A

MAJOR PROJECT REPORT

On

NETWORK INTRUSION DETECTION SYSTEM USING DATA MINING TECHNIQUES

Submitted in Partial fulfillment of the requirement
For the Degree Of
MASTER OF TECHNOLOGY
(SOFTWARE ENGINEERING)

Submitted By:
KAUSHAL KUMAR
ROLL NO- 2K12/SWE/14

Under the Guidance of:
Mrs. Abhilasha Sharma

DEPARTMENT OF COMPUTER ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
2012-2014

CERTIFICATE

DELHI TECHNOLOGICAL UNIVERSITY
(Govt. of National Capital Territory of Delhi)
BAWANA ROAD, DELHI-110042
Date: ____________________

This is to certify that the thesis entitled ***Network Intrusion Detection System using Data Mining Techniques*** submitted by *Kaushal Kumar (Roll Number: 2K12/SWE/14*), in partial fulfilment of the requirements for the award of degree of Master of Technology in Software Engineering, is a work carried out by him under my guidance.

Mrs. Abhilasha Sharma
Assistant Professor
Department of Computer Engineering
Delhi Technological University
Delhi

Signature……………………………….

ACKNOWLEDGEMENT

I take this opportunity to express my deepest gratitude and appreciation to all those who have helped me directly or indirectly towards the successful completion of this thesis.

Foremost, I would like to express my sincere gratitude to my guide ***Mrs. Abhilasha Sharma***, *Assistant Professor, Department of Computer Engineering, Delhi Technological University, Delhi* whose benevolent guidance, constant support, encouragement and valuable suggestions throughout the course of my work helped me successfully complete this thesis. Without her continuous support and interest, this thesis would not have been the same as presented here.

Besides my guide, I would like to thank the entire teaching and non-teaching staff in the Department of Computer Engineering, DTU for all their help during my course of work.

Kaushal Kumar
Master of Technology
(Software Engineering)
College Roll No. 2K12/SWE/14
Department of Computer Engineering
Delhi Technological University
Bawana Road, Delhi-110042

Signature………………………………

ABSTRACT

In present scenario intrusion in the network is very critical concern for integrity, confidentiality, liability and many other security related term of the network. Due to this need of better network intrusion detection system is very important and critical need. Every Intrusion Detection System has some special features and these are totally depending on methodology used for development and in other word, training phase. There are two important techniques used in development of Intrusion Detection System, clustering and classification. Accuracy of the system is highly depended on classification although need of clustering cannot be change or replaceable. So for achieving more accuracy, high detection rate, high performance and low false alarm rate, this paper is using combination of clustering and classification technique. In this paper better approach of combination of techniques used that overcome short-fall of previous approach used for implementation for Intrusion detection system. This paper is using Fuzzy C-Mean algorithm as a clustering technique and Support vector machine as classification technique. KDD cup dataset is used for training and testing purpose in order to evaluate performance, accuracy, detection rate, false alarm rate and other important parameter. This paper presents a detail comparative analysis between combination of Fuzzy C-Mean and Support vector machine and combination of K Means and Support vector machine. Experiments and analyses show that the new approach is better in increasing the detection rate and accuracy as well as in decreasing the false positive rate. In previous paper's related to intrusion detection system, all are based on detection of abnormal data's. But these systems are highly depend on type of attacks, so it is better approach to track normal data's because behaviour of normal data's do not change in any time duration. That gives some advantage to the system on the basis of reliability, performance and accuracy.

List of Figure(s)

List of Table(s)

Table of Contents

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

In today's world, our regular life is totally depends upon Internet. With the speedy involvement of the Internet in our everyday life, it is very important to make it secure, with the dramatically growth of internet and network technologies increase the number of attacks and threats. In early days, only static defense techniques such as virtual private network, firewall and data information encryption etc. are used for network security but they are not enough to secure network completely. So, there is a need of dynamic defence techniques. Extensive research has been done to ensure the security of the computer networks for more than 25 years. As a result, dynamic approach is introduced and known as intrusion detection systems (IDS), which is used for detect and stop network intrusions.

Security: Security means Degree of protection provided to information against danger, damage, loss and crime. Every security technique fell into at least one of the 3 elements: prevention, detection and response.

Network Security: Network security means Protection of data during their transmission. When we communicate, we are sharing information between two devices via some form of transmission medium. Network security has been one of the most important problems in Computer Network Management [12].

Goals of Security:

i) **Data Integrity:** It ensures that the information which is transmitted from the sender to receiver is not modified during its transmission until it reaches to the intended receiver. It maintains and assures the consistency and accuracy of the data during its transmission.

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

iii) **Data Availability:** It ensures that the required information is accessible and usable by the authorized users whenever they need it.

Some other goals are:

iv) **Authentication:** It ensures that the identity of a receiver is confirmed.

v) **Non-Repudiation:** It ensures that the denial of integrity and authenticity of information is not possible.

Intrusion: The most publicized threats on network traffic is considered as Intrusion. It is defined as any set of malicious actions that tries to compromise the security goals means integrity, confidentiality or availability of system.

Intrusion Detection: The process of analysing and monitoring the events occurring in network traffic in order to detect suspicious activities is known as Intrusion detection. In recent years, intrusion detection has emerged as an important field for network security.

Intrusion Detection System: Intrusion detection system acts an important role in Intrusion detection in computer and network systems. Intrusion Detection system is a combination of hardware and software which is used to detects intrusions in the network.

Why Data Mining Techniques are used?

In recent years, one of the well accepted techniques for intrusion detection is Data Mining techniques. Intrusion Detection System monitors all the events taking place in the network by gathering and analysing information from various areas within the network. In Intrusion Detection System, Data is collected from various sources such as network log data, host data etc. Since the network traffic is large, analysis of data is too hard. So this gives rise to the need of Intrusion Detection System along with Different Data Mining Techniques for detection of suspicious data. Data mining techniques such as Classification, Clustering and Association Rule easily extract the information from large dataset. To apply data mining techniques in intrusion detection, pre-processing is the first step to be done on the collected data. In pre-processing step, data is converted into a specific format for the mining process. Next, the specific formatted data is used for clustering and classification. The clustering model based on different partitioning algorithm, hierarchical algorithm, grid based algorithm and density based algorithm. The classification model can be Decision tree based, Rule based, Bayesian network based , Support vector machine based ,Fuzzy logic based or neural network based [2][3][8]. Data mining technique provides the guarantee that no intrusion will be missed while checking the real time data in the network, thus ensuring the accuracy and efficiency in the detection process. Data mining techniques also helps in intrusion prevention mechanisms. They can detect both known and previous unknown patterns of attacks.

1.2 INTRUSION DETECTION SYSTEM

One of the traditional technologies which are used for defences is firewall but it is a static defence's technique and manual passive system compared to Intrusion Detection System. Thus, IDS was deployed to improve the network security of organization. The concept of Intrusion detection system was firstly proposed by Denning (1987), to identify, detect and trace the intrusion. An Intrusion detection system is a combination of software and hardware, which are used for detecting malicious activities. It recognizes possible security breaches, which include intrusion from within and outside the organization and hence can detect the patterns of intrusions. Whenever any malicious activity is detected in the network traffic, the main objective of Intrusion Detection System is to alarm the system administrator. In general, Intrusion Detection System makes two assumptions about the data set used as input for intrusion detection as follows [32]:

i) The amount of normal data exceeds the abnormal or attack data quantitatively.

ii) The attack data differs from the normal data qualitatively.

Intrusion Detection System could collect information online from the network traffic. After gathering the information, Intrusion Detection System will analyse and monitor this information. So, Intrusion Detection System acts as the "second line of defence". Finally, it will provide the detecting results and warning message for system administrator. The detecting results could be either normal or attack behaviour. An ideal Intrusion Detection System has a 100% attack detection rate, means 100% accuracy along with a 0% false positive rate, but it is hard to achieve [7]. The main objective of Intrusion detection system is to detect illegal behaviours of the host or network. The goal of intrusion detection is to build a system which would automatically examine network activity and detect such attacks. Once an attack is detected, the system administrator could be informed and thus take corrective action.

Intrusion Detection System can also perform the following tasks.

- Keep an eye on the system and user activities.
- Verification of the system errors.
- Evaluating the integrity of systems and data files.
- Note down any abnormal behaviour make statically records.
- Recognition activity model mapping known attacks and alerts.

1.2.1 Classification of Intrusion Detection System

A) Classification According to Technique

Each suspicious activity 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. Accordance with analytical methods, the Intrusion Detection System can be divided into two categories, one is Misuse Detection or Signature based Detection, and the other is Anomaly Detection or behaviour based detection.

1) Misuse Detection

Misuse detection is signature based Intrusion detection system. Detection of intrusions based on a pattern or signature of the malicious activity. During development stage of model, Misuse detection sets up the attack behaviours based on known attack behaviours. The misuse detection is similar to antivirus software. The antivirus software compares the scanned data with known virus code. If system finds abnormal attributes, the virus is existence and removes it. Hence, misuse detection collects the known attack behaviours from attribute database [8]. If the attack behaviour is similar to the one in database, the misuse detection can defend it before the intruder destroys our system. It can be very helpful for known attack patterns. 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. Signature based systems, by definition, are very accurate on known attacks which are included in their signature database. Moreover, since signatures are associated with specific misuse behaviour, it is easy to determine the attack type. However, their detection capabilities are limited to those within signature database. As the new attacks are discovered, a signature database requires continuous updating to include the new attack signatures. For the known attack, it may report a detailed and accurate report; for the unknown attack, its function is limited, addition, the characteristic repository is renewed continually. This kind of detection is characterised by a low rate of false alarm and a high rate of missing report [9].

2) Anomaly Detection

It is different from Misuse Detection. It consists of building model from normal data which can be used to detect variations in the observed data from the normal model. 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 behaviour and looks for malicious behaviours. When user has misbehaviours, the system notifies users that has an intruder. However, if the attack is similar to the normal behaviours, it may not be detected. Moreover, it is difficult to associate deviations with specific attacks. As the users change their behaviours, normal behaviours should be redefined. It can detect both known and unknown attacks. Although it can detect unknown intrusions, rate of missing report is low and the rate of false alarm is high. The main drawback of anomaly detection is that the detection is depended on the latest attack models, so it can't identify new attack behaviours [2][9].

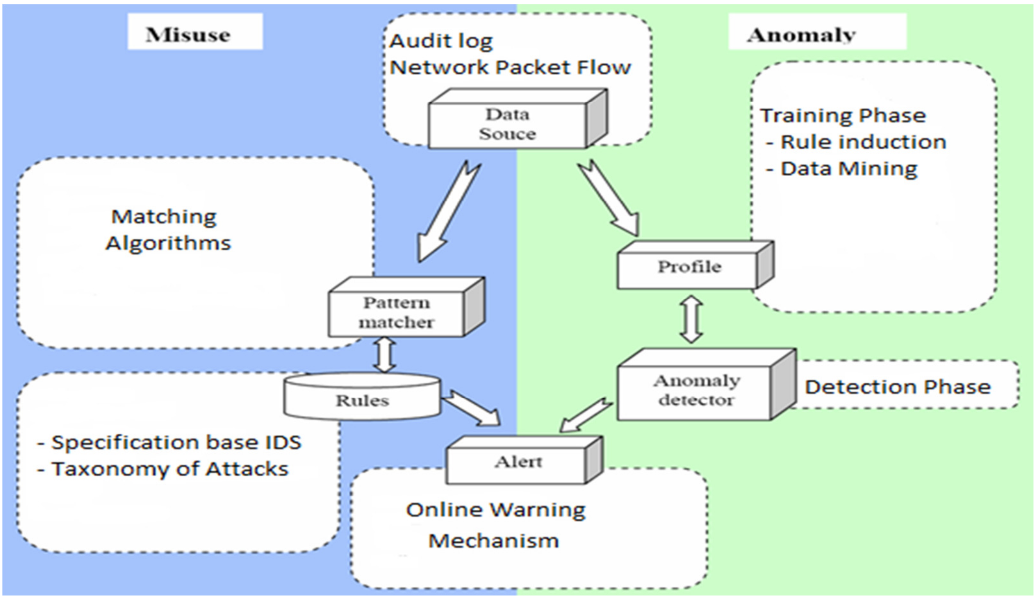


Fig 1.1: The Flow Chart of Misuse Detection and Anomaly Detection [2]

Comparison between Signature–Based and Behaviour–Based Intrusion Detection System

A signature based (misuse detection) IDS is used to monitor packets on the network and compares the packets against a library of signature or characteristics from known malicious threats. Behaviour – based IDS (anomaly detection system) analyses packets of data in the network at the very first instant. The main objective is to distinguish between normal and abnormal traffic examining the fundamental behaviour of a system [18].

Table1.1: Comparison between Misuse Detection and Anomaly Detection

Techniques	Advantages	Disadvantages
Signature – Based(Misuse Detection)	-Higher Detection rate, Accuracy for known behaviours. -Simplest and effective method. -Low False alarm rate.	- It can detect only known attacks. - Needs a regular update of the rules which are used. - Often no differentiation between an attack attempt and a successful attack. - Rate of Missing report is high.
Behaviour–Based(Anomaly Detection)	-can examine unknown and more complicated intrusions. - Rate of Missing report is low. -Detect new and unforeseen vulnerabilities.	- Needs to be trained and tuned model carefully, otherwise it tends to false – positives -low detection rate and high false alarm rate. - It can't identify new attacks because intrusion detection depends upon latest model.

A. Classification According to Architecture

For general purpose, the Intrusion Detection System (IDS) has classified into two major architectures.

1) Network-Based IDS

Network based IDS are best suited for alert generation of intrusion from outside the perimeter of the enterprise. The network based IDS are inserted at various points on LAN and observe packets traffic on the Network information is assembled into packets and transmitted on LAN or Internet. Network based IDS are valuable if they are placed just outside the firewalls, thereby alerting personals to incoming packets that might circumvent to the firewall. Some Network-Based IDS take or allows taking input of Custom signatures taken from user security policy which permits limited detection security policy violation. This limitation is due to packets traffic information that does not work well today in switched and encrypted environments where packets analysis is weak in detecting, attacking or originating from authorized Network users [10] [15] [25]. Network-Based Intrusion Detection Systems (IDS) use raw network packets as the data source. The IDS typically uses a network adapter in promiscuous mode that listens and analyses all traffic in real-time as it travels across the network. It is used to supervise and investigate network transfer to protect a system from network based threats. It tries to detect malicious activities such as denial-of-service (Dos) attacks and network traffic attacks. Network based IDS includes a number of sensors to monitors packet traffic, one or more servers for network management functions, and one or more management relieves for the human interface.

2) Host based IDS

It refers to intrusion detection that takes place on a single host system. It collects the data from each and every single host. It gets audit data from host audit trails and monitors activities such as integrity of system, file changes, host based network traffics, and system logs. If there is any unlawful change or movement is detected, it alerts the user by a pop-up menu and informs the central management server. Central management server blocks the movement or a combination of the above three. The judgment should be based on the strategy that is installed on the local system [12] [41]. Host-based Intrusion detection system places monitoring sensors also known as network resources nodes to monitor audit logs which are generated by Network operating system or application program. Audit logs contain records for events and activities taking place at individual Network resources. Because this Host-based Intrusion detection system can detect attacks that can't be seen by network based Intrusion detection system such as intrusions and can be misuse by trusted insider. Host based can be overcome the problem associated with network based Intrusion detection system immediately after alarming. It can also verify if any attack was unsuccessful, either because of immediate response to alarm or any other reason. It also maintains user login and user logoff action and all activity that generates audit records.

The Need for Both Types

As we can clearly see both network and host-based IDS solutions have unique strengths and benefits over one another and that is why the next generation IDS must evolve to include a tightly integrated host and network component. There are no Silver Bullets when it comes to network security but adding these two required components will greatly enhance our resistance to attack [21] [24].

Table1.2: Comparison between Network – based and Host – based IDS

Techniques	Advantages	Disadvantages
Network – Based IDS	-Detection of distributed attacks e.g. denial of service (DoS) - No impact on end system - Invisible configuration (stealth mode)	-High requirements on computing performance to scan every packet - Detection of attacks that manifest themselves in the network -Cannot be used if encrypted communication is allowed (unless a workaround like an SSL proxy is available)
Host – Based IDS	- Monitors the actual reaction of the host - Access to host - specific information e.g. integrity checker process or system call monitoring - Monitoring on all protocol layers- Encryption is no hindrance (except on the application layer	- installation on every single host - Adaptation to the different platforms and operating systems - Performance requirements on every supervised host - No detection of distributed attacks on multiple targets

1.2.2 Working of Intrusion Detection Systems

- **Data Acquisition:** Data is collected from network traffic using specific software and thus helps to get the information about the traffic like types of packets, hosts and protocol details,
- **Feature Selection:** Because of the huge network traffic, collected data is considerably large. So we generate feature vectors that contain only necessary information. In

network-based intrusion detection, it can be IP header information is treated as feature vector, which consists of packet type, source and destination IP address, protocol type and other flags.

- **Analysis:** In this step, the collected data is analysed to determine whether data is suspicious or not. Here uses various methods for Intrusion detection.
- **Action:** Intrusion Detection System alarms the system administrator that abnormal information has found and it tells about the behaviour. IDS also participate in controlling the attacks by closing the network port or killing the processes.

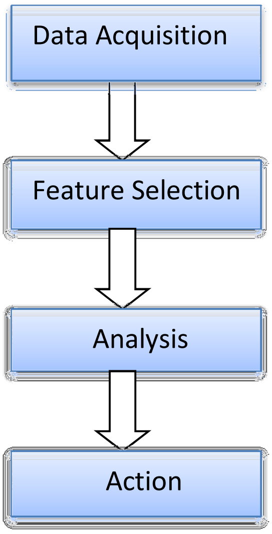


Fig 1.2: Working of Intrusion Detection System

1.2.3 Uses of an Intrusion Detection System

With increased usage and more cases of intrusion taking place now-a-days than ever before, we need something to enhance the security. This is where Intrusion detection system comes into the picture.

Following are the uses of Intrusion detection system [35]:

Backing up Firewalls: In many cases, intruders try to and penetrate firewalls to gain unauthorized access to corporate networks. This is done by attacking the firewall itself and breaking it down by tweaking its rules and signatures. In this case, the Intrusion detection system can decrease the risk of such attacks by temporarily backing up firewalls [14]. The Intrusion detection system of this type filters packets based on their IP packet header. This enables the network administrator to deploy Intrusion detection systems with functionality comparable to that of very advanced firewalls. Further, this type of Intrusion detection system can also be used while the general firewall is down for maintenance or when the firewall software is being updated or for any other reason.

Controlling File Access: Generally functions of controlling file access are done to specialized systems, such as Secret Net, which are intended specifically for protecting network information from unauthorized access. However, protection of some critically important files such as database files and password files cannot be done by such systems. Moreover, such systems are mainly developed for the Windows and NetWare platforms. So such systems fail in UNIX environments which are used for network applications in many organizations. So in such types of cases a Intrusion detection system comes to the rescue of network administrators. Mainly host based Intrusion detection systems are used in such cases which are based both on log-file analysis (Real Secure Server Sensor) and IDSs analysing system calls (Cisco IDS Host Server).

Controlling the Administrator's Activities: Intrusion detection systems can act as an additional control tool which can check unauthorized configuration changes by the hosts that have been granted administrative privileges.

Protection against Viruses: There has been an alarming increase in the number of viruses and worms that now invade the internet and affect numerous computers every day. Worm epidemics like the Red Code, Blue Code, Nimada etc. has demonstrated the danger of underestimating the dangers of such malicious programs.

Detecting unknown Devices: An Intrusion detection system can help in identifying the address of unknown/external hosts within the protected network segments. It can also detect increased traffic and special kind of activities from specific workstations which were not involved in such kind of activities before [15]. Such activities can be a hint to malicious activities from the hosts and the network administrator must be informed about this.

1.2.4 Performance Measurement of IDS

There are some primary factors which are used during performance measurement of Intrusion detection system [19].

- **False Positive (FP):** Or false alarm, Corresponds to the number of detected attacks but it is in fact normal.
- **False Negative (FN):** Corresponds to the number of detected normal instances but it is actually attack, in other words these attacks are the target of intrusion detection systems.
- **True Positive (TP):** Corresponds to the number of detected attacks and it is in fact attack.

• **True Negative (TN):** Corresponds to the number of detected normal instances and it is actually normal.

Performance of IDS is very important in order to ensure the capability of the system to detect intruders. Parameters those are used for performance analysis are detection rate, false alarm rate and accuracy.

Equation 1 shows how to calculate the percentage of detection rate:-

Detection Rate (DR) = (TP/TP+FN) x **100% ……………………………… (1)**

The number of false alarm rates can be obtained from the formula in Equation 2.

***False Alarm Rate (FAR) =FP/*Number of Attacks …………………………..(2)**

Then, the accuracy of the system is determined by Equation 3.

***Accuracy = (TP+TN/TP + TN + FP + FN)* x 100% …………………………..(3)**

1.3 ORGANISATION OF THESIS

This thesis consists of 5 chapters. The chapter deals with.

- **Chapter 1** introduces the theme of research work and introduction about Intrusion detection system.
- **Chapter 2** provides description about Data Mining based Intrusion Detection System.
- **Chapter 3** provides review of literature in order to create an adequate framework for conducting this research work.
- **Chapter 4** deals with the research methodology followed to achieve the goals of the work.
- **Chapter 5** provides the results of the thesis work, & discussion about the result.
- Conclusion and Future Work

CHAPTER 2
DATA MINING BASED INTRUSION DETECTION SYSTEM

2.1 INTRODUCTION

The term Data mining is the technique which is used for extracting useful information from the large amount of data and also used to classify, clean and examine large amount of data. Data mining technology is advanced because:

•It can deal with large amount of data.

•It can find out the hidden and mistreated information.

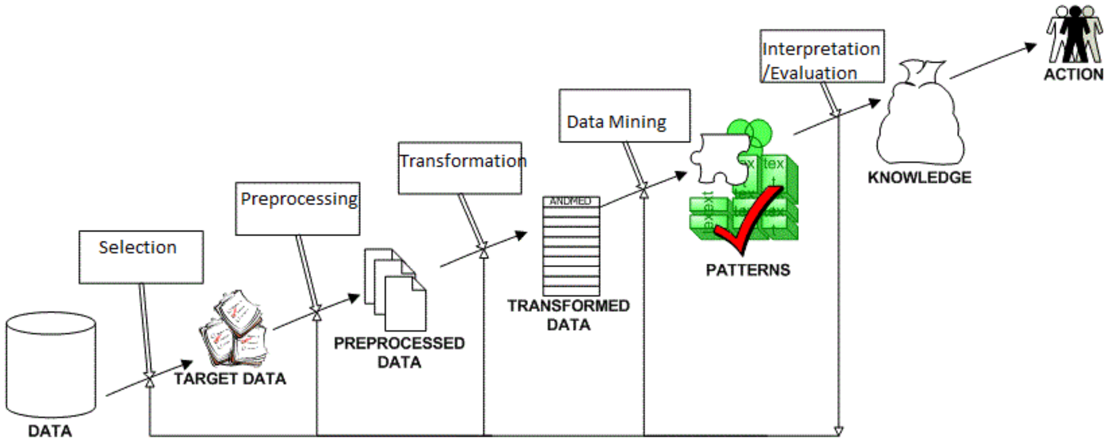


Fig 2.1: Data Mining [14]

Intrusion detection system is the region where data mining extensively concentrated. There are two reasons for this, first an IDS is very common and very popular and extremely critical activity, Second It deals with large volume of the data on the network. In intrusion detection system, information comes from various sources like host data, network log data, alarm messages etc. Since the variety of different data sources is too complex, the complexity of the operating system also increases. Also network traffic is huge, so the data analysis is too hard. The data mining technology have the capability of extracting large databases; it is of great importance to use data mining techniques in intrusion detection [37]. By applying data mining technology, intrusion detection system can widely verify the data to obtain a model, thus helps to obtain a comparison between the abnormal pattern and the normal behaviour pattern. Manual analysis is not required for this method [39]. The data mining technology has the huge advantage in the data extracting characteristic and the rule, so it is of great importance to use data mining technology in the intrusion detection. One of the main advantages is that same data mining tool can be applied to different data sources. Lee and Salvatore J. Stolfo, Columbia University was the two person who applied the data mining technology first time in the intrusion detection research area. An important problem in intrusion detection is how effectively can separate the attack patterns and normal data patterns from a large number of network data and how effectively generate automatic intrusion rules after collected raw network data. To achieve this, various data mining techniques are used such as classification, clustering, association rule mining etc [41].

2.2 ARCHITECTURE FOR IDS USING DATA MINING TECHNIQUES

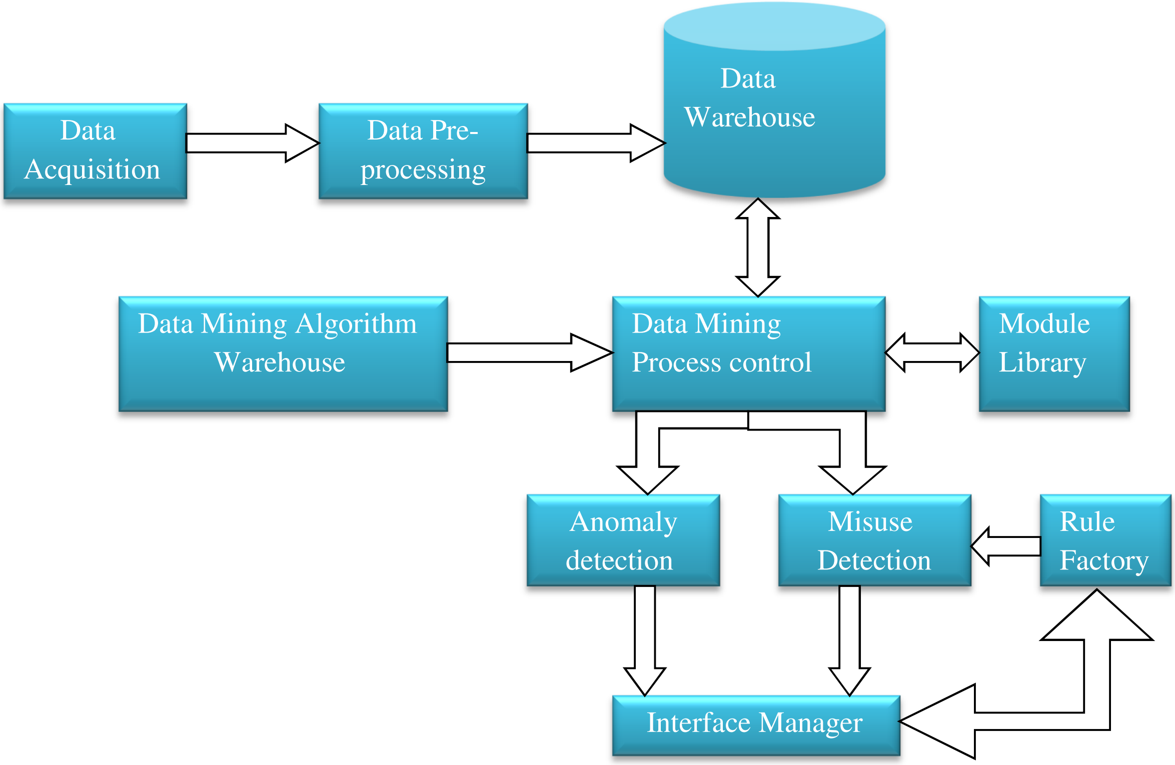


Fig 2.2: Architecture for Intrusion Detection System based on Data Mining [7]

Data Acquisition

Collect the data from various sources like Log, network, Syslog, command line, etc., which will be used for monitoring and analysis of the network [17].

Data Pre-process

Data packet or information is transformed into suitable forms using many methods such as data cleaning, data normalization, data integration, and standardized data analysis and data reduction techniques. By using data pre-processing, improve the accuracy and efficiency of data mining algorithms. Data pre-processing is important because the network data to be analysed by data mining techniques are [19]:

- Noisy: containing errors
- Incomplete: lacking of certain important attributes.
- Inconsistent: containing inconsistency between different information.
- Enhancing mining process: to enhance performance, reduce the number of data sets.
- Improve data quality: It can improve the data quality and helping to improve the efficiency and accuracy of the subsequent mining process. Data cleaning is a approach of data pre-processing which deals with the removal of imperfect, such as noise for continuous data attributes.

Data Mining Processor

The data which are obtained after data pre-processing is stored in Data warehouse. Data mining algorithm library module is a collection of a different Data mining algorithms. The efficiency of each techniques in the library of algorithms is improved by making sure that each technique has good time complexity, extensible, in order to initiate search process of data mining, and be used for forecast next time. Data mining process control is responsible for choosing a suitable mining algorithm from the data mining algorithm library. The traditional data mining intrusion detection system classify the data as normal data or suspicious behaviour, meaning for training data sets, it needs well labelled data. During the process of marking training data sets, the data mining control module may extract the features and detection characteristics from the library of algorithms. The mining control module can also be used when system starts running before training data sets is available, by calling data from a data warehouse and using clustering algorithm, the data is labelled as normal or abnormal data and returns data warehouse as a training data. Finally, results obtained from the mining process are carried to intrusion detection module to produce appropriate communication to the security officer or for the system to take a suitable response [7] [22] [34] [38].

Intrusion Detection Module

Rules Factory Keeps the rules that intrusion detection system needs for matching with data mining output modules. Anomaly Detection Module is responsible for generating a normal behaviour characteristic, comparing the rules from current network data streams with the rules of Rules Factory, if detected data is beyond the threshold that is malicious data, else normal data. Misuse detection module believes an intrusion has occurred if detected system behaviour matches records of Rules Factory.

Interface Manager

This module is responsible for produce decision on normal or abnormal pattern. If the decision reached is normal pattern, it adds it to close normal pattern in the regular library, but if the decision reached indicates abnormal pattern, adds it to close abnormal pattern in the regular library.

2.3 DATA MINING TECHNIQUE

Different data mining techniques like Classification, Clustering and Association rules are frequently used to acquire information about intrusions by observing network data.

2.3.1 Classification

Classification is analysis of data, which takes each and every instance of a dataset and assigns it to a specific class normal and abnormal. A classification based Intrusion Detection System will categorize all the information of network traffic into predetermined set either normal or abnormal class. Data classification consists of two steps – learning and classification. In learning step, a classifier is formed and this classifier model is used to predict the class labels for a given data in the classification step. To define the classes, Each and every record in the dataset already has value for the attribute used. The objective of a classifier is not to explore the data to discover different classes, but to find how new records should be arranged into classes [26]. Different types of classification techniques such as Decision tree induction, Genetic algorithm, Naive Bayesian networks, K-nearest neighbour classifier, Support vector machine, fuzzy logic etc. are used along with Intrusion Detection System. In the area of intrusion detection, classification technique is less efficient than clustering technique. Classification method can be effective for both anomaly detection and misuse detection, but it is frequently used for misuse detection.

Here we describe different classification algorithm which are used in Intrusion detection system.

A. Decision Tree

Decision tree is a classification technique in data mining for predictive models. Decision tree is a flowchart tree like structure where internal node represents a test on attribute, branch represents an outcome of the test and leaf node represents a class label. Initially decision tree is created by pre-classified data set. The important approach is, It uses divide and conquer method for splitting of data and divide the data items into their respective classes. It is a model that can be used to show possible results for particular occurrences in which the conditional probabilities are assigned for each occurrence [27]. Those occurrences of intrusions form a tree based structure that contains root node and a number of leaf nodes. Decision tree is very useful even for the large amount of data and it provides high accuracy. Nowadays, different enhanced version of Decision Tree algorithm is used for Intrusion detection.

1) ID3 Algorithm

ID3 is one of the famous Inductive Logic Programming methods, developed by Quinlan. It is basically an attribute based machine-learning algorithm that constructs a decision tree according to training set of data and an entropy measure to build the leaves of the tree. The informal formulation of ID3 is as follows [28]:

- Determine the attribute which has the highest information gain on the training set.
- Use this attribute as the root of the tree; create a branch for each of the values that the attribute can take.
- For each of the branches, repeat this process with the subset of the training set that is classified by this branch.

2) J48 Algorithm

J48 (enhanced version of C4.5) is based on the ID3 algorithm developed by Ross Quinlan, with additional features to address problems that ID3 was unable to deal. In practice, C4.5 uses one successful method for finding high accuracy hypotheses, based on pruning the rules issued from the tree constructed during the learning phase. However, the principal disadvantage of C4.5 rule sets is that it takes more CPU time and memory. Given a set S of cases, J48 first grows an initial tree using the divide-and-conquer algorithm as follows [29]:

- If all the cases in S belong to the same class or S is small, the tree is leaf labelled with the most frequent class in S.
- Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test as the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1,S2,… according to the outcome for each case, and apply the same procedure recursively to each subset.

3) Random Forest

Concept of Random Forest is introduced by Lepetit et. al. It is ensemble classification technique which consists of many decision trees. In random forest every tree is prepared by randomly select the data from dataset. By using Random forest, improve the accuracy and prediction power because it is less sensitive to outlier data. It can easily deals with high dimensional data [13].

Advantages

1. Construction of decision tree does not require any knowledge of the domain.
2. It can handle large data or high dimensional data.
3. Representation of decision tree is easy to understand.
4. It can handle both categorical and numerical data.

Disadvantages

1. Output attribute produce by decision tree must be categorical.
2. It is Limited to one output attribute.
3. Decision tree algorithms are volatile.
4. For numeric datasets, creation of tree can be complex.

B. Bayesian Method

It is a graphical model which contains the set of the random variables and their conditional dependencies; in which each node represents the random variable and the non-connected node represent the variables which are independent from each other. The major benefit of this technique is to deal with the incomplete data [30].

1) Naïve Bayes Algorithm

Naïve Bayes is an extension of Bayes theorem in that it assumes independence of attributes. The *Naïve Bayes* classifier is probabilistic classifier, it predict the class according to membership probability. Naive Bayes sometimes also called as idiot's Bayes, simple Bayes, and independence Bayes because to derive conditional probability, it analyses the relation between independent and dependent variable [29].

Bayes Theorem:

$$P\ (H/X) = P\ (X/H).\ P\ (H)\ /\ P(X)$$

Where, X is the data record, H is the hypothesis which represents data X, P(H) is prior probability, P(H/X) is the posterior probability of H conditioned on X and P(X/H) is the posterior probability of X conditioned on H.

Naive Bayes classifier is important for several reasons because it is easy to build without any complicated iterative parameter. This means it may be willingly applied to huge data sets. It is robust and easy to interpret in any particular application. In spite of this assumption, naive Bayesian classifiers give satisfactory results because focus is on identifying the classes for the instances, not the exact probabilities. Application like spam mail classification, text classification can use naïve Bayesian classifiers. The limitation is the requirement of the prior probabilities. The amount of probability information required is exponential in terms of number of attribute, number of classes and the maximum cardinality of attributes. With increase in number of classes or attributes, the space and computational complexity of Bayesian classifiers increases exponentially [18].

Advantages

1. The computations of Naïve Bayesian classifier are simple.
2. When it applied to large set of data, exhibit high speed and accuracy.

Disadvantages

1. The assumptions made in class conditional independence.
2. It is deficient in available probability data.

C. Neural Networks

Neural Network was usually used to refer a network or biological neurons. IDS neural network has been used for both misuse and anomaly intrusion detection. They can identify the arbitrary patterns in input data, and compare such patterns with an outcome, which can be a binary indication of whether an intrusion has occurred or not. The connections have weights that determine how one unit will affect other. Subset of such units act as input nodes, output nodes and remaining nodes constitute the hidden layer. By assigning activation to each of the input node and allowing them to propagate through the hidden layer nodes to the output nodes, neural network performs a functional mapping from input values to output values. The mapping is stored in terms of weight over connection. It has high degree of accuracy to recognize known suspicious events. Generally, it is used to learn complex nonlinear input output relationships [36].

Advantages

1. It requires less formal arithmetical training.
2. It can detect the nonlinear relationships between independent and dependent variables.
3. It can tolerate the noisy set of data.
4. Here multiple training algorithms are available.

Disadvantages

1. Computational burden is more.
2. Over fitting.
3. Training time is more.

D. Genetic Algorithms

Genetic algorithms such as Particle swarm optimization algorithm used in the ID. The application of GAs in IDS research started in early 1995 and involves evolving a signature that indicates intrusion. Another similar method is the Learning Classifier System in which binary rules are evolved, that collectively recognizes patterns of intrusion [37].

Advantages

1. It can easily solve every optimization dilemma.
2. It solves the problems with multiple results.
3. It can easily shift to existing models.

Disadvantages

1. Here, no global optimum.
2. It has no constant optimization response time.

E. K-Nearest Neighbour Algorithm

K-Nearest Neighbour (k-NN) is one of the simplest classification technique and a type of Lazy learning, it simply stores a given training tuple and waits until it is given a test tuple. It is an instance based learner that classifies the objects based on closet training data. For a given unknown tuple, a k-Nearest neighbour looks the pattern space for the k-training tuples that are closest to the unknown tuple. Here the object is classified by a majority vote of its neighbours. It calculates the distance between different points of input vectors and labelled it. Only In the case of K=1, the object is simply assigned to the class of its neighbour. When value of K is large then it takes more prediction time and influence the accuracy. This technique is computationally expensive and requires efficient storage for implementation of parallel hardware [39].

Advantages

1. It can analytically tractable.
2. Implementation of algorithm is simple.
3. Adaptive behaviour is high.
4. Parallel Implementations is easy.

Disadvantages

1. It requires high storage.
2. Highly susceptible to the curse of dimensionality.
3. Classification and testing task is slow.

2.3.2 Clustering

Human labelling is expensive and time-consuming in case of classification, because the available network information is too large. So clustering has attracted curiosity from researchers in the area of intrusion detection. Clustering is the process of labelling data and assigning into groups of similar objects without using known structure of data. Each group is called cluster, where members of the same clusters are utterly similar and members of the different clusters are dissimilar from each other. Hence clustering methods can be useful for classifying network data for detecting intrusions [14] [38] [45]. The main advantage of clustering algorithm is it can be able detect intrusions in the audit data without known signature of intrusion. There are two different methods for clustering-based intrusion detection. The first method is called unsupervised clustering, where the model of anomaly detection is trained using unlabelled data that consists of both attack as well as normal traffics. The second method is called semi-supervised clustering, where the intrusion detection model is trained using only normal activity. The main idea behind the first method is that amount of malicious or attack data is small as compared to total data. Based on this supposition, using cluster size easily detects anomalies and attacks, because large cluster is normal data cluster and the data which is not in large cluster correspond to attacks. Clustering algorithms can be categorized into four groups: hierarchical algorithm, partitioning algorithm, grid based algorithm and density-based algorithm. Clustering can be employed on both Misuse detection and Anomaly detection.

Here we describe different types of clustering techniques which are used in Intrusion detection system.

A. K-Means Clustering Algorithm

Concept of K-Means is firstly proposed by James Macqueen. In all machine learning algorithm K-Means clustering algorithm is one of the easiest and simplest clustering algorithm. In this clustering algorithm number of clusters (e.g. *k* clusters) is pre-defined, which is specified by the user. The first step is randomly select a set of k instances as centroids, which is centres of the k clusters, usually choose one for each cluster as far as possible from each other. Next, the algorithm scans each and every instance from the data set and assigns instances to the nearest cluster. Different methods are used to measure the distance between the centroid and instance but the most popular method is Euclidian distance measurement. The centroid of each clusters are always recalculated after every insertion of instance. This process is iterated until no more changes are made [9] [12].

The k-Means clustering algorithm is explained using following pseudo code.

```
Step 1:  Select the total number of clusters (k).
Step 2:  Randomly select k points for centroid of k clusters.
Step 3:  Next measure the distance from each data point to all centroids using Euclidean
         distance method.
Step 4:  Assign each and every instance to its closest centroid.
Step 5:  Recalculate the positions of all the centroids.
Step 6:  Repeat step 3 to 5 until no alteration in any centroids.
```

Advantages

1. It is easy to understand & implement and instances automatically assigned to clusters.
2. Interpretation is easy.
3. Relatively scalable and efficient in processing large data sets.

Disadvantages

1. When increases the number of data points to maximum, It takes maximum execution time.
2. It is sensitive to outlier and noise because if there is an object with a very large value, the data distribution may be biased or distorted.
3. All instances forced to be in a cluster.

B. K-Medoids Clustering Algorithm

Similar to K-Means, K-Medoids is also clustering by partitioning algorithm, which attempts to minimize the distance between data points and its centroid. The most centrally located instance or data point is considered as centroid in place of taking mean value. This centrally situated object is called medoid or reference point [42].

The pseudo code of k-Medoids is explained below:

Step 1: Input a data set D which consists of n objects.

Step 2: Input the number of clusters K.

Step 3: Randomly select k objects for initial cluster centres or medoids.

Step 4: Assign each object to the cluster with the nearest medoid.

Step 5: Calculate the total distance between the object and its cluster medoid.

Step 6: Swap the medoid with non-medoid object.

Step 7: Recalculate the positions of the k medoids

Step 8: Repeat step 4 to 7 until the medoids become fixed.

Advantages

1. This algorithm is more robust in the presence of noise and outliers because a medoid is less influenced by outliers.
2. It performs better when the number of data points increases to maximum.

Disadvantage

1. Its processing is more costly.

C. EM Clustering Algorithm

Expectation Maximization (EM) clustering is density based clustering algorithm. It is a modification of K-Means clustering algorithm and used for estimate the density of data points. In the EM clustering algorithm, we use an EM algorithm to find the parameters which maximize the probability of the data, assuming that the data is generated from k normal distributions. The algorithm gathers both the covariance and means of the normal distributions. This clustering method requires a number of inputs which are the total number of clusters, the data set, the maximum number of iteration and the maximum error tolerance. The EM can be separated into two steps which are Expectation (E-step) and Maximization (M-step). In E-step, calculate the expectation of the cluster probabilities for each data point in the dataset and then re-label the data points based on their probability estimations. In M-step, to re-estimate the parameters values from the E-step results. The outputs of M-step are then used as inputs for the following E-step. Until the result is not found, these two steps performed iteratively [21] [41].

2.3.3 Association Rule

The association rule considers each and every attribute/value pair as an item. Collection of items referred as an item set in a single network request. The algorithm searches to find an item set from large number of dataset that frequently appears in network. The main aim of association rule is to derive multi-feature correlations from a database table. Association rule mining determines association rules and/or correlation relationships among large set of data items. Association rule shows conditions for attribute values that occur frequently in the dataset. An example of association rule mining is Market Basket analysis. Association rules are obtained from the dataset and they are in the form of "if-then" statements. Apriori was the first scalable algorithm developed for association rule mining. Association rule mining in intrusion detection is very useful in many ways [27] [35] [44].

Basic steps for incorporating association rule for intrusion detection as follows:-

1. First network data need to be arranged into a database table where each row is an audit record and each column is a field of the audit records.
2. It is always shows that the intrusions and user activities shows frequent correlations among network data. Consistent behaviour's in the network data can be captured in association rules.
3. Also rules based on network data can continuously merge the rules from a new run to the aggregate rule set of all previous runs.
4. Thus with the association rule, we get the capability to capture behaviours in association rule for correctly detecting intrusions and hence lowering the false alarm rate.

CHAPTER 3
LITERATURE REVIEW

3.1 INTRODUCTION

A Literature Review is a body of text that aims to identify, evaluate, and synthesize the critical points of current knowledge and methodological contributions to a particular topic by the author. Literature reviews are secondary sources, and as such, do not report any new or original experimental work. It should give a theoretical basis for the research and helps to determine the nature of research. It identifies what is already known about an area of research.

For the present study, literature has been collected from different sources such as journal articles, Internet and conference proceeding paper.

3.2 REVIEW ON DATA MINING TECHNIQUE BASED IDS

Xiang M. Y. Chong et. al. (2004) [40], designed and proposed a multiple level tree classifier for intrusion detection which contains three-level of decision tree classification to increase detection rate. This model is more efficient in detecting known attacks but a serious shortcoming of this approach is the low detection rate for unknown attacks and generation of high false alarm rates.

Witchai Chimphlee et. al. (2005) [39], apply genetic algorithm and Fuzzy C-means methods in Intrusion detection system. Genetic algorithm identifies subset of features for network security and Fuzzy C-means is used to avoid a hard definition between normal class and certain intrusion class. Features selection methods aim at selecting a small or pre-specified number of features leading to the best possible performance of the entire classifier and reduce the error rate. The main goal of feature subset selection is to reduce the number of features used in classification while maintaining acceptable classification accuracy.

Fabrice Colas et. al. (2006) [7], observed different findings. They stated that Naïve Bayes is advantageous for a small number of samples, but as number of samples increases, the difference diminishes. SVM is, however, disadvantageous in terms of processing times. The processing time tends to grow quadratic ally with the number of samples in the training set.

Peddabachigiri S. et. al. (2007) [27], proposed a model of intrusion detection system using a hierarchical hybrid intelligent system combining decision tree and support vector machine (DTSVM) that produces high detection rate while reduces different attacks from normal behaviour .

Mrutyunjaya Panda et. al. (2007) [25] applies one of the efficient data mining algorithms called naïve bayes for anomaly based network intrusion detection. Experimental results on the KDD cup'99 data set show the novelty of approach in detecting network intrusion. It is observed that the proposed technique performs better in terms of false positive rate, cost, and computational time when applied to KDD'99 data sets compared to a back propagation neural network based approach. As a naïve Bayesian network is a restricted network that has only two layers and assumes complete independence between the information nodes.

R. Gaddan Shekhar et. al. (2007) [31], present a method to cascade k-Means clustering and the ID3 decision tree learning methods for classifying anomalous and normal activities in a computer network, an active electronic circuit, and a mechanical mass-beam system. The k-Means clustering method first partitions the training instances into k clusters using Euclidean distance similarity. On each cluster, representing a density region of normal or anomaly instances, build an ID3 decision tree. Results show that the detection accuracy of the K-Means with ID3 method is as high as 96.24 percent at a false-positive-rate of 0.03 percent, the total accuracy is as high as 80.01 percent.

Mrutyunjaya Panda et. al. (2008) [26] proposed a comparative study of data mining algorithm for network intrusion detection system. In this paper, effectiveness of three data mining classifier namely Decision tree with the ID3 & J48 algorithm and Naïve bayes algorithm are evaluated. Experimental results use the KDD Cup'99 online data set. Accuracy & performance of classification of Naïve bayes for all classes is better than the overall accuracy obtained in the case of different Decision tree algorithm. Decision tree is robust in detecting new intrusions in comparison to Naïve bayes classification algorithm.

QU Zhiming et. al. (2009) [29], proposed a model by using rough set and clustering algorithm in network security management. They uses k-means clustering algorithm. Rough set algorithm of classification is used for attribute reduction. It gives best results in terms of performance. It is fail for core attributes (if the matrix element only includes a single attribute) because it is not reduced.

Chunhua Gu et. al. (2009) [4], proposed an Intrusion Detection classifier using rough set and Support Vector Machine. Rough set is used for attribution reduction and support vector machine for intrusion detection classification.

M. Govindarajan et. al. (2009) [21], proposed new K-nearest neighbour classifier applied on Intrusion detection system and evaluate performance in term of Run time and Error rate on normal and malicious dataset. This new classifier is more accurate than existing K-nearest neighbour classifier.

Rung-Ching Chen et. al. (2009) [30], proposed a system for network intrusion detection using rough set and support vector machine. First, RST is used to pre-process the data and reduce the dimensions. Next, the features were selected by RST will be sent to SVM model to learn and test respectively.

Xiaohui Bao et. al.(2009) [41], proposed Network Intrusion Detection System combining Anomaly Intrusion Detection and Misuse Intrusion Detection, based on Support Vector Machine. Support Vector Machine is a classification method possessing better learning ability for small samples, which has-been widely applied in many fields such as web page identification and face identification. The technology of Support Vector Machine applied in intrusion detection possesses such advantages as high training rate and decision rate, insensitiveness to dimension of input data, continuous correction of various parameters with increase in training data which endows the system with self-learning ability, and so on.

T. Velmurugan et. al. (2010) [38], compute the complexity between k-means and k-medoids clustering algorithm for normal and uniform distribution of data points. They analysed the efficiency of k-Means and k-Medoids clustering algorithms by using large datasets in the cases of normal and uniform distribution; and found that the average time taken by k-Means algorithm is greater in both the cases. They further stated when the data points are increased to maximum, the k-Means algorithm takes maximum time.

Snehal A. Mulay et. al. (2010) [35], proposed an Intrusion Detection System using support vector machine and Decision tree. This model decreases the training and testing time & increasing the efficiency. This gives better result than individual models.

Mohammadreza Ektela et. al. (2010) [22], used Support Vector Machine and classification tree Data mining technique for intrusion detection in network. They compared C4.5 and Support Vector Machine by experimental result and found that C4.5 algorithm has better performance in term of detection rate and false alarm rate than SVM, but for U2R attack SVM performs better.

Guanghui Song et. al. (2011) [9], proposed an intrusion detection method based on multiple kernel support vector machine classifier. For using multiple kernel support vector machine classifier improve the accuracy and detection rate.

P. Amudha et.al.(2011) [28], observed that Random forest gives better detection rate, accuracy and false alarm rate for Probe and DOS attack & Naive Bayes Tree gives better performance in case of U2R and R2L attack. Also the execution time of Naive Bayes Tree is more as compared to other classifier.

Fatin Mohd Sabri et. al. (2011) [8], designed a hybrid of Rough set theory and Artificial immune recognition system as a solution to decrease false alarm rate. Rough set theory is expected to be able to reduce the redundant features from huge amount that are capable to increase the performance of the classification. AIS will minimize the duplications. Hybrid of AIS with Rough set capable to reduce the number of False alarm rate, but number of false negative detection is increases.

Amuthan Prabhakar Muniyandi et. al. (2011) [28], propose an anomaly detection method cascading k -Means clustering and the C4.5 decision tree methods for classifying anomalous and normal activities in a computer network. The k-Means clustering method is first used to partition the training instances into k clusters using Euclidean distance similarity. On each cluster, representing a density region of normal or anomaly instances, build decision trees using C4.5 decision tree algorithm. The decision tree on each cluster refines the decision boundaries by learning the subgroups within the cluster. The final conclusion is obtained by exploiting the results derived from the decision tree on each cluster. The proposed algorithm gives impressive detection accuracy in the experiment results.

Z. Muda et. al. (2011) [46], proposed hybrid approach of Intrusion detection based on k-Means clustering and Naive Bayes classification. They used KDD Cup '99 dataset as evaluation data and found out that their hybrid learning approach produced low false alarm rate below 0.5%, while keeping accuracy and detection rate higher to 99%. The approach was capable to correctly classify Normal data type, and also attack data types like Probe and DoS but not for U2R and R2L.

Roshan Chitrakar et. al. (2012) [31], proposed a hybrid approach to anomaly based intrusion detection by using k-Medoids clustering with Naïve Bayes classification and produced better performance compared to k-Means with Naïve Bayes classification. The approach, using Kyoto 2006+ datasets, showed around 4% of improvement in both Accuracy and Detection Rate while reducing False Alarm Rate by 1% as compared to the k- Means clustering approach followed by Naïve bayes classification technique. If the number of data points is less, then the K-Means algorithm takes lesser execution time. But when the data points are increased to maximum the K-Means algorithm takes maximum time and the K-Medoids algorithm performs reasonably better than the K-Means algorithm.

Deepthy k Denatious et. al. (2012) [6], describe different data mining techniques applied for detecting intrusions. Also describe the classification of Intrusion detection system and its working. For large amount of network traffics, clustering is more suitable than classification in the domain of intrusion detection because enormous amount of data needed to collect to use classification.

Om et. al. (2012) [2], propose a hybrid intrusion detection system that combines k-Means, and two classifiers: K-nearest neighbour 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.

Sanjay Kumar Sharma et. al. (2012) [35], improved network intrusion detection technique based on k-means clustering via naive bayes classification for anomaly based network intrusion detection. Compared to the Naive based approach, this approach achieve higher detection rate. However, it generates somewhat more false positives.

Manish Jain et. al. (2012) [19], improved technique based on naive bayes for intrusion detection system. The proposed algorithm achieved high detection rates (DR) and significant reduce false positives (FP) for different types of network intrusions using limited computational resources. Proposed Naïve Bayesian classifier gives higher detection rate and reduce false alarm. The main propose of this paper was to improve the performance of naïve Bayesian classifier for intrusion detection.

N. S. Chandolikar et. al. (2012) [21], proposed an efficient classification algorithm for Intrusion attacks. In this paper, comparison between Decision Tree, Rule induction & Bayesian belief network classification algorithm proposed. It uses the KDD Cup 99 data set for classification of intrusion attacks using data mining algorithm. Here analysis suggests that the true positive rate is high in Decision Tree algorithm rather than Rule based & Bayesian belief network algorithm. Decision Tree is most suitable for Intrusion attacks & it shows better performance.

Roshan Chitrakar et. al. (2012) [32], proposed a hybrid approach to anomaly based intrusion detection by using k-Medoids clustering with Support Vector Machine classification technique and produced better performance compared to k-Medoids with Naïve Bayes classification. The approach shows improvement in both Accuracy and Detection Rate while reducing False Alarm Rate as compared to the k- Medoids clustering approach followed by Naïve bayes classification technique.

Koc. L et al. (2012) [14] has proposed a network intrusion detection system based on a Hidden Naive Bayes multiclass classifier. Experimental results show that the HNB model exhibits a superior overall performance in terms of accuracy, error rate and misclassification cost compared with the traditional Naïve Bayes model and performed better than other models, such as SVM, in predictive accuracy.

A. M. Chandrasekhar et. al. (2013) [24], proposed a model using k-means, Fuzzy Neural Network and Support Vector Machine classifier for intrusion detection. With the help of K-means make large heterogeneous training data set into the number of homogenous subsets. So complexity of each subset reduces. They use KDD Cup 99 dataset for experiment.

Yogita B. Bhavsar et. al. (2013) [43], proposed Intrusion Detection System using support vector machine classification. Support vector machine is one of the most prominent classification algorithms in data mining area but extensive training time. Here they overcome the drawback of support vector machine.

Alka Shrivastava et. al. (2013) [24], proposed an Intrusion Detection Model based on SVM classification technique and K-means Clustering technique. It is not only capable of attack situation but can also classifying the individual attacks. K-means clustering filters the un-useful similar data points hence reduces the training time also hence provides an overall enhanced performance by reducing the training time. The Detection accuracy of the system is up to 90% which is excellent also the algorithm have very low FPR (max 8.3%) hence reduces the chances of false alarming. The results also shows that it takes only 0.0075 seconds to identify the intrusion hence fast enough to prevent any loss due to delayed action. Further it could achieve much better performance by increasing the number of samples taken and increasing the number of characteristics parameter selected.

Rachnakulhare et.al.(2013) [34], presents an intrusion detection system based on fuzzy C-means clustering and probabilistic neural network which not only reduces the training time but also increases the detection accuracy. The Detection accuracy of the system is up to 99% which is excellent also the algorithm have very low FPR (max 8.3%) hence reduces the chances of false alarming.

CHAPTER 4
RESEARCH METHODOLOGY

4.1 INTRODUCTION

In order to maintain the high accuracy, high detection rate and lower down the false alarm rate, we propose a Hybrid Intrusion Detection System which is combination of Fuzzy C-Means clustering technique and Support Vector Machine classification technique of Data mining.

By using Fuzzy C-Means clustering technique, Data points are divided into number of clusters of similar data instance based on membership value.

Next using Support Vector Machine classification technique, classify the resulting clusters into normal and abnormal classes. It is possible that some data instances are misclassified during FCM clustering, so by using SVM technique on resulting cluster data instances may be classified correctly.

So Proposed Hybrid Intrusion Detection System improves the accuracy & detection rate and lowers down the false alarm rate.

Next section describes the work flow of proposed model and also algorithm for proposed model.

4.2 PROPOSED WORK FLOW DIAGRAM

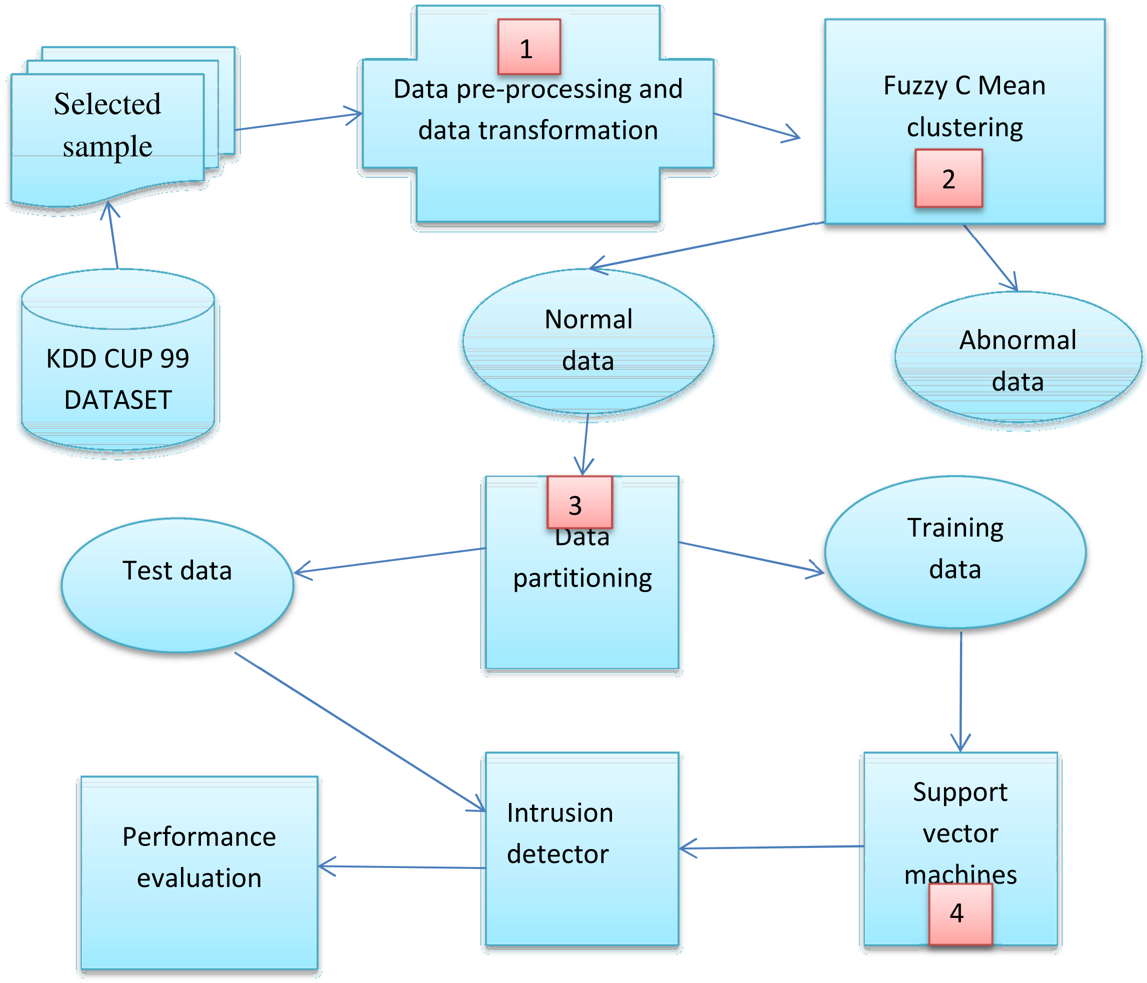


Fig 4.1: Work Flow Diagram of Proposed Algorithm

After Applying Fuzzy C-Means clustering technique, similar data instances are grouped based on their membership and divided into 2 clusters

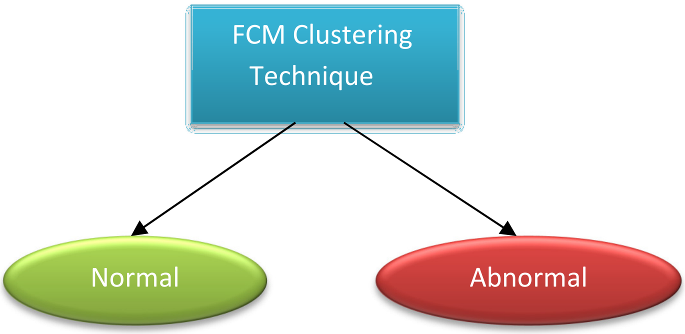


Fig 4.2: FCM Clustering

The resulting clusters are then classified into normal and abnormal classes using SVM classifiers.

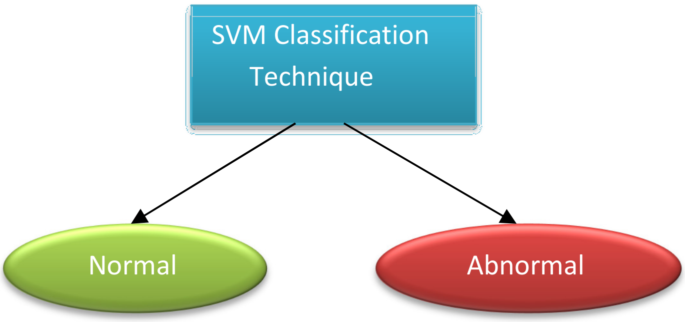


Fig 4.3: SVM Classification

4.3 PROPOSED ALGORITHM

Step 1: Collect the KDD Cup99 dataset online.

Step 2: Select the samples of data from KDD Cup99 dataset,

Step 3: In Pre-processing, Transform the selected features from the dataset for further processing.

Step 3: Cluster the dataset using the Fuzzy C-means Clustering method.

Step 4: Partition the data into training and testing sets.

Step 5: Now Apply Support Vector Machine Classification method on training set.

Step 7: Test the trained Intrusion detection model by the testing dataset.

Step 8: Evaluate the performance of the trained Intrusion detection system.

4.4 DATA COLLECTION

The data set used to perform the experiment is taken from KDD Cup ’99, which is widely accepted as a benchmark dataset .The data set was chosen to evaluate rules and to detect intrusion.

4.4.1 Description about KDD Cup99 Dataset

KDD dataset details:

- Training set - 5 million connections.
- 10% training set - 494,021 connections

- Test set - 311,029 connections

In KDD Cup '99 dataset the input data flow contains the details of the network connections, such as protocol type, connection duration, login type etc. The entire KDD Cup '99 data set contains 41 features and Connections are labelled as normal or attacks and Type C is continuous, while D is discrete.

Table 4.1: List of Features (KDD-CUP-99 task description)

Sl. No.	**Feature name**	**Description**	**Type**
1	Duration	Length (# of seconds) of the connection	C
2	protocol type	Type of the protocol, e.g. tcp, udp, etc.	D
3	service	Network service on the destination, e.g., http, telnet, etc.	D
4	flag	Normal or error status of the connection	D
5	src_bytes	# of data bytes from source to destination	C
6	dst_bytes	# of data bytes from destination to source	C
7	land	1 if connection is from/to the same host/port; 0 otherwise	D
8	wrong_fragment	# of "wrong" fragments	C
9	urgent	# of urgent packets	C
10	hot	# of "hot" indicators	C
11	num_failed_logins	# of failed login attempts	C
12	logged in	1 if successfully logged in; 0 otherwise	D
13	num_compromised	# of compromised conditions	C
14	root_shell	1 if root shell is obtained; 0 otherwise	D
15	su_attempted	1 if "su root" command attempted; 0 otherwise	D
16	num_root	# of "root" accesses	C
17	num_file_creations	# of file creation operations	C
18	num_shells	# of shell prompts	C
19	num_access_files	# of operations on access control files	C
20	num_outbound_cmds	# of outbound commands in an ftp	C

		session	
21	is_host_login	1 if the login belongs to the “hot” list; 0 otherwise	D
22	is_guest_login	1 if the login is a “guest’ login; 0 otherwise	D
23	count	# connections to the same host as the current one during past two seconds	C
24	srv_count	# of connections to the same service as the current connection in the past two seconds	C
25	serror_rate	% of connections that have “SYN” errors	C
26	srv_serror_rate	% of connections that have “SYN” errors	C
27	rerror_rate	% of connections that have “REJ” errors	C
28	srv_rerror_rate	% of connections that have “REJ” errors	C
29	same_srv_rate	% of connections to the same service	C
30	diff_srv_rate	% of connections to different services	C
31	srv_diff_host_rate	% of connections to different hosts	C
32	dst_host_count	# of destination host count	C
33	dst_host_srv_count	# of destination host services	C
33	dst_host_srv_count	# of destination host services	C
34	dst_host_same_srv_rate	% of connections of destination host to same services	C
35	dst_host_diff_srv_rate	% of connections of destination host to different services	C
36	dst_host_same_src_port_rate	% of connections of destination host to same service port	C
37	dst_host_srv_diff_host_rate	% of connections of destination host to different service port	C

38	dst_host_serror_rate	% of connections of destination host that have “SYN” errors	C
39	dst_host_srv_serror_rate	% of connections of destination host service that have “SYN” errors	C
40	dst_host_rerror_rate	% of connections of destination host that have “REJ” errors	C
41	dst_host_srv_rerror_rate	% of connections of destination host service that have “REJ” errors	C

Generally, there are four categories of attacks as follows [33]:

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.

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.

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.

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.

Test data has attack types that are not present in the training data .Problem is more realistic. Train set contains 22 attack types. Test data contains additional 17 new attack types that belong to one of four main categories.

Table 4.2: Different Attacks in KDD Cup’99 Dataset

Types	**Attacks in the Training Data**	**Attacks in the Testing Data**
Probe	Ipsweep, Nmap, portsweep, satan	Mscan, sain
DOS	Back, Land, Neptune, Pod, smurf, Teardrop	Apache2,Mailbomb,processtable,upstorm
U2R	Buffer_overflow, Loadmodule, Perl, Rootkit	Httptunnel, Ps, Worm, Xterm
R2L	Ftp_write, Guess_passwd, Imap,	Named, sendmail, snmpgetattack,

	Multihop, phf, spy, Warezclient, Warezmaster	snmpguess, sqlattack, xlock, xsnoop

4.5 DATA PREPROCESSING

Pre-processing of original KDD Cup'99dataset is necessary to make it as a suitable input for Hybrid Intrusion Detection System. Data set pre-processing can be achieved by applying [43]:

Data Set Transformation: The training dataset of KDD Cup'99 consist of approximately 4,900,000 single connection instances. Each connection instance contains 42 features including attacks or normal. From these labelled connection instances, we need to transform the nominal features to numeric values so as to make it suitable input for Hybrid Intrusion Detection System.

Data Set Normalization: Dataset normalization is essential to enhance the performance of intrusion detection system when datasets are too large.

Data Set Discretization: Dataset discretization technique is used for continuous features selection of intrusion detection and to create some homogeneity between values, which have different data types.

4.6 DATA MINING METHODS USED FOR DESIGNING OF HYBRID IDS

4.6.1 Fuzzy C-Means Clustering Algorithm

There are essentially two types of clustering methods: hierarchical algorithms and partitioning algorithms.

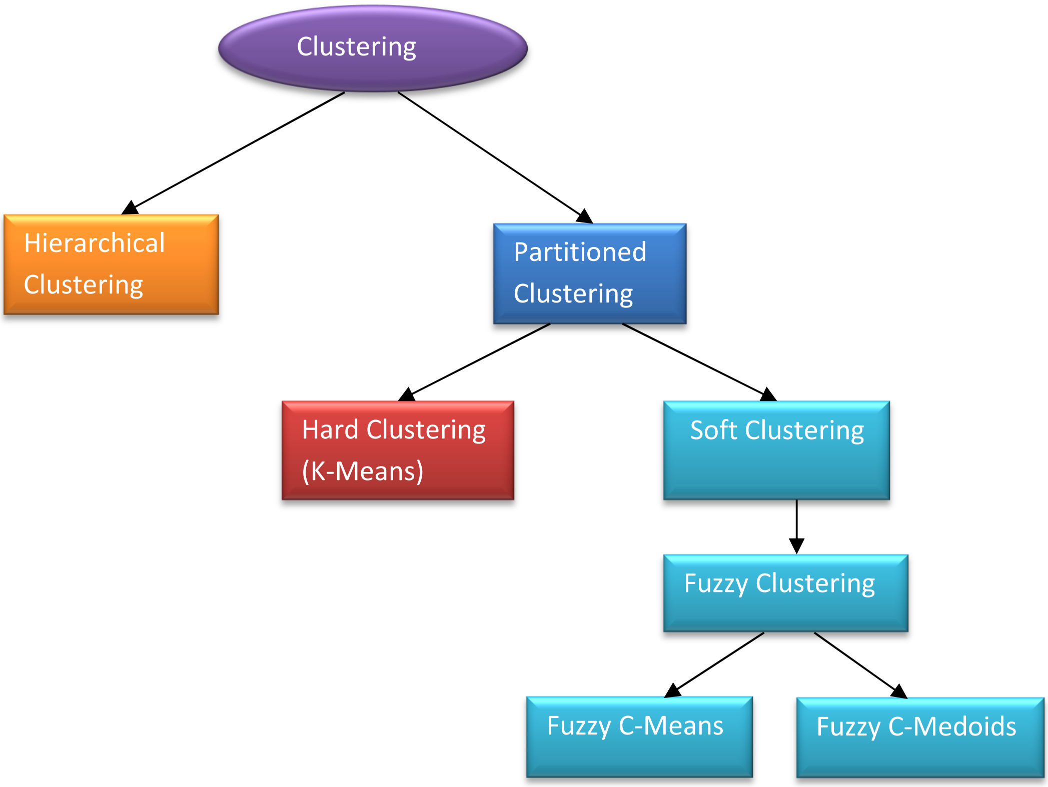


Fig 4.4: Classification of Clustering [25]

Hard Clustering mainly based on mathematical set theory i.e. either a data point belong to a particular Cluster or not that means hard partition divides the input space into the number of partitions defined by the user and as such each data point belongs to exactly one cluster of the partition. But Soft Clustering based on fuzzy set theory i.e. a data point may partially belong to a cluster.

The Fuzzy C-Means (FCM) algorithm is one of the most widely used methods in fuzzy clustering. It is based on the concept partitioning, introduced by Ruspini (1970), Dunn (1974) and Bezdek (1981). In fuzzy clustering each data point belongs to every cluster by some membership value and the process of grouping is iterated till the change in the membership values of each data point stops changing. It is a method of the clustering which allows one piece of the data to be belonged to two or more clusters. It is based on minimization of the following objective function

$$J\,(\mathrm{U,c_1,........,c_c}) = \sum_{j=1}^{c} \sum_{i=1}^{n} uij^{m}\, d_{\mathrm{ij}}^{\;2}$$

In many situations, fuzzy clustering is more natural than hard clustering. Fuzzy c-means clustering involves two processes: the calculation of cluster centres and the assignment of points to these centres using a form of Euclidian distance. This process is repeated until the cluster centres stabilize. The algorithm is similar to k-means clustering in many ways but it assigns a membership value to each data points for the clusters within a range of 0 to 1 on the basis of distance between the cluster centre and the data point. So it incorporates fuzzy set's concepts of partial membership and forms overlapping clusters to support it [35].

Fuzzy Logic + K-Means Partition = Fuzzy C-Means

In fuzzy clustering we make a fuzzy partition of the data set.

Algorithm

Let X = {x_1, x_2, x_3 ..., x_n} be the set of data points and C= {c_1, c_2, c_3 ..., c_c} be the set of centres.

Step 1: Randomly select *'c'* cluster centres and initialize the membership matrix (U)

$$\sum_{j=1}^{c} uij = 1 \qquad \forall i = 1,2 \ldots\ldots..n$$

Where, uij represents the membership function of i^{th} data point to j^{th} cluster centre.

Step 2: Calculate the fuzzy membership $'uij'$ using:

$$uij = \frac{1}{\sum_{k=1}^{c}(\frac{dij}{dik})^{(2/(m-1))}}$$

Where k is the iteration step, m€[1,∞] is fuzziness index(m is a real number which is bigger than 1. In most of the cases, m=2. If m=1, the non-fuzzy c-mean of main clustering function is obtained) and d_{ij} & d_{ik} represents euclidean distance between ith data point to jth cluster center and kth iteration respectively.

Step 3: Calculate centroids (c_j) by using

$$c_j = \frac{\sum_{i=1}^{n} uij^m xi}{\sum_{i=1}^{n} uij^m}$$

Step 4: Compute dissimilarity between centroid and data points using

$$J\,(U,c_1,\ldots\ldots\ldots,c_c)=\sum_{j=1}^{c} Jj$$

$$=\sum_{j=1}^{c}\sum_{i=1}^{n} uij^m\, d_{ij}^2$$

Where *'J'* is the objective function

Step 5: Repeat step 2 and 3 until the minimum value of *J* is achieved

$$\text{Or } \| U^{(k+1)} - U^{(k)} \| < \beta$$

Where U= $(u_{ij})_{n*c}$ is membership matrix and β€[0,1]

Advantages:

1. Data points are partially belong to one set and partially to one or more other sets but in hard partitioning each data point either belongs in a partition or is strictly excluded; there is no chance for the data points to be a part of more than on partition at the same time.
2. Gives best result for overlapped data set and comparatively better then k-means algorithm.

4.6.2 SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik. The Support Vector Machine is one of the most successful classification algorithms in the data mining area. SVM uses a high dimension space to find a hyper-plane to perform binary classification. It is based on the idea of hyper plane classifier. The goal of SVM is to find a linear optimal hyper plane so that the margin of separation between the two classes is maximized. SVM uses a high dimension space to find a hyper-plane to perform binary classification, where the error rate is minimal. The basic idea is to find a hyper-plane which separates the d-dimensional data perfectly into its two classes. However, since example data is often not linearly separable, SVM's introduce the notion of a "kernel induced feature space" which casts the data into a higher dimensional space where the data is separable. The SVM uses a portion of the data to train the system. It finds several support vectors that represent the training data. These support vectors will form a SVM model [24]. According to this model, the SVM will classify a given unknown dataset into target classes. However, for many problems they are not easy to find hyper planes to classify the data. The SVM has several kernel functions that users can apply to solving different problems. Selecting the appropriate kernel function can solve the problem of linear inseparability.

A. Basic Concept

A basic input data format and output data domains are listed as follows.

$(x_i, y_i), \ldots, (x_n, y_n), x \in R^m, y \in \{+1,-1\}$ Where $(x_i, y_i), \ldots, (x_n, y_n)$are a train data, n is the numbers of samples, m is the inputs vector, and y belongs to category of +1 or -1 respectively

[38]. Consider a hyper-plane defined by (w, b), where w is a weight vector & normal vector to the hyper plane and b is a bias. If the training data are linearly separable, we can select two hyper planes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin".

The hyper plan formula is: $(w \ . \ x) + b = 0$

The category formula is:

$(w \ . \ x) + b \geq 1$ $\quad$ *if* $yi = +1$ $\quad$ *s*

$(w \ . \ x) + b \leq -1$ $\quad$ *if* $yi = -1$

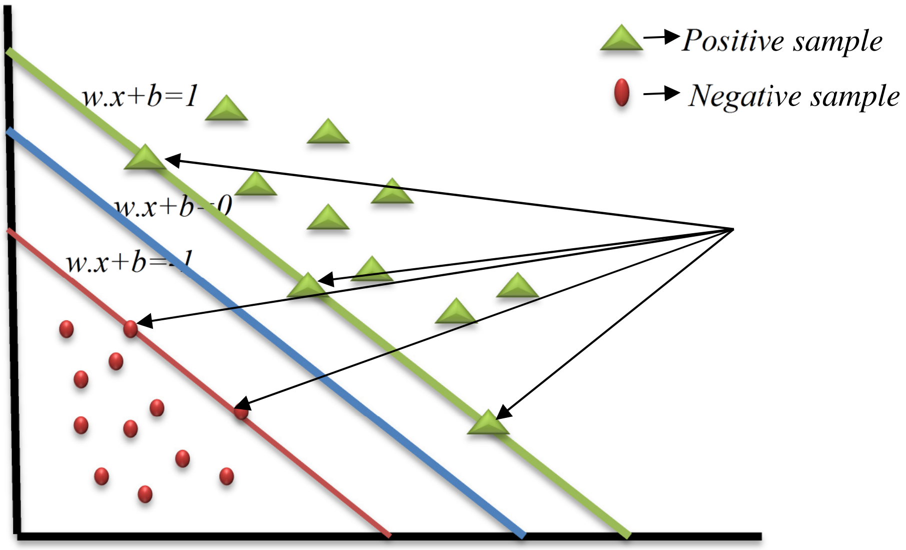


Fig. 4.5: Representation of Hyper planes of SVM

The classification of a new object x is done by the function

$$f(x)= \mathrm{sign}(w.x + b)= \mathrm{sign}(\textstyle\sum_{i=1}^{N} \alpha_i \, y_i \, (x_i . \, x)+b)$$

The training vectors x_i occur only in the form of a dot product. For each training point, there is a Lagrangian multiplier αi. The Lagrangian multiplier values αi reflects the importance of each data point. When the maximal margin hyper-plane is found, only points that lie closest to the hyper-plane will have αi > 0 and these points are called support vectors. All other points will have αi = 0. That means only those points that lie closest to the hyper plane, give the representation of the hypothesis/classifier. These data points serve as support vectors. Their values can be used to give an independent boundary with regard to the reliability of the hypothesis/classifier.

B. SVM for Multiclass Classification

The idea of using a hyper-plane to separate the feature vectors into two groups works well when there are only two target categories, but for more than two categories different approach required and many have been suggested, but two are the most popular: "one against many" and, "one against one".

1) One against Many

The earliest used implementation for SVM multiclass classification is probably the one-against-many method [31]. It constructs SVM models where is the number of classes. The ith SVM is trained with all of the examples in the ith class with positive labels, and all other examples with negative labels. Thus given training data $x1$,$y1$,….., xl,yl where xi∈Rn,i=1,…,l and yi∈(1,…,k) is the class of xi, the ith SVM.

2) One against One

The 1A1 approach on the other hand involves constructing a machine for each pair of classes resulting in (N−1)/2 machines. When applied to a test point, each classification gives one vote to the winning class and the point is labelled with the class having most votes. This approach can be further modified to give weighting to the voting process. From machine learning theory, it is acknowledged that the disadvantage the 1AA approach has over 1A1 is that its performance can be compromised due to unbalanced training datasets (Gualtieri and Cromp, 1998) [32] however, the 1A1 approach is more computationally intensive since the results of more SVM pairs ought to be computed.

Advantages

1. Support Vector Machine (SVM) that can produce high detection rate and performance with a smaller data distribution.
2. SVMs do not require a reduction in the number of features.
3. It is an effective, robust, efficient and accurate method of network action classification.
4. Faster execution
5. The major strengths of SVM are the training is relatively easy.
6. It scales relatively well to high dimensional data and the trade-off between classifier complexity and error can be controlled explicitly.

4.7 ADVANTAGES OF PROPOSED INTRUSION DETECTION SYSTEM

1. Proposed Intrusion detection system is faster than existing system in terms of execution time.
2. Proposed Intrusion detection system gives less false alarm rate, high detection rate and high accuracy.
3. It is an effective, robust and efficient Intrusion Detection System.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 INTRODUCTION

All the experimental work is done in Matlab R2010 and system information is as follows, OS-Window 7 ultimate service pack 1, processor-Intel(R) core(TM) i3 2328 CPU 2.20 GHz, Physical memory-2.00 GB, and System type- 32 bit processor.

Here we discuss and show the result of our experiment**.**

5.2 SELECT SAMPLES

For implementation of Intrusion Detection Systems, Select 1454 Training data instances from KDD Cup'99 Dataset in which 890 Normal Data and 564 Abnormal Data.

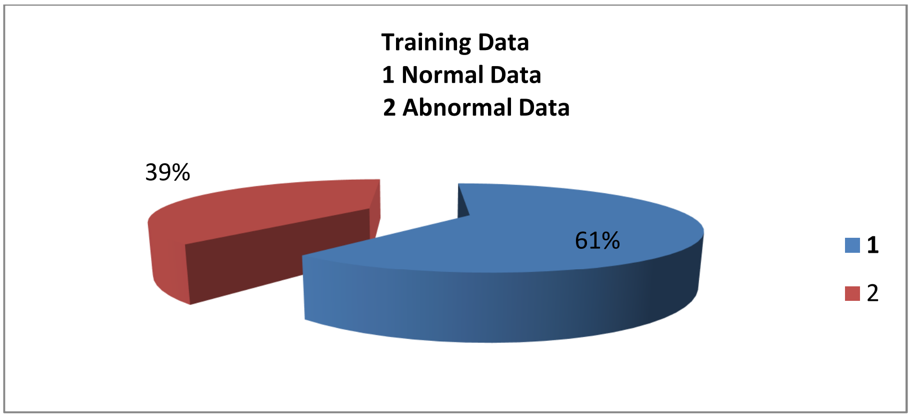


Fig 5.1: Distribution of Training data

And 836 Testing Data points in which 500 normal data and 336 abnormal data.

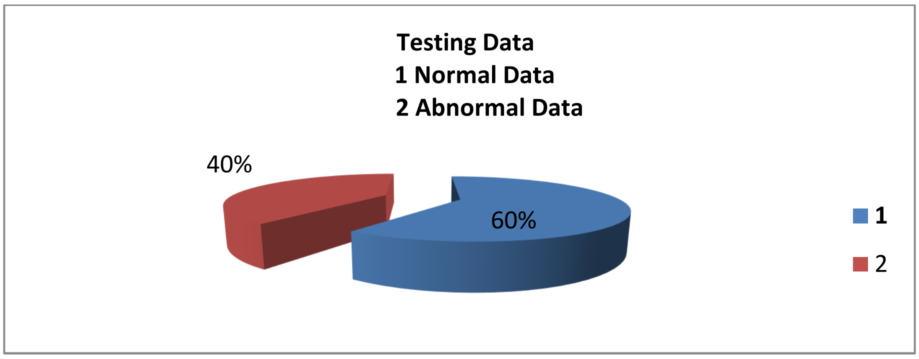


Fig 5.2: Distribution of Testing Data

5.3 DATA TRANSFORMATION

There are several text words in the KDD Cup'99 Dataset. So we need to transform the nominal features to numeric values so as to make it suitable input for Hybrid Intrusion Detection System. Here we use a Table for data transformation and according to this table nominal data is transform from nominal to numerical.

Table 5.1: Data Transformation Table

Type	Feature name	Numeric Value
Attacks or Normal	Normal	0
Attack or Normal	Attack(Abnormal)	1
Protocol type	TCP	2
	UDP	3
	ICMP	4
Flag	OTH	5
	REJ	6
	RSTO	7
	RSTOS0	8
	RSTR	9
	S0	10
	S1	11
	S2	12
	S3	13
	SF	14
	SH	15
Services	All Services	16 to 81

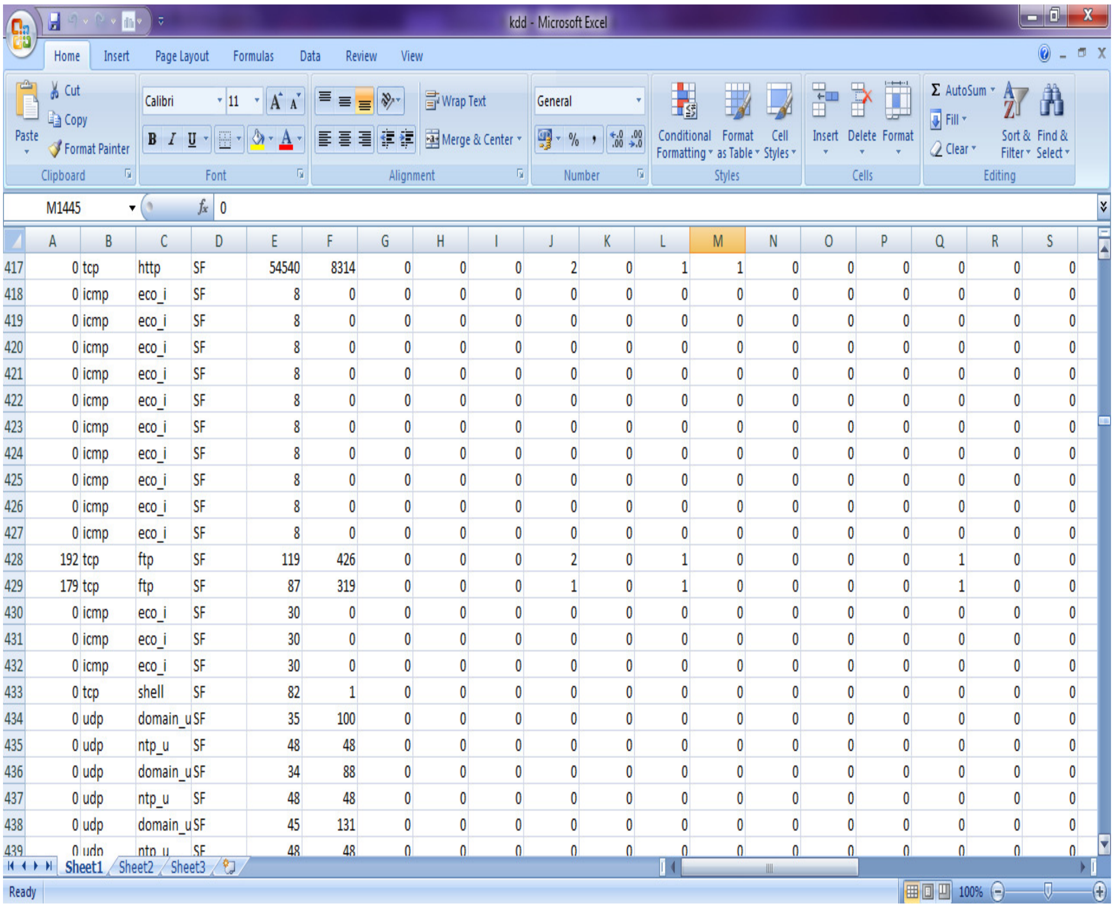
Example of Data Transformation:

The data of KDD Cup'99 Dataset

0,tcp,http,SF,181,5450,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.0s0,0.00,
9,9,1.00,0.00,0.11,0.00,0.00,0.00,0.00,0.00,normal.
0,udp,domain_u,S0,31,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,4,0.00,0.00,0.00,0.00,1.00,0.00,0.75,
85,41,0.22,0.06,0.22,0.05,0.00,0.00,0.00,0.00,smurf.

And it is transformed into nominal to numerical data using Transformation matrix and the result is:

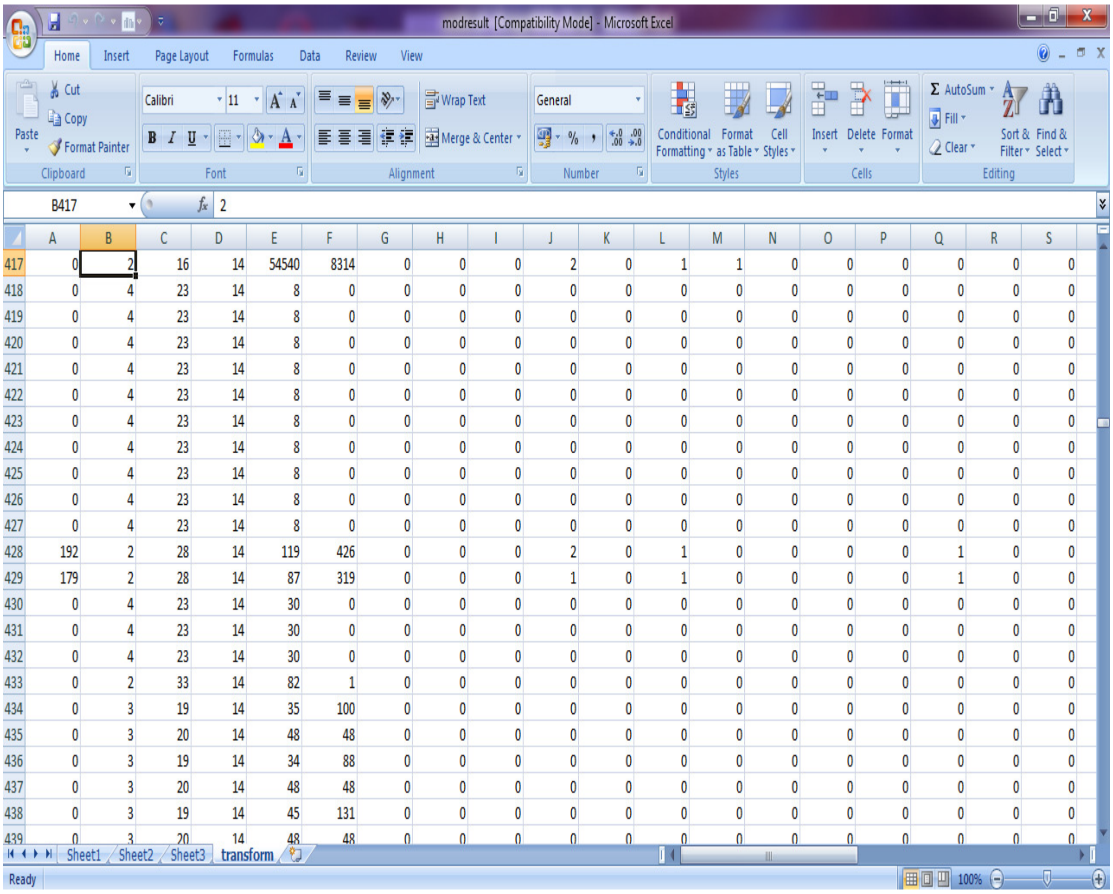
0,2,16,14,181,5450,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0
.00,9,9,1.00,0.00,0.11,0.00,0.00,0.00,0.00,0.00,0.
0,3,19,10,31,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,4,0.00,0.00,0.00,0.00,1.00,0.00,0.75,85,41,0.2
2,0.06,0.22,0.05,0.00,0.00,0.00,0.00,1.



M1445 | 0

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
417	0	tcp	http	SF	54540	8314	0	0	0	2	0	1	1	0	0	0	0	0	0
418	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
419	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
420	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
421	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
422	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
423	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
424	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
425	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
426	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
427	0	icmp	eco_i	SF	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
428	192	tcp	ftp	SF	119	426	0	0	0	2	0	1	0	0	0	0	1	0	0
429	179	tcp	ftp	SF	87	319	0	0	0	1	0	1	0	0	0	0	1	0	0
430	0	icmp	eco_i	SF	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0
431	0	icmp	eco_i	SF	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0
432	0	icmp	eco_i	SF	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0
433	0	tcp	shell	SF	82	1	0	0	0	0	0	0	0	0	0	0	0	0	0
434	0	udp	domain_u	SF	35	100	0	0	0	0	0	0	0	0	0	0	0	0	0
435	0	udp	ntp_u	SF	48	48	0	0	0	0	0	0	0	0	0	0	0	0	0
436	0	udp	domain_u	SF	34	88	0	0	0	0	0	0	0	0	0	0	0	0	0
437	0	udp	ntp_u	SF	48	48	0	0	0	0	0	0	0	0	0	0	0	0	0
438	0	udp	domain_u	SF	45	131	0	0	0	0	0	0	0	0	0	0	0	0	0
439	0	udp	ntp_u	SF	48	48	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig 5.3: Screen Shot of Data before Transformation



B417 f_x 2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
417	0	2	16	14	54540	8314	0	0	0	2	0	1	1	0	0	0	0	0	0
418	0	4	23	14	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
419	0	4	23	14	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
420	0	4	23	14	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
421	0	4	23	14	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
422	0	4	23	14	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
423	0	4	23	14	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
424	0	4	23	14	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
425	0	4	23	14	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
426	0	4	23	14	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
427	0	4	23	14	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
428	192	2	28	14	119	426	0	0	0	2	0	1	0	0	0	0	1	0	0
429	179	2	28	14	87	319	0	0	0	1	0	1	0	0	0	0	1	0	0
430	0	4	23	14	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0
431	0	4	23	14	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0
432	0	4	23	14	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0
433	0	2	33	14	82	1	0	0	0	0	0	0	0	0	0	0	0	0	0
434	0	3	19	14	35	100	0	0	0	0	0	0	0	0	0	0	0	0	0
435	0	3	20	14	48	48	0	0	0	0	0	0	0	0	0	0	0	0	0
436	0	3	19	14	34	88	0	0	0	0	0	0	0	0	0	0	0	0	0
437	0	3	20	14	48	48	0	0	0	0	0	0	0	0	0	0	0	0	0
438	0	3	19	14	45	131	0	0	0	0	0	0	0	0	0	0	0	0	0
439	0	3	20	14	48	48	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig 5.4: Screen shot of Data after Transformation

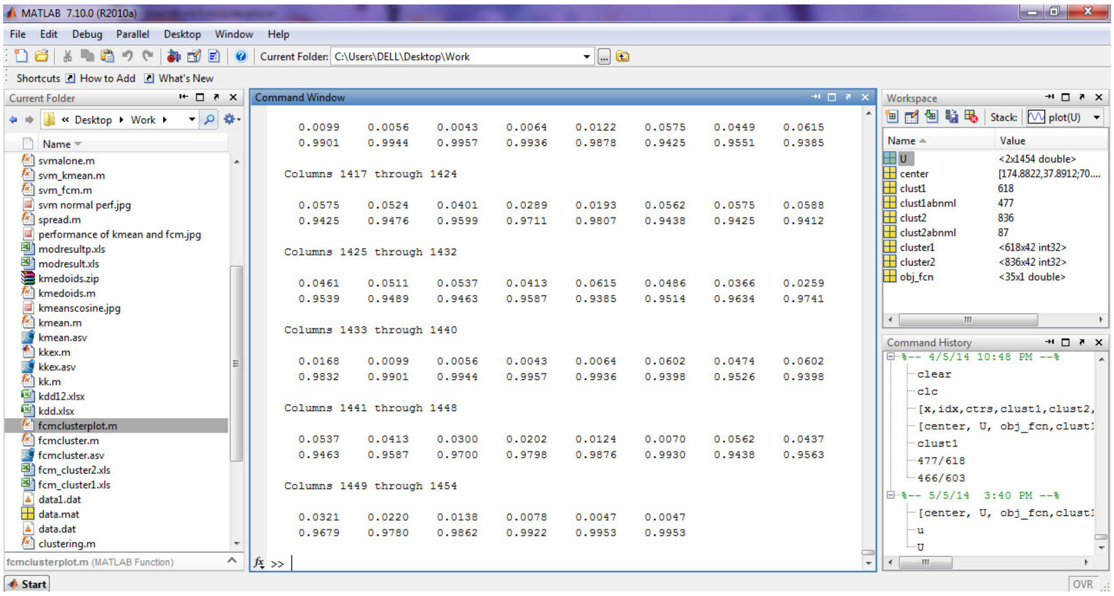
5.4 HYBRID INTRUSION DETECTION SYSTEM

Combination of Fuzzy C-Means clustering technique and Support Vector Machine classification technique is used for designing Hybrid Intrusion Detection System.

5.4.1 Fuzzy C-Means clustering technique

Step 1: Select number of cluster C=2

Step 2: Calculate Membership matrix U



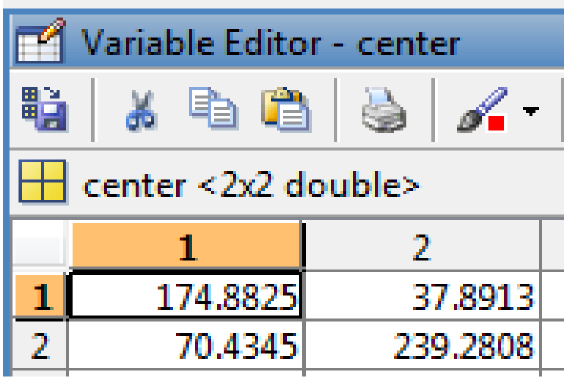
```
    0.0099    0.0056    0.0043    0.0064    0.0122    0.0575    0.0449    0.0615
    0.9901    0.9944    0.9957    0.9936    0.9878    0.9425    0.9551    0.9385

  Columns 1417 through 1424

    0.0575    0.0524    0.0401    0.0289    0.0193    0.0562    0.0575    0.0588
    0.9425    0.9476    0.9599    0.9711    0.9807    0.9438    0.9425    0.9412

  Columns 1425 through 1432

    0.0461    0.0511    0.0537    0.0413    0.0615    0.0486    0.0366    0.0259
    0.9539    0.9489    0.9463    0.9587    0.9385    0.9514    0.9634    0.9741

  Columns 1433 through 1440

    0.0168    0.0099    0.0056    0.0043    0.0064    0.0602    0.0474    0.0602
    0.9832    0.9901    0.9944    0.9957    0.9936    0.9398    0.9526    0.9398

  Columns 1441 through 1448

    0.0537    0.0413    0.0300    0.0202    0.0124    0.0070    0.0562    0.0437
    0.9463    0.9587    0.9700    0.9798    0.9876    0.9930    0.9438    0.9563

  Columns 1449 through 1454

    0.0321    0.0220    0.0138    0.0078    0.0047    0.0047
    0.9679    0.9780    0.9862    0.9922    0.9953    0.9953

fx >>
```

Name	Value
U	<2x1454 double>
center	[174.8822,37.8912;70....
clust1	618
clust1abnml	477
clust2	836
clust2abnml	87
cluster1	<618x42 int32>
cluster2	<836x42 int32>
obj_fcn	<35x1 double>

```
%-- 4/5/14 10:48 PM --%
  clear
  clc
  [x,idx,ctrs,clust1,clust2,
  [center, U, obj_fcn,clust1
  clust1
  477/618
  466/603
%-- 5/5/14  3:40 PM --%
  [center, U, obj_fcn,clust1
  u
  U
```

Fig 5.5: Screen shot of Membership Matrix U

Step 3: Calculate centroids (c_j)

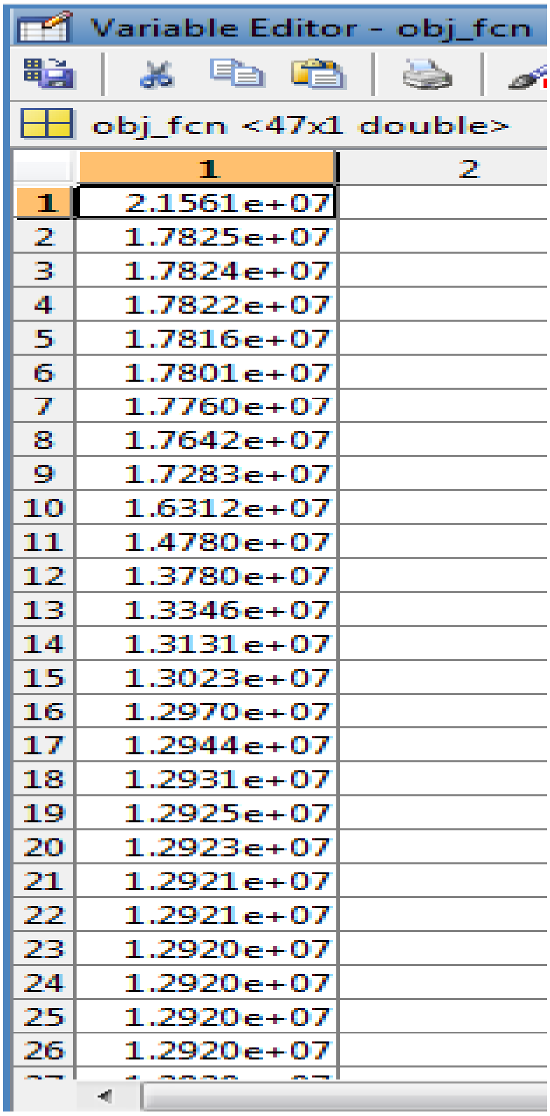


center <2x2 double>

	1	2
1	174.8825	37.8913
2	70.4345	239.2808

Fig 5.6: Screen shot of Centroids

Step 4: Compute dissimilarity between centroid and data points by using objective function *J*



Variable Editor - obj_fcn

obj_fcn <47x1 double>

	1	2
1	2.1561e+07	
2	1.7825e+07	
3	1.7824e+07	
4	1.7822e+07	
5	1.7816e+07	
6	1.7801e+07	
7	1.7760e+07	
8	1.7642e+07	
9	1.7283e+07	
10	1.6312e+07	
11	1.4780e+07	
12	1.3780e+07	
13	1.3346e+07	
14	1.3131e+07	
15	1.3023e+07	
16	1.2970e+07	
17	1.2944e+07	
18	1.2931e+07	
19	1.2925e+07	
20	1.2923e+07	
21	1.2921e+07	
22	1.2921e+07	
23	1.2920e+07	
24	1.2920e+07	
25	1.2920e+07	
26	1.2920e+07	

Fig 5.7: Screen Shot of Objective Function

Step 5: Data points are divided into 2 clusters normal and abnormal

836 data points cluster into normal category and 618 data points cluster into abnormal category.

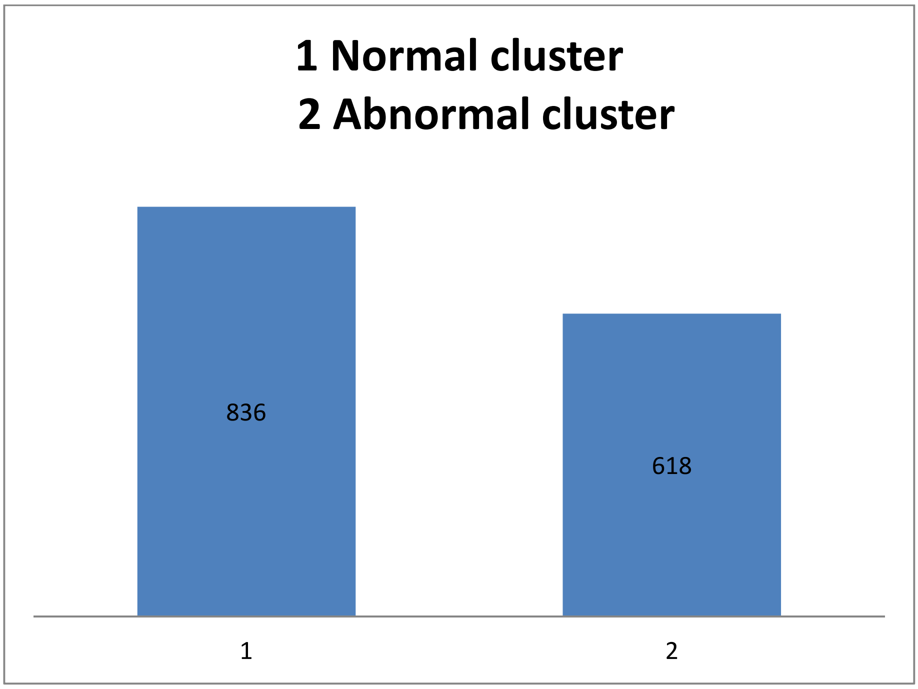


Fig 5.8: Data after FCM Clustering

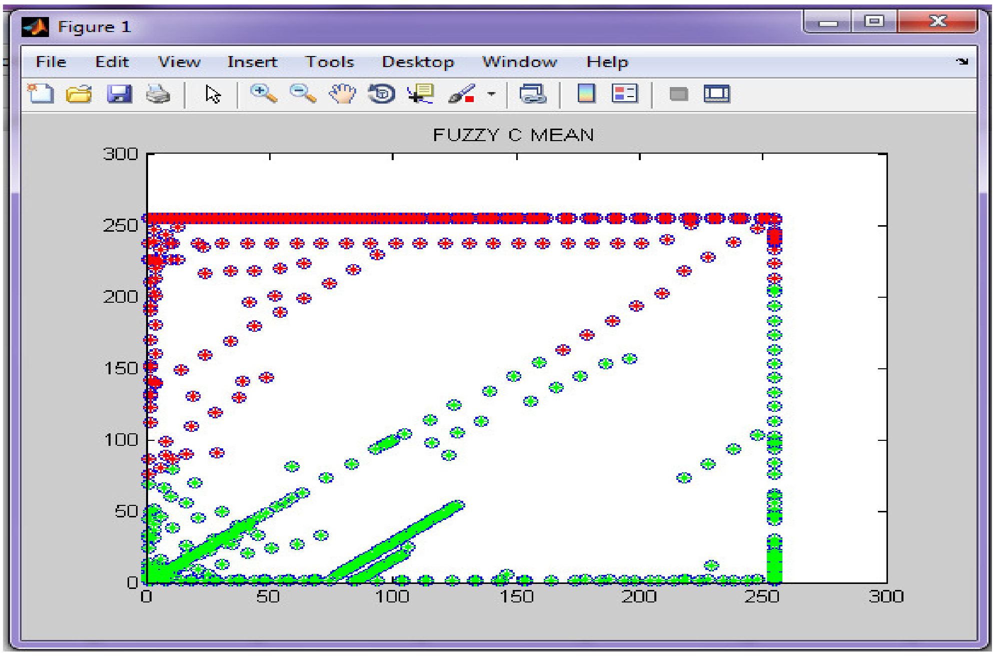


Fig 5.9: Data Distribution Plotting after Applying FCM Clustering

5.4.2 Support Vector Machine Classification Technique

When we use support vector machine alone for implementation of IDS then the performance of the system varies from 79.00 to 85.15, although performance of combination of SVM and FCM varies from 99.20 to 99.76 and combination of SVM and K-Mean varies from 98.20 to 99.25. So it is better approach to go with combination of FCM and SVM data mining algorithms. Pairing of these data mining techniques should be as follows, FCM +SVM and K-Means + SVM. On the basis of performance of these two combinations, they become two important and must be comparable algorithms on the basis of Accuracy, Detection rate and False alarm rate.

By using K-Means clustering technique, dataset is divided into 2 clusters, one is normal cluster and another is abnormal cluster according to behavior of data instances.

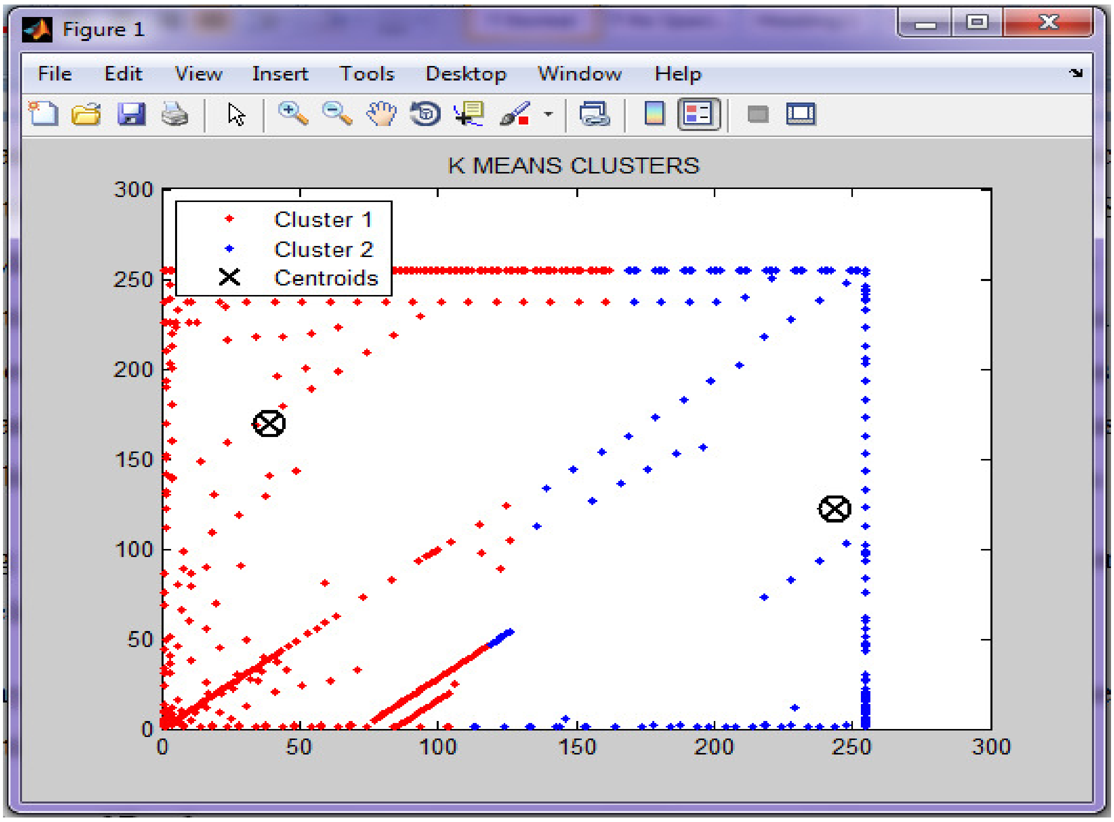


Fig 5.10: Data Distribution Plotting after Applying K –Means Clustering

After that apply SVM classification method on both FCM and K-Means clustering algorithm separately. Hence my further representation will be based on comparative analysis of these two combinations of algorithms.

Comparison of Performance:

Performance measurement is the process of collecting, analysing and/or reporting information regarding the performance of a system.

Table 5.2: Performance Comparison

SVM+FCM	SVM+K-Means
0.9952	0.9925
0.9976	0.9928

Comparison of Detection Rate:

The detection rate is defined as the number of intrusion instances detected by the system (True Positive) divided by the total number of intrusion instances present in the test set.

Table 5.3: Detection Rate Comparison

KDD dataset	Clustering techniques			
Classification	K mean	K- medoid	FCM	None
SVM	99.20-99.43	99.75-99.97[20]	99.73-99.97	71.6[22]
Naïve bayes	91.03-98.15	97.43-98.69[20]	97.86-98.13	None
Rough set fuzzy SVM	None	None	None	85.76[22]
Two phase classification	None	None	None	91.35[22]

Comparison of Accuracy: Accuracy is also used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition. It is the proportion of true results (both true positives and true negatives)

Table 5.4: Accuracy Comparison

KDD dataset	Clustering techniques			
Classification	K mean	K medoid	FCM	None
SVM	98.20-99.76	99.33-99.79[20]	99.56-99.98	86[22]
Naïve bayes	91.59-97.90	96.83-97.70[20]	96.16-96.40	
Rough set fuzzy SVM	None	None	None	90[22]
Two phase classification	None	None	None	95.59[22]

Comparison of False Alarm Rate:

False Alarm Rate: defined as the number of 'normal' patterns classified as attacks (False Positive) divided by the total number of 'normal' patterns.

Table 5.5: False Alarm Rate Comparison

KDD dataset	Clustering techniques			
classification	K mean	K medoid	FCM	None
SVM	0.0023-0.0028	0.05-0.52[20]	0.0025-0.0027	28.4[22]
Naïve bayes	0.05-0.053	1.99-2.37[20]	0.018-0.024	None
Rough set fuzzy SVM	None	None	None	14.24[22]
Two phase classification	None	None	None	1.80[22]

Confusion Matrix

Confusion Matrix is usually called a matching matrix. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. Performance of such systems is commonly evaluated using the data in the matrix.

	Predicted		
Actual		0	1
	0	TP	FP
	1	FN	TN

Confusion Matrix of FCM+SVM:

For accuracy 99.52%, C= $\begin{bmatrix} 373 & 1 \\ 1 & 42 \end{bmatrix}$

For accuracy 99.76%, C=$\begin{bmatrix} 374 & 1 \\ 0 & 42 \end{bmatrix}$

Confusion Matrix of K-Means+SVM:

For accuracy 99.25%, C= $\begin{bmatrix} 167 & 2 \\ 1 & 258 \end{bmatrix}$

For accuracy 99.28%, C=$\begin{bmatrix} 167 & 1 \\ 1 & 259 \end{bmatrix}$

CONCLUSION AND FUTURE SCOPE

In this paper, it operates on normal data’s and all performance related attributes are calculated on the basis of these normal data’s .Comparison of these two leading clustering techniques (K-Means and Fuzzy C-Mean) along with Support vector machine has some variation in result but more weightage come under the pocket of Fuzzy C-Mean and Support vector machine. A comparatively better clustering method, Fuzzy C-Mean, is combined with Support Vector Machine to produce better classification performance in terms of Accuracy, Detection Rate, performance and False Alarm Rate. From the experiments and analyses, it is shown that the proposed hybrid approach has outperformed the approach of combining Support vector machine with k-Means. Therefore, intrusion detection can be more effective and efficient with the new proposed approach.

The proposed approach can further be made more efficient and effective, by applying multiple kernel based SVM classification schemes, in detecting various known and unknown attacks and thus separating them into correct categories.

REFERENCES

[1] B.A. Nahla, B. Salem, and E. Zied, “Naïve Bayes vs Decision Trees in Intrusion Detection Systems”, ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004.

[2] Carlos A. Catania, Facundo Bromberg, Carlos Garcia Garino, “An Autonomous Labelling Approach to Support Vector Machine Algorithms for Network Traffic Anomaly Detection”, Expert Systems with Applications: An International Journal Archive, vol. 39 Issue 2,February 2012.

[3] C. F. Tsai, and C.Y Lin, “A Triangle Area-Based Nearest neighbours Approach to Intrusion Detection,” Pattern Recognition, 43(1), 2010.

[4] Chunhua Gu and Xueqin Zhang , A Rough Set and SVM Based Intrusion Detection Classifier, Second International Workshop on Computer Science and Engineering, 2009

[5] C.H. Tsang, S. Kwong, and H. Wang, “Genetic-Fuzzy Rule Mining Approach and Evaluation of Feature Selection Techniques for Anomaly Intrusion Detection,” Pattern Recognition, 40:2373–2391, 2007.0

[6] Deepthy K Denatious & Anita John, “Survey on Data Mining Techniques to Enhance Intrusion Detection”, International Conference on Computer Communication and Informatics, Coimbatore, INDIA, IEEE, Jan 10-12 2012.

[7] Fabrice Colas and Pavel Brazdil, “Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks”, Proceedings - IFIP AI Proceeding, pp.169-178, 2006.

[8] Fatin Norsyafawati Mohd Sabri, Norita Md Norwawi, Kamaruzzaman Seman, “Hybrid of Rough Set Theory and Artificial Immune Recognition System as a Solution to Decrease False Alarm Rate in Intrusion Detection System”, 7th International Conference on Information Assurance and Security (IAS) ,pp 134-138, 2011.

[9] G. Meera Gandhi, Kumaravel Appavoo, S.K. Srivatsa, “Effective Network Intrusion Detection using Classifiers Decision Trees and Decision rules”, Advanced Networking and Applications, vol. 2, no. 3, pp: 686-692, 2010.

[10] Hai Nguyen, Katrin Franke, Slobodan Petrovic, “Improving Effectiveness of Intrusion Detection by Correlation Feature Selection”, Proc. of IEEE Intl. Conference on Availability, Reliability and security,pp.17-24, 2010.

[11] Jiawei Han, Micheline Kamber, " Data Mining - Concepts and Techniques" Elsevier (Second Edition),pp 359 to 367, 2003.

[12] Jin Huang, Jingjing Lu, Charles X. Ling, "Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy", The Third International Conference on Data Mining, 2003.

[13] KDDCup'99 The Third International Knowledge Discovery and Data Mining Tools Competition Dataset KDD Cup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/ kddcup99.html, 1999.

[14] L. Bo, and C. Y. Yuan, "The research of Intrusion Detection based on Support Vector Machine", International Conference on Computer and Communications Security, Chicago, USA, pp. 21-23, November, 9-13, 2009.

[15] Maria Muntean, Honoriu Vălean, Liviu Miclea, Arpad Incze, "A Novel Intrusion Detection Method Based on Support Vector Machines", International Symposium on Computational Intelligence and Informatics, Budapest, Hungary , IEEE, pp-47-52, 18–20 November, 2010

[16] Meng Jianliang, Shang Haikum, Bian Ling, "The Application on Intrusion Detection Based on K-means Cluster Algorithm", International Forum on Information Technology and Applications, IEEE, 2009.

[17] Ming Xue, Changjun Zhu, "Applies Research on Data Mining Algorithm in Network Intrusion Detection", International Joint Conference on Artificial Intelligence ,IEEE, 2009.

[18] M.Bahrololum, E. Salahi, M. Khalegi, "Machine Learning Techniques for feature reduction in Intrusion Detection Systems: A Comparison", Proc. of IEEE Intl. Conf. on Computer Science and Convergence Information Technology, pp.1091-1095, 2009.

[19] M. Hong, G. Yanchun, W. Yujie, and L. Xiaoying, "Study on classification method based on Support Vector Machine", First International Workshop on Education Technology and Computer Science, Wuhan, China, pp.369-373, March, 7-8, 2009.

[20] M. Gao, J. Tian, and M. Xia, "Intrusion Detection Method Based on Classify Support Vector Machine", Second International Conference on Intelligent Computation Technology and Automation, Zhangjiajie, China, pp. 391-394, October, 10-11, 2009.

[21] M.Govindarajan, Rlvl.Chandrasekaran, "Intrusion Detection Using k-Nearest Neighbor", pp 13-20, IEEE, 2009.

[22] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi, Lilly Suriani Affendey, “Intrusion Detection Using Data Mining Techniques”,Proc. of IEEE Intl. Conf on Information Retrieval & Knowledge Management,pp.200-203, 2010.

[23] Milan kumari and sunila godara, "Comparative Study of Data Mining classification Methods in Cardiovascular Disease Prediction," International Journal of Computer Science &Technology, pp.1-3, June 2011.

[24] M. Chandrasekhar and K. Raghuveer, “Intrusion detection technique by using K-Means, Fuzzy Neural network and SVM Classifiers”, International conference computer communication and informatics, 2013.

[25] Mrutyunjaya Panda and Manas Ranjan Patra, “Network intrusion detection using naïve bayes , IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12, pp.258-263, December 2007.

[26] Mrutyunjaya Panda and Manas Ranjan Patra, “ A comparative study of data mining algorithm for network intrusion detection” , IEEE conference on Emerging trends in engineering and technology, 2008

[27] Peddabachigiri S., A. Abraham., C. Grosan and J. Thomas, “ Modeling of Intrusion Detection System Using Hybrid intelligent systems” , Journals of network computer application, pp 114-132, 2007

[28] P Amudha, H Abdul Rauf, “Performance Analysis of Data Mining Approaches in Intrusion Detection”, IEEE, 2011.

[29] QU Zhiming and Wang Xiaoli, “ Study of Rough set and Clustering algorithm in network security management” , International conference on network security, 2009

[30] Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh, “using rough set and support vector machine for network intrusion detection”, International Journal of Network Security & Its Applications (IJNSA),Vol 1, No 1, April 2009

[31] Roshan Chitrakar and Huang Chuanhe, “Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-Medoids Clustering and Naïve Bayes Classification”, The 8th International Conference on Wireless Communication, Networking and Moblie Computing, Shanghai, China, 2012.

[32] Roshan Chitrakar and Huang Chuanhe, “Anomaly Detection using Support Vector Machine Classification with k-Medoids Clustering”, IEEE 2012

[33] R.China Appala Naidu, P.S.Avadhani, “Comparison of Data Mining Techniques for Intrusion Detection”, International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), IEEE, 2012.

[34] Rachnakulhare ,Divakar Singh, “Intrusion Detection System based on Fuzzy C Means Clustering and Probabilistic Neural Network”, International Journal of Computer Applications (0975 – 8887) Volume 74– No.2, July 2013.

[35] Song Naiping , Zhou Genyuan , “A study on Intrusion Detection Based on Data Mining”, International Conference of Information Science and Management Engineering, pp 135-138, IEEE 2010.

[36] S. Wu, E. Yen, “Data mining-based intrusion detectors,” Elsevier computer Network, 2009.

[37] S. Rubin, S. Jha, and B. Miller, “Automatic Generation and Analysis of NIDS Attacks”, Proceedings of 20th Annual Computer Security Application Conference, IEEE Computer Society, pp.28-38, 2004.

[38] T. Velmurugan and T. Santhanam, “Computational Complexity between k-Means and k-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points”, Journal of Computer Science, 6 (3), pp 363-368, 2010.

[39] WitchaiChimphlee, Abdul Hanan Abdullah, Mohd Noor Md SapSiriporn Chimphleel and Surat Srinoy, “Integrating Genetic Algorithms and Fuzzy c-Means for Anomaly Detection, IEEE Indicon Conference, Chennai, India, pp 575-579, Dec. 2005.

[40] Xiang, C., M.Y. Chong and H.L.Zhu, “Design of Multiple-Level Tree Classifiers for Intrusion Detection System”. IEEE Conference on Cybernetics and Intellligent Systems (CCIS 2004), Singapore, pp: 873-878, 2004.

[41] Xiaohui Bao, Tianqi Xu, Hui Hou, “Network Intrusion Detection Based on Support Vector Machine”, Proceedings - International Conference on Management and Service Science, 2009.

[42] Yihua Liao, V. RaoVemuri, "Use of K-Nearest Neighbor classifier for intrusion detection" Computers & Security, Vol 21, No 5, pp 439-448, 2002

[43] Y. Li and L. Guo, “An Active Learning Based on TCM-KNN Algorithm for Supervised Network Intrusion”, Computer and Securtiy, 26: 459-467, 2007.

[44] 25Yinhui Li, Jingbo Xia, Silan Zhang, Jiakai Yan, Xiaochuan Ai, Kuobin Dai, “An efficient intrusion detection system based on support vector machines and gradually

feature removal method", Expert Systems with Applications, v 39, n 1, pp 424-430, January 2012.

[45] Yuping Li , Weidong Li, Guoqiang Wu, "An Intrusion Detection Approach Using SVM and Multiple Kernel Method", International Journal of Advancements in Computing Technology, v 4, n 1, p 463-469,January 2012.

[46] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir , "Intrusion Detection based on k-Means Clustering and Naïve Bayes Classification" , 7th International Conference on IT in Asia (CITA) , 2011.