setup: WEKA platform was selected for the implementation of the selected algorithms. WEKA[19] is open source software issued under General Public License, developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. We conducted experiments on NSL-KDD[20] dataset (20%). NSL-KDD labeled dataset used in the experiment contains 25,192 records and data is either labeled as normal.Each record in NSL-KDD data set is a network linking record. Each link consists of 41 attribute properties containing 3 symbolic variables (protocol_type, service and flag). In order not to affect the clustering result, the attribute values need to be pre-treated. Firstly, these three symbolic attributes are removed and then all the remaining numerical attributes (39 attributes) are normalized in the range of [0 1] so that the attributes having higher values do not dominate over attributes with low values. The standard deviation transform is shown asfollows:

$$\dot{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{S_k} \quad (1)$$

The normalized transform is as follows:

Table III. Cluster representation in terms of percentage

	Normal (%)	Anomaly (%)	Incorrectly Classified (%)
K-Means	61	39	10.15
Density Based Cluster	56	44	13.41
EM	73	27	19.28
Cobweb	60	40	17.82

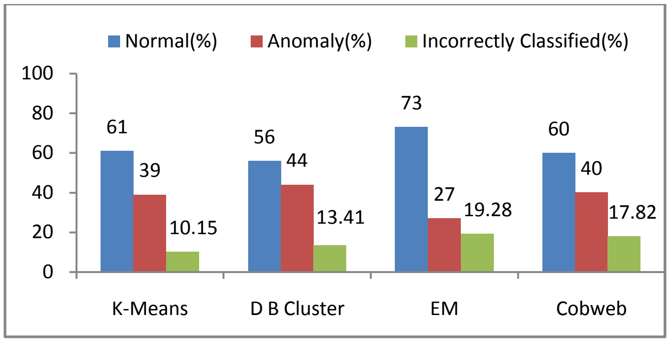

Fig IV. Percentage of classified clusters

Table IV. Time taken to build clusters

Algorithm	Time(seconds)
K-Means	3.39
Density Based Cluster	3.27
EM	25.68
Cobweb	161.19

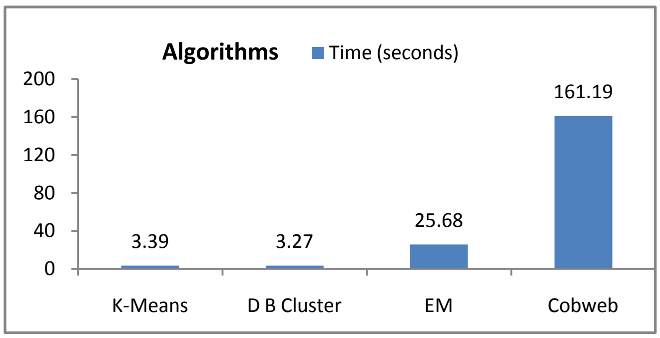

Fig V. Time Taken to build cluster model

$$\ddot{x}_{ik} = \frac{\dot{x}_{ik} - \dot{x}_{k_{min}}}{\dot{x}_{k_{max}} - \dot{x}_{k_{min}}} \quad (2)$$

Maximum iterations are set to 100 and number of cluster is set to 2. Table 1 consists of all 42 attributes of the NSL-KDD dataset. Three of the symbolic data attributes (protocol_type, service and flag) are removed as non-contributing. Clustering is performed over (25192x39) metrics.

B. Results: Intrusion detection dataset has been clustered in two modes: normal data and anomaly data. Results of the four clustering algorithms have been compared and time complexity to build the cluster model is evaluated. Total no of instances in the training dataset is 25192. Four Algorithms K-Means, Density Based Cluster, EM and Cobweb are implemented on this dataset and results of these algorithms are compared. It represents the results obtained by the experiment and represented in figure 2. Table 2 represents the four algorithms based on the values of clustered instances.

C. Analysis: Results are compared on the basis of Time and correctly classified instances. The study analyses the NSL-KDD dataset and shows that K-Means algorithm out performs on the bases of time (3.39 sec) and lowest number of incorrectly classified instances (10.15%). EM also performs well in terms of correctly clustered instances but time taken in building the clusters is very high as compared to K-Means clustering algorithm. The key drawback of DB Cluster is that they expect some kind of density drop to detect cluster borders. A common use case in artificial intelligence data is that the cluster borders produced by these algorithms will often look arbitrary, because the cluster density decreases continuously. So Density-Based Cluster algorithm results in lower accuracy i.e. 56%.

V. RESEARCH CHALLENGES

In dynamic network environment, where the traffic patterns are always changing and huge amount of data are coming every second, it is really difficult to process this data. Some of the research challenges faced are discussed below:

a) The number of clusters: Identifying the number of clusters is a difficult task if the number of class labels is not known beforehand. A careful analysis of number of clusters is necessary to produce correct results. Else, it is found that heterogeneous tuples may merge or similar type tuples may be broken into many.
b) High dimensionality**:** The number of features is very high and may even exceed the number of samples. So one has to face the curse of dimensionality [20].
c) Large number of samples: The number of samples to be processed is very high. Algorithms have to be very conscious of scaling issues. Like many interesting problems, clustering in general is NP-hard, and practical and successful data mining algorithms usually scale linear or log-linear. Quadratic and cubic scaling may also be allowable but a linear behavior is highly desirable.
d) Sparsity: Most features are zero for most samples, i.e. the object-feature matrix is sparse. This property strongly affects the measurements of similarity and the computational complexity.
e) The identification of distance measure: For numerical attributes, distance measures that can be usedare standard equations like Euclidian, manhattan, and maximum distance measure. All the three are special cases of Minkowski distance. But identification of measure for categorical attributes is difficult.
f) Significant outliers**:** Outliers may have significant importance. Finding these outliers is highly non-trivial, and removing them is not necessarily desirable.

VI. CONCLUSION AND FUTURE WORK

As an application to intrusion detection, we have clustered the NSL-KDD dataset, which is a modified version of KDD' 99 intrusion detection dataset, into two clusters (i.e. normal and anomaly) and also identified the corresponding cluster percentage. This work is done to label the data so that in future it can be used as class labels to correctly classify new instances. The clusters are formed according to the distance between data points and cluster centers are formed for each cluster. We have also done a comparative analysis of four clustering techniques (i.e. K-means, EM, Density Based Cluster, and Filtered cluster) and list out the challenges faced in this process. In this paper performance analysis of four clustering algorithms is carried out on NSL-KDD dataset. Results show that K-Means outperforms in time and accuracy to classify the dataset.

In future, this work can be extended to classify the dataset into four of its major attack types (DoS, R2L, Probe and U2R). To increase the accuracy in clustering process data reduction techniques can be applied along with hybrid of these algorithms. Ensemble of single models could be a better solution.

REFERENCES

[1]. J. T. Yao, S. L. Zhao, and L. V. Saxton, "A study on fuzzy intrusion detection," SPIE: Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security, vol. 58,no. 12, pp. 23-30, 2005.
[2]. J. S. R. Jang, C. T. Sun, and E. Mizutani, "Neuro-Fuzzy and Soft Computing-A Computational Approach to Learning and Machine Intelligence [Book Review]," Automatic Control, IEEE Transactions on, vol. 42,no. 10, pp. 1482-1484, 1997.
[3]. Z. Muda, W. Yassin, M. N. Sulaiman, and N. I. Udzir, "Intrusion detection based on K-Means clustering and Naive Bayes classification," in Information Technology in Asia (CITA 11), 7th International Conference on, Kuching, Sarawak, 2011, pp.1-6.
[4]. H. Om and A. Kundu, "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system," in Recent Advances in Information Technology (RAIT), 1st International Conference on, Dhanbad, 2012, pp. 131-136.
[5]. G. V. Nadiammai and M.Hemalatha, "An Evaluation of Clustering Technique over Intrusion Detection System," in International Conference on Advances in Computing, Communications and Informatics (ICACCI'12) Chennai, 2012, pp. 1054-1060.
[6]. Q. Ye, X. Wu, and G. Huang, "An intrusion detection approach based on data mining," in Future Computer and Communication (ICFCC), 2nd International Conference on, Wuhan, 2010, pp. V1-695-V1-698.
[7]. M. J. Haque, K. W. Magld, and N. Hundewale, "An intelligent approach for Intrusion Detection based on data mining techniques," in Multimedia Computing and Systems (ICMCS), International Conference on, Tangier, 2012, pp. 12-16.
[8]. P. Poonam and M. Dutta, "Performance Analysis of Clustering Methods for Outlier Detection," in Advanced Computing & Communication Technologies (ACCT), Second International Conference on, Rohtak, Haryana, 2012, pp. 89-95.
[9]. D. K. Denatious and A. John, "Survey on data mining techniques to enhance intrusion detection," in Computer Communication and Informatics (ICCCI), International Conference on, Coimbatore, 2012, pp. 1-5.

[10]. L. Wenke and J. S. Salvatore, "A framework for constructing features and models for intrusion detection systems," ACM Trans. Inf. Syst. Secur., vol. 3, no. 4, pp. 227-261, 2000.
[11]. P. Dokas, L. Ertoz, V. Kumar, A. Lazarevic, J. Srivastava, and P.-N.Tan, "Data Mining for Network Intrusion Detection," MINDS, pp. 21-30, 2002.
[12]. J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967, pp. 281-297.
[13]. T. Velmurugan and T. Santhanam, "Performance Evaluation of K-Means and Fuzzy C-Means Clustering Algorithms for Statistical Distributions of Input Data Points," European Journal of Scientific Research, vol. 46,no. 3, pp. 320-330, 2010.
[14]. S. Borah and M. K. Ghose, "Performance analysis of AIM-K-Means and K-Means in quality cluster generation," Journal of Computing, vol. 1,no. 1, 2009.
[15]. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society.Series B (Methodological), vol. 39,no. 1, pp. 1-38, 1977.
[16]. Expectation–maximization algorithm [Available Online] http://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm.
[17]. H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based Clustering," WIREs Data Mining and Knowledge Discovery vol. 1,no. 3, pp. 231-240, 2011.
[18]. Fisher and H. Douglas, "Knowledge acquisition via incremental conceptual clustering," Machine Learning, vol. 2,no. 2, pp. 139-172, 1987.
[19]. WEKA – Data Mining Machine Learning Software, [Available Online] http://www.cs.waikato.ac.nz/ml/weka/.
[20]. NSL-KDD dataset, [Available Online] http://iscx.ca/NSL-KDD/