



Data Mining Approach IDS K-Mean using Weka Environment

Richa *

Saurabh Mittal

Research Scholar Galxy Global Imperial Tech.Campus
Department of Computer Sc & Engineering
Kurukshetra University
India

Associate Professor Galxy Global Imperial Tech.Campus
Department of Computer Sc & Engineering
Kurukshetra University
India

Abstract— Intrusions detections systems from position of analysis of security procedure are a second line of defense; they have a decision-making role to observe the activities of our network or hosts to identify attacks in real time. In our days, electronics attacks can reason a very destructive damage for nations which make necessary the use of completed security policy to minimize the prospective threats. IDS it is a very main factor to defend against this exposure, in our works, we use a wired data base Knowledge Discovery Data Mining (KDD) CUP 10Percent and a Data Mining Tools Waikato Environment for Knowledge Analysis (WEKA) we check the results by using a several evaluations parameters. The results illustrate that a very high detection rate for certain attacks types

Keywords— Clustring, K-Mean, KDD Cup, Weka Environment

I. INTRODUCTION

An Intrusion Detection System (IDS) [1] is a defense system that plays an important role to keep or secure a network system and its main goal is to view network activities automatically to identify malicious attacks. Intrusion detection system (IDS) is increasingly becoming a vital and critical component to secure the network in today’s world of Internet.

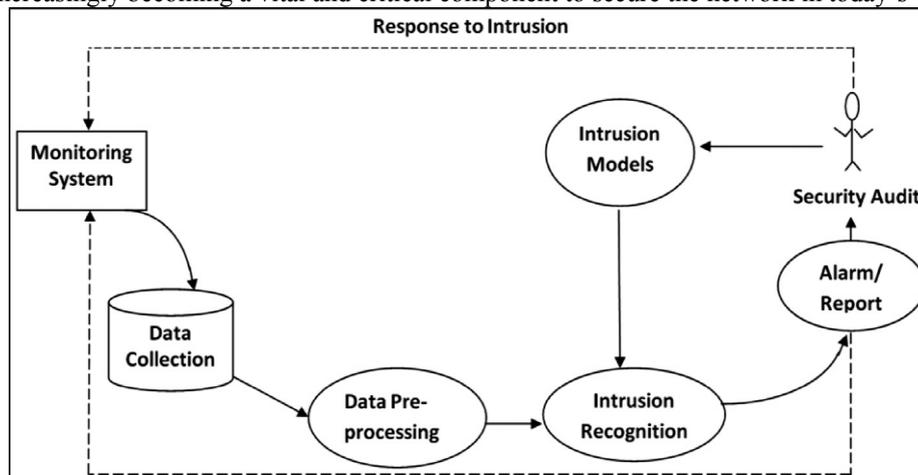


Fig1: Structure of IDS

Data mining can improve variants detection rate, manage false alarm rate, and decrease false dismissals. Data mining based on intrusion detection systems can be generally categorize into major two groups :misuse detection and anomaly detection. Network intrusion detection is the procedure of monitor the events occurring in a computing system or network and analyzing them for signs of intrusions, distinct as attempt to compromise the confidentiality. The intrusion attacks can be divided into four categories: Probe (e.g. IP sweep, vulnerability scanning), denial of service (DoS) (e.g. mail bomb, UDP storm), user-to-root (U2R) (e.g. buffer overflow attacks, root kits) and remote-to-local (R2L) (e.g. password guessing, worm attack).

Data mining based IDS can efficiently recognize these data of user significance and also predicts the results that can be utilize in the prospect. Data mining or knowledge discovery in databases has gain a large deal of awareness in IT industry as well as in the society. Data mining has been concerned to evaluate the useful information from great volumes of data that are noisy, fuzzy and dynamic. Fig. 1 illustrates the overall architecture of IDS. It has been to be found centrally to capture all the incoming packets that are transmitted over the network. Data are collectively and send for pre-processing to eliminate the noise; irrelevant and missing attributes are replaced. Then the preprocessed data are analyzed and classified according to their severity measures. If the record is ordinary, then it does not require any more change or else it send for report generation to raise alarms. base on the position of the data, alarms are raise to make the supervisor to handle the state in advance. The attack is modeled so as to enable the classification of set-up data. All the above method continues as soon as the transmission starts.

Intrusion detection (ID) is a type of protection management system for computers and networks. An ID system gather and analyze in order from various area within a computer or a network to identify possible security breaches, which contain both intrusions (attacks from outside the organization) and misuse (attacks from within the organization). ID uses *vulnerability estimation* (sometimes referred to as *scanning*), which is a technology developed to review the security of a computer system or network.

Intrusion Detection Systems are divided into two types [2,3,4] according to the detection approaches: Misuse Detection and Anomaly Detection.

A. Misuse detection

Misuse detection first build prototype for malicious behavior and then identify intrusion based on this known pattern i.e. it finds intrusions by look for activity corresponding to known techniques for intrusions. The main advantage of misuse detection is its higher detection accuracy to all known attack. The drawback of this approach is that it can only identify intrusions that follow predefined patterns.

B. Anomaly detection

Anomaly detection defines the probable behavior of the network or profile in advance. Any important deviations from such defined expected behavior are reported as possible attacks. But not all such deviation are attacks. The main improvement of this approach is that it can inspect unknown and more complicated intrusions. The shortcoming of this approach is its low detection rate and high false alarm rate.

II. CLUSTERING

Clustering techniques [8] can be helpful for detect intrusions from network data, because clustering methods can determine composite intrusions over a different time period. Clustering is an unsupervised machine learning method for discovering pattern and deals with unlabeled data with many dimensions. It is the process of passing on the data into groups of similar objects and each group is called as cluster. Each group consists of member from the same cluster that are similar and members from the different clusters are different from each additional. Anomaly based IDS have the facility to detect new attacks, as any attack will modify from the standard behavior. In order to detect attacks, a number of clustering based detection methods has been proposed. K-means [10] is one of the straightforward portioning algorithms that solve the clustering problem. The process of K-means algorithm follows a very easy and easy way to classify a given data set through a certain number of k clusters that are fixed a prior.

III. KDD CUP 99 Data Set

The function of Intrusion Detection Systems (IDSs), as special-purpose devices to detect anomalies and attacks in the network, is attractive more important. The examine in the intrusion detection field has been mostly focused on anomaly-based and misuse-based detection techniques for a long time. The first significant deficiency in the KDD data set is the huge number of redundant records. analyze KDD train and test sets, we found that about 78% and 75% of the records are duplicate in the train and test set, respectively. This big amount of unnecessary records in the train set will cause learning algorithms to be biased towards the more common report, and thus prevent it from learning unfrequent records which are usually more harmful to networks such as U2R attacks.

KDD CUP 99 DATA SET DESCRIPTION:-

Since 1999, KDD'99 [3] has been widely used data set for the estimate of anomaly detection methods is built based on the data captured in DARPA'98 IDS evaluation program.

The replicated attacks fall in one of the following four categories:

- 1) Denial of Service Attack (DoS): is an attack in which the invader makes some computing or memory resource too busy or too full to feel legal re-quests, or denies legal users access to a machine.
- 2) User to Root Attack (U2R): is a class of develop in which the attacker starts out with contact to a normal user account on the system (perhaps gain by sniff passwords, a vocabulary attack, or common engineering) and is able to develop some vulnerability to gain root access to the system.
- 3) Remote to Local Attack (R2L): occur when an attacker who has the facility to send packets to a mechanism over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.
- 4) Probing Attack attempt to gather in order about a network of computers for the apparent purpose of circumventing its security controls.

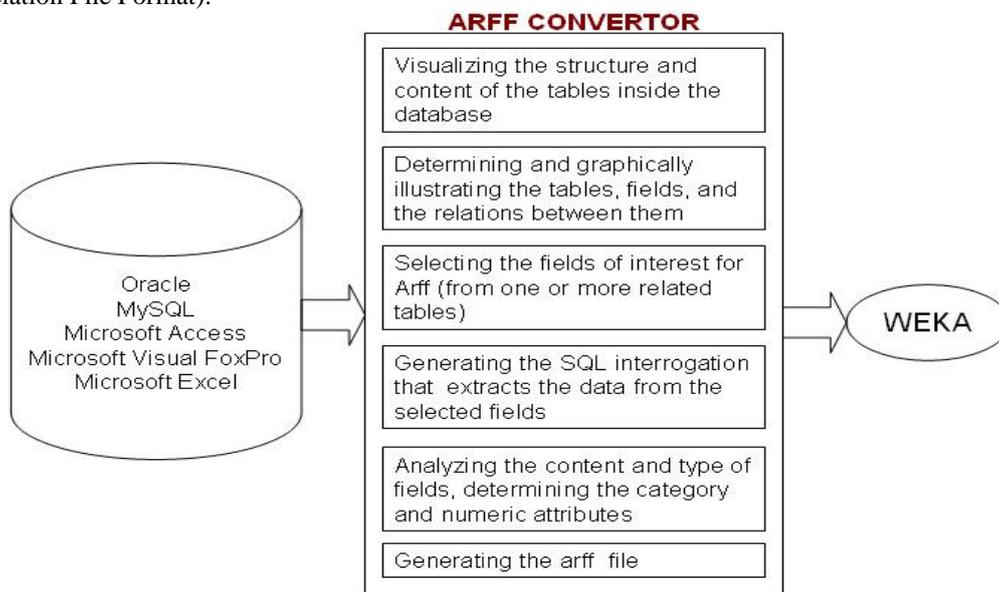
The datasets enclose a total number of 24 training attack type, with an extra 14 types in the test data only. KDD'99 features can be classified into three groups:

- 1) **Basic features:** this category encapsulate all the attribute that can be extract from a TCP/IP connection. Most of these features important to an implicit delay in detection.
- 2) **Traffic features:** this group includes features that are computed with respect to a window interval and is divided into two groups:

- a) "Same host" features: inspect only the connections in the past 2 seconds that have the same target host as the existing relation, and calculate statistics related to protocol behavior, service, et
- b) "same service" features: inspect only the connections in the past 2 seconds that have the same service as the existing connection.

IV. WEKA Environment

WEKA is an open-source mechanism knowledge data mining software, developed in Java by the Waikato University from New Zealand. The first intern description of WEKA was begin in 1994, and its first open version, that is version 2.1 was launched on the market in 1996. At there, the last constant version is version 3. WEKA is actually helpful software for the educational system, research and applications [1]. WEKA version 3.6 offers 49 preprocessing instruments (discreteness, reduction of noise, selection of attributes, etc.), 79 organization and regression algorithms (among which one can find J48, NaïveBayes[2], Random Forest), 8 clustering algorithms (such as SimpleKMeans, XMeans), 3 algorithms use for finding connection rules, including the Apriori algorithm and 3 graphic interface: the Explorer, Experimenter and information Flow. WEKA was downloaded additional than 1,4 million times, since was positioned on Source -Forge, in April, 2000 [3]. WEKA's short introductory production described above is destined to highlight the fact that WEKA is a powerful instrument for the exploratory analysis of data. Arff files are WEKA associated data files (Attribute-Relation File Format).



In practice there are frequent situations in which we dispose of related databases (Oracle, MySQL, Microsoft Access, Microsoft SQL Server, Microsoft Fox Pro, etc) which we want to analyze through the data mining techniques offered by WEKA. Loading the data from these databases into WEKA, as well as saving these data in arff format, is fairly difficult and requires knowledge of the SQL language.

V. PURPOSED METHOD

We have used KDD Trains 10 Percent data set available for research on network intrusion detection. This data set of the five million connection records was used as the data set for the 1999 KDD intrusion detection contest and is called the KDD Cup 99 data.,

We Analyze with help of Weka Platform. Weka is real world platform for Data Mininig.

Algorithm:

Step 1: initialize clustering P .

Step 2: Take data set;

Step 3: if P is null, build a new cluster centered on d, and add it to P. Go to Step 7;

Step 4: find a cluster C_j from P, which is the closest to d among all created clusters.

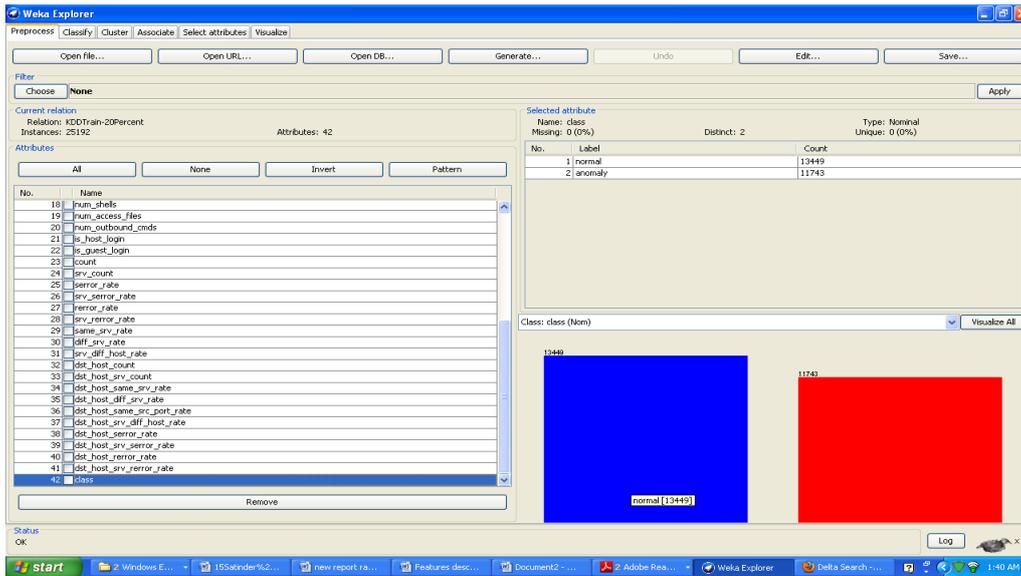
Step 5: if $\text{dist}(C_k, d) \leq Q$, add d to C_k. Go to Step 7.

Step 6: else, build a new cluster centered on d, add it to P;

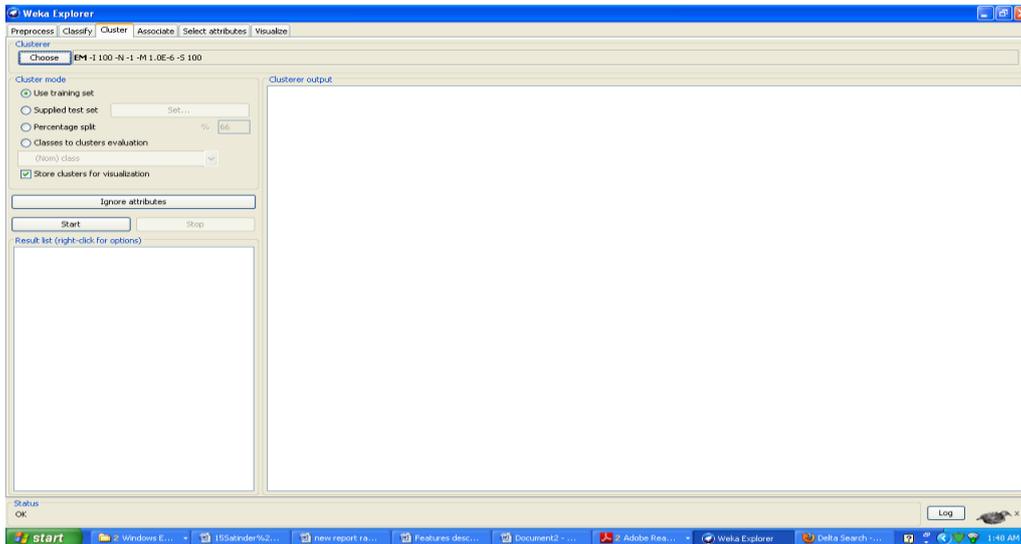
Step 7: repeat (2) (3) until all vectors of data set are processed.

Now ,We have to analyzed data set using Weka tool.

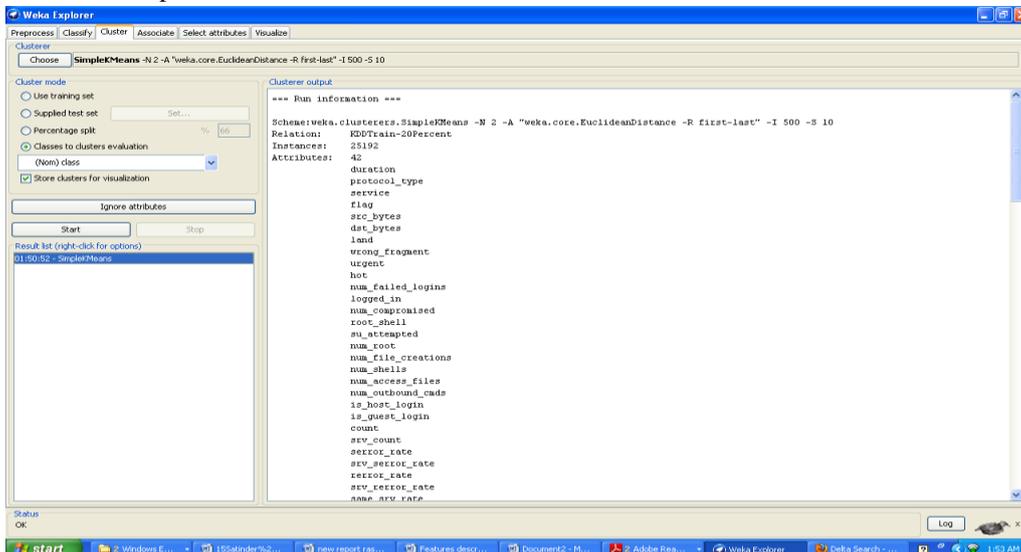
Step1 Weka3.6.2 is used for kdd Train 10 percent weka explorer with dataset loaded .We used K-mean Clustering algorithm K=2,Normal and Anomaly.



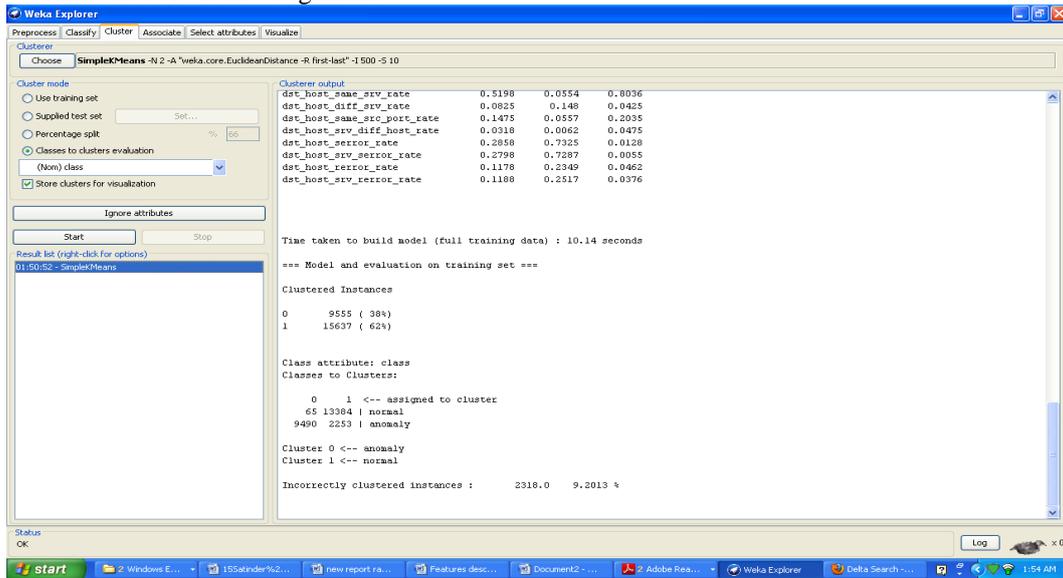
Step2 Now we find out Centroid, We select weka 3.6.2 in cluster.



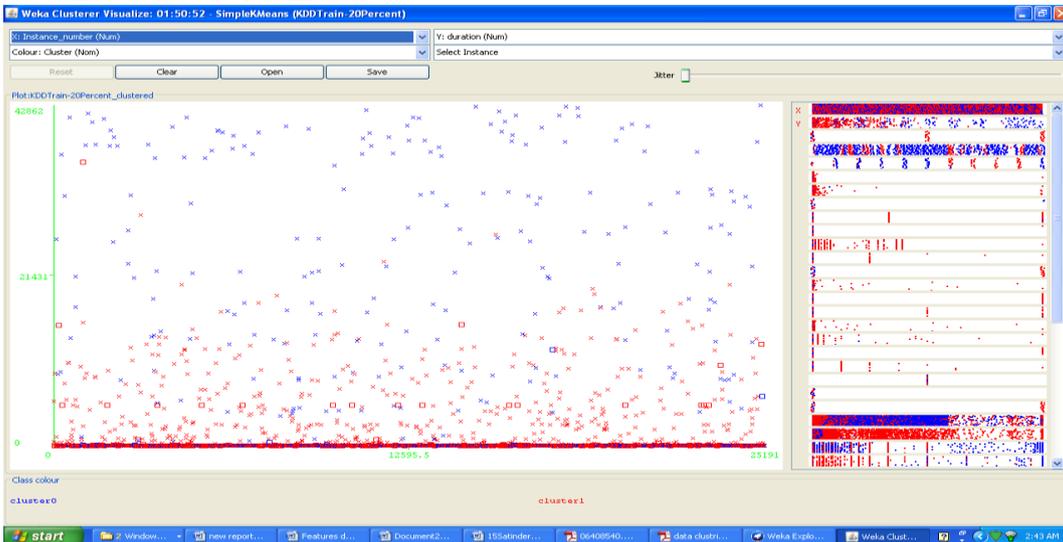
Step3: Now we select classes to cluster evaluation. Then we choose k-mean and we find out accuracy. We use 42 Attribute in KDD Train 10 percent



Step 4: Now we find cluster visual assignment



Step-5 Now we find out cluster visualization



Standard measures which were developed for evaluating IDSs include detection rate (DTR), false positive rate (FPR), and overall accuracy (OA). These three performance metrics may be defined as follows [21]:

$$DTR = \frac{TP}{TP+FN} \times 100\%$$

$$FPR = \frac{FP}{TN+FP} \times 100\%$$

$$OA = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

The Measurement used for Evaluation TP, TN, FP, and FN .

TP- True Positive indicate the number of attack record that are correctly classified.

TN- True Negative indicate the number of valid record that are correctly classified.

FP- False Positive indicates record that are incorrectly classified as attack.

FN- False Negative indicates record that are incorrectly classified as valid activities where in fact they are attack.

Accuracy:

The accuracy (AC) is the proportion of the total number of predictions that were accurate.

- P is the number of correct predictions that an instance is negative,
- Q is the number of incorrect predictions that an instance is positive,

- R is the number of incorrect of predictions that an instance negative, and
- S is the number of correct predictions that an instance is positive.

$$\begin{aligned} \text{acc} &= (P+Q)/(P+Q+R+S) \\ &= (65+2253)/(65+2253+9490+13384)=9.2013\% \\ \text{TP} &= S/R+S \\ 13384/9490+13384 &= 58\% \\ \text{FP} &= Q/P+Q \\ 2253/65+2253 &= 97\% \\ \text{TN} &= P/P+Q \\ 65/65+2253 &= 2\% \\ \text{FN} &= R/R+S \\ 9490/9490+13384 &= 41\% \end{aligned}$$

VI. CONCLUSIONS

The proposed method described in section v aims to achieve high accuracy, high detection rate and very low or no false alarm rate. This segment discusses the limitations of previous existing methods and advantages of proposed scheme over them. With the plan to improve detection rate and decrease false alarm rate, this paper presents a KDD 10 percent approach for intrusion detection system. Feature range helps in selecting important and relevant features from the data set and reduces the time required to process the data set.

REFERENCES

- [1] Mouaad KEZIH*, Mahmoud TAIBI” **Evaluation Effectiveness of Intrusion Detection System with Reduced Dimension Using Data Mining Classification Tools**” 2013 2nd International Conference on Systems and Computer Science (ICSCS) Villeneuve d’Ascq, France, August 26-27, 2013.
- [2] A.M Chandrasekhar “**Intrusion Detection Technique Using K-Mean Fuzzy Neural Network and SVM Classifier**” International Conference on Computer Communication and Informatics Jan4-6-2013 India.
- [3] Kapil Wankhade, Sadia Patka, Ravindra Thool” **An Efficient Approach for Intrusion Detection Using Data Mining Methods**” Volume 2 Issue 4 IEEE 2013
- [4] G.V. Nadiammal, M. Hemalatha” **Effective approach toward Intrusion Detection System using data mining techniques**” Elsevier 2013
- [4] T.R. Gopalakrishnan Nair, K.Lakshmi Madhuri” **DATA MINING USING HIERARCHICAL VIRTUAL K-MEANS APPROACH INTEGRATING DATA FRAGMENTS IN CLOUD COMPUTING ENVIRONMENT**” IEEE 2011
- [5] Feixiang Huang, Shengyong Wang, and Chien-Chung Chan” **Predicting Disease By Using Data Mining Based on Healthcare Information System**” 2012 IEEE International Conference on Granular Computing.
- [6] Cheung-Leung Lui, Tak-Chung Fu, Ting-Yee Cheung” **Agent-based Network Intrusion Detection System Using Data Mining Approaches**” Proceedings of the Third International Conference on Information Technology and Applications (ICITA’05) 2005 IEEE.
- [7] R. Robu* and V. Stoicu-Tivadar” **Arff Converter Tool for WEKA Data Mining Software**” 2010 IEEE.
- [8] Abhay Kumar, Rannish Sinha, Daya Shankar Verma” **Modeling using K-Means Clustering Algorithm**” 1st Int’l Conf. on Recent Advances in Information Technology | RAIT-2012 | 2012 IEEE.
- [9] Safwan Mawlood Hussein, Fakariah Hani Mohd Ali, Zolidah Kasiran” **Evaluation Effectiveness of Hybrid IDS Using Snort with Naïve Bayes to Detect Attacks**” 2012 IEEE.
- [10] Muamer N. Mohammada, Norrozila Sulaimana, Osama Abdulkarim Muhsinb” **A Novel Intrusion Detection System by using Intelligent Data Mining in Weka Environment**”