A Hybrid Approach to improve the Anomaly Detection Rate Using Data Mining Techniques

Thesis submitted in partial fulfillment of the requirements for the award of

Degree of

Master of Engineering
in
Information Security

Submitted By
Priya Bansal
801333018

Under the supervision of:
Dr. Deepak Garg
Associate Professor & Head

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

July 2015

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "*A Hybrid Approach to Improve the Anomaly Detection Rate Using Data Mining Techniques*" in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Information Security* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Deepak Garg* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Priya Bansal

(Priya Bansal)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Deepak Garg)
Associate Professor
Computer Science and Engineering Department

Countersigned by:

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. S. Bhatia)
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgment

First of all, I am extremely thankful to my respective guide Dr. Deepak Garg Associate professor, CSED, Thapar University for his valuable guidance, advice, motivation, encouragement, moral support, sincere effort and positive attitude with which he solved my queries and provide delightful ambiance for learning, exploring and making this thesis possible. He always set high goals for me and help me to find the direction to get those goals. It has been a great experience to work under his sanctuary.

I would like to thank my family members and my friends who are dearest and precious to me for their love, encouragement, blessings and support in all respects. Most importantly, none of this would have been possible without the love and patience of my family. To my brother Varun Bansal and friends for showing me the right direction. They are a constant source of love, concern, support and strength for me all these years.

Finally, I would like to thank the management of Thapar University for proving a great platform for learning, not just for academics but also for sports and many other creative things.

Priya Bansal
801333018
ME (IS)

Abstract

An Intrusion Detection System is a device or software application that monitors events occurring on the network and analyzes it for any kind of malicious activity that violates computer security policy. With an increase in dependency rate on the internet, there is a significant increase in the number of internet attacks as well. The challenges arise towards the network security due to the introduction of new methods of attacks. To identify these attacks, a new hybrid approach using data mining based on C4.5 and Meta algorithm is proposed.This approach provides a classifier which improves the overall accuracy of detection. Various data mining techniques have been developed for detecting intrusion. For detection of anomalies a hybrid technique based on C4.5 and meta-algorithm is proposed that provides better accuracy and reduces the problem of high false alarm ratio. The comparison of the proposed approach is made with other data mining techniques. With this proposed approach detection rate is improved considerably. The experimentation is implemented in WEKA tool using KDD Cup 1999 dataset

Table of Contents

List of Figures

List of Tables

Abbreviations

IDES	Intrusion Detection Expert System
IDS	Intrusion Detection System
IPS	Intrusion Prevention System
NIDS	Network based Intrusion Detection System
HIDS	Host based Intrusion Detection System
AIDS	Application based Intrusion Detection System
SPADE	Statistical Packet Anomaly Detection Engine
P-BEST	Production based Expert System Toolset
NATE	Network Analysis of Anomalous Traffic Events
ADAM	Audit Data and Analysis and Mining
FER	Frequent Episode Rules
SMF	System Management Facility
PCA	Principal Component Analysis
JAM	Java Agents for Meta-learning
MADAM-ID	Mining Audit Data for Automated Models for Intrusion Detection
SVM	Support Vector Machine
DT	Decision Tree
DGSOT	Dynamically Growing Self-Organizing Tree
SA	Simulated Annealing
CCNN	Cluster Center and Nearest Neighbor
IG	Information Gain
Weka	Waikato Environment for Knowledge Analysis
GPL	General Public License
GUI	Graphic User Interface
CLI	Command Line Interface
DOS	Denial of Services
U2R	User to Root
R2L	Remote to Local
IQR	Interquartile Range
IDE	Integrated Development Environment

JVM	Java Virtual Machine
PPV	Positive Predictive Value
FPR	False Positive Rate
TPR	True Positive Rate

Chapter 1

Introduction

1.1 Brief Summary

Over the last decade, we have become too much technology dependent. Now-a-days we rely on networks to receive emails, banking, stock price, news and online shopping. The excessive use of the communication networks leads to terrorism. Due to that it raises the need of secure and safe system. Because of the dependence on the computer technology, we have to significantly improve computer network security, so that data integrity, confidentiality and availability do not hamper. All computers are vulnerable to compromise and every network is at risk to unauthorized access and leakage of private and sensitive information.

A firewall is widely and actively used in security mechanisms. It configures security policy, but that has also been in vain, because it cannot prevent from all kinds of malicious intent of the intruder or attacker. In case of a firewall, only its header content is examined whereas in Intrusion Detection System (IDS) both content and header of packet are examined. So, IDS is much more dynamic as compared to firewall in order to secure our private and sensitive data. IDS has proven an important tool for security, but we cannot replace firewall completely with IDS as they work complementary to each otherby analyzing all kinds of misuse patterns on networks.

An intrusion is defined as any kind of action that compromises the integrity, confidentiality or Availability. Although it plays a very significant role to define and protect in security architecture, but IDS is still immature and not considered as a complete defense,. IDS identifies or monitors any kind of intrusion and notify immediately in the form of alert so that resources never get compromised. An IDS is also used in legal proceedings as forensic evidence against the intruder because it provides recording of any kind of intrusion involved in cybercrime. An IDS is deployed to cover unauthorized access to resources or data.It can be hardware and/or software. An IDS can be used to protect a single host or a whole computer network. IDS provides user friendly interface to non-expert staff for managing the systems easily.

.

1.2 Intrusion Detection System

Intrusion is any kind of unauthorized activity on a computer network .It is achieved passively or actively. In passive, intrusion takes place by information gathering and eavesdropping,whereas in case of active intrusion takes place through harmful packet forwarding, packet dropping and by hole attacks[1]. An IDS is a process or device that monitors events occurring on a network and analyzing it to detect any kind of activity that violate computer security policies. The IDS device can be hardware, software or a combination of both that monitors the computer network against any unauthorized access[2]. The main motive of the IDS is to catch the intruder before a real and serious damage to computer network.

To protect from the attack and malicious activity IDS provides following features [3] [4]:

- Monitoring and analyzing network user and computer system activity
- Auditing computer system policy configurations and vulnerabilities
- Accessing integrity of critical data server and file system
- Statistical analysis of pattern matching to the known attacks
- Unauthorized activity analysis
- Operating system auditing
- Record information on abnormal events
- Alert administrators about malicious activity
- Producing reports

1.2.1 Types of Alert

An IDS notifies the administrator by using alerts about the unauthorized access to the computer system. IDS generates a huge number of alerts. IDS classify alerts into four categories [5]:-

- True Positive: A real intrusion for which IDS generates an alert.
- False Positive: No intrusion, but IDS generate an alert.
- False Negative: A real intrusion, but IDS never generate any kind of alert.
- True Negative: No intrusion and IDS never generate an alert.

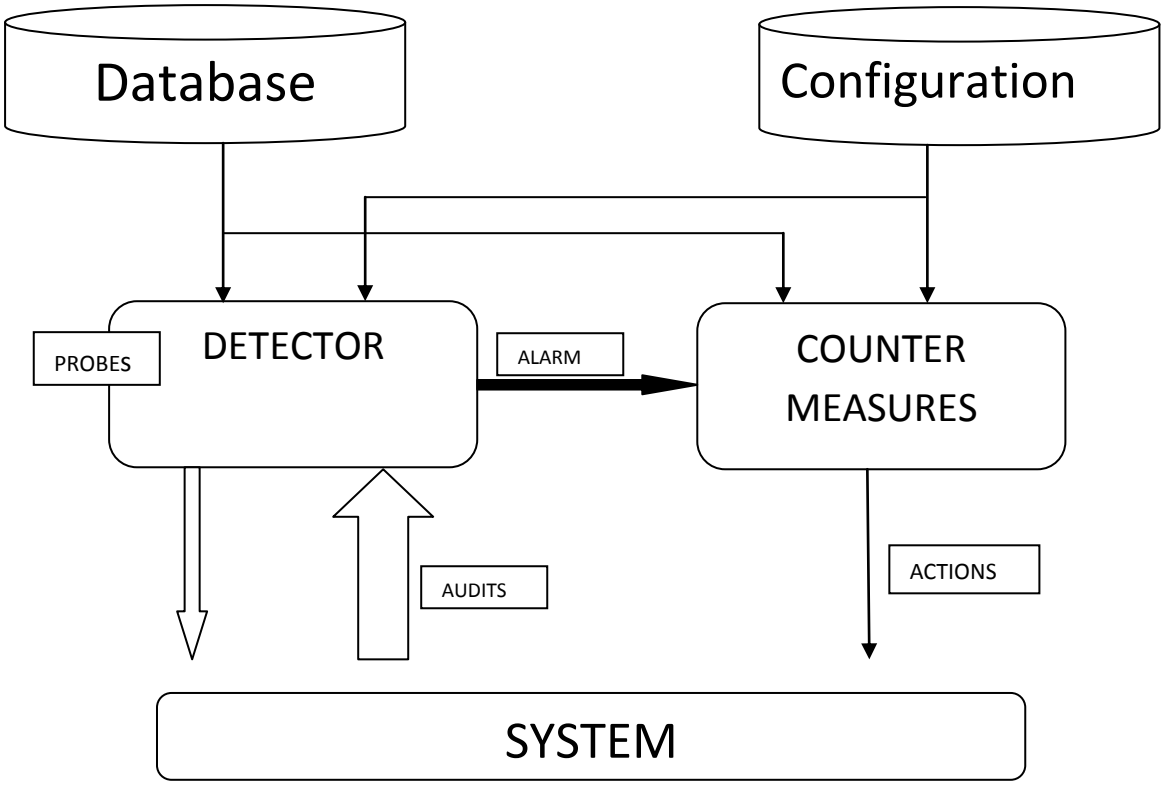

Figure 1. Simple Intrusion Detection System [4]

1.2.2 Types of IDS

IDS uses different types of technologies. In IDS we can classify the data into the three parts for the basic understanding[6]. IDS technologies are classified as [7:

1. Network Based Intrusion Detection System (NIDS)
2. Host Based Intrusion Detection System (HIDS)
3. Application-based Intrusion detection System (AIDS)

1.2.2.1 Network based Intrusion Detection System (NIDS)

It monitors network traffic for a particular network area or devices and search protocol action to identify different kinds of malicious activities. It is deployed at a boundary between two networks. This is also known as Traffic Based Intrusion Detection System[8].

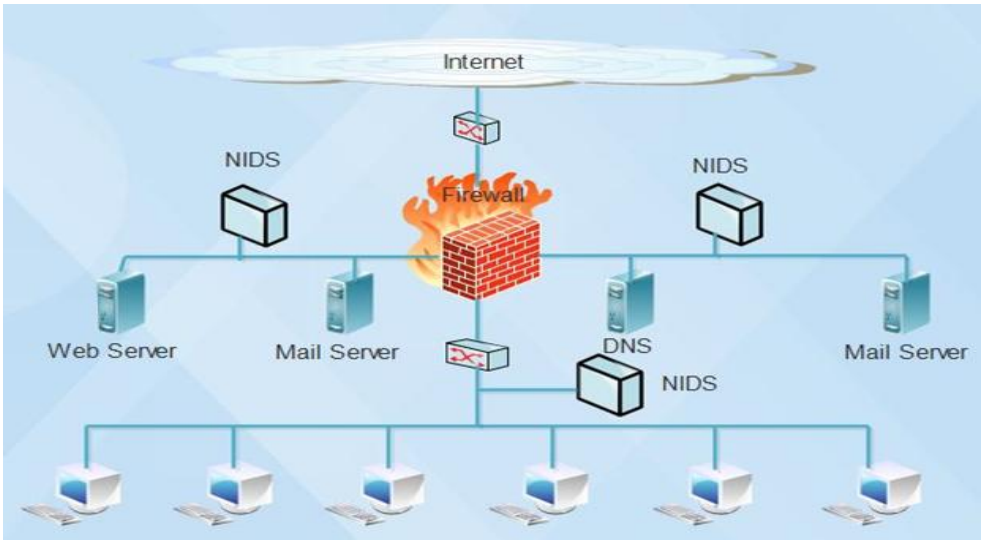

Figure 2. Architecture of Network based IDS

1.2.2.2 Host based Intrusion Detection System (HIDS)

It monitors the traffic of a single host and the events occurring within that host of the malicious activities [9]. It monitors the log files ,various running processes and applications;file access and modification;system and application configuration changes on a single host[10].

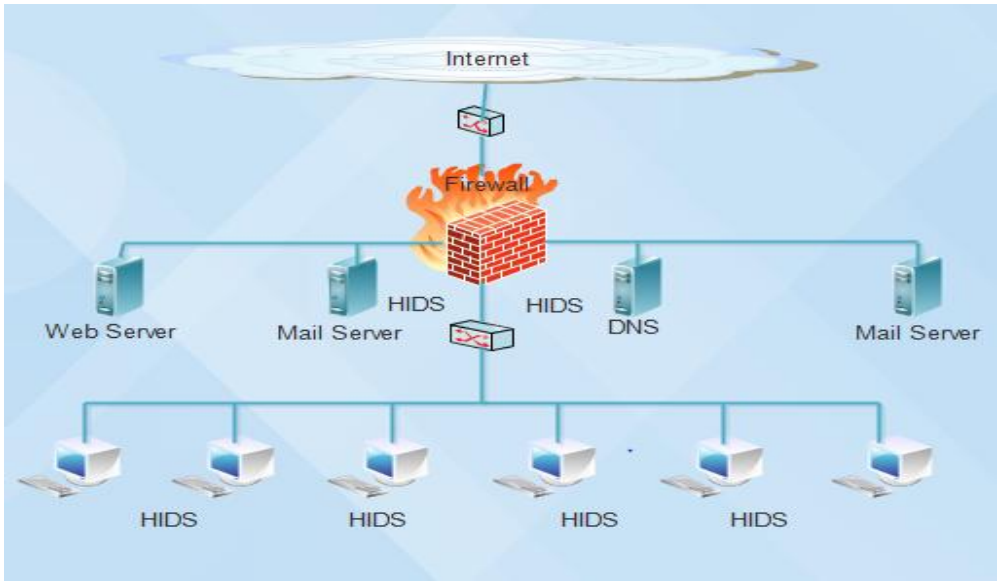

Figure 3. Architecture of Host based IDS

1.2.2.3 Application based Intrusion Detection System (AIDS)

It monitors the traffic on a network based on events that occur in some kind of specific applications. This may be Host based , Network based or hybrid. It monitors

the traffic of a single host and the events happening within that host for malicious activities[11]. It monitors the log file, various running processes, running, applications, file access and modification, system and application configuration changes on a single host. It is basically deployed on critical hosts like publicly accessible servers and those servers having critical information[12].

1.2.3 Types of Detection

In real-time detection, when an IDS monitor for any kind of intrusion in the computer network it immediately notifies in the form of alert and proper action istaken accordingly so that system security never breaches. The real-time IDS works in off-line mode by using the previous or historical data collected from intrusion identified in the past[13]. Non-real-time IDS, as the name suggests analyzes the data with delay. In this approach, data can be collected from different locations on the network and after that it is collected in a centralized or on a single system for further usage. In IDS currently there are a several varieties of detection approaches, but the two main approaches are [9]:

1.2.3.1 Misuse/Signature based Detection

This approach system has the collection of abnormal patterns of unauthorized activities.In this detection engine identify only well-known patterns of attacks which exploit the vulnerabilities and weakness of software. These patterns are known as signatures. Therefore, this detection technique is known as a signature based detection[14]. This detection technique is effective for detecting known attacks. It generates a low number of false positive alerts but the main limitation of this technique is that it detects only the known intrusion and never detects the unknown intrusion which may lead to the zero day attack[15]. It consists of some knowledge based system which includes the database of attack signature. It is also known as a knowledge based detection approach[16].

1.2.3.2 Anomaly/Statistical based Detection

This approach detection engine identifies something that is a rare and abnormal pattern of behavior[9]. A statistical technique is used to measure is the intrusion.If the behavior of any user slightly differs from that of notifying line than it is an intrusion.

This detection technique is having the ability to detect unknown attacks through the statistical analysis, but the limitation of anomaly based is that. It never detects the well-known attacks.This is because of the fact that it generates a large number of false positive alerts. Its detection engine detects normal and abnormal behavior of user that's why it is also known as behavior based detection approach[11].

Table1.Summary of IDS detection techniques [20]

IDS/IPS technique	**Characteristics/advantages**	**Limitations/challenges**
Signature based detection	• Identifies intrusion by matching captured patterns with preconfigured knowledge base. • High detection accuracy for previously known attacks. • Low computational cost.	• Cannot detect new or variant of known attacks. • High false alarm rate for unknown attacks
Anomaly detection	• Uses statistical test on collected behavior to identifyintrusion. • Can lower the false alarm rate for unknown attacks.	• More time is required to identify attacks. • Detection accuracy is based on amount of collectedbehavior or features.

1.3 Detection Models in Anomaly Detection

As previously discussed there are advantages and limitations of both the techniques. In signature based IDS there is a high accuracy of detecting known attack,but now-a-days security is the main concern in every field and every day a new type of attack is introduced. Therefore,signature based IDS are not able to detect that intrusion and system leads to zero-day attacks. To overcome this problem anomaly detection is the best method to detect the new type of attack on the basis of their behavior. There is some limitation in anomaly detection also but it protects the system from zero-day attacks.

In anomaly detection consist of two phases a) training phase and b) testing phase. On the basis of that behavior is differentiated. In a training phase, we train our dataset about the normal and abnormal traffic profile. After that this training profile is tested on the dataset like KDD Cup 1999 or many more to check the accuracy of the detection approach [23].

Detection models used in anomaly detection [23] [24]. To detect anomalies various different kinds of techniques has been proposed. These Techniques are classified into three categories:

1.3.1 Statistical Anomaly Detection

On the basis of audit records, incoming package and CPU usage etc. profiles are generated to represent the behavior. In these two profiles are maintained: a) current profile and b) stored profile. IDS are continuously updating the current profile periodically according to the capture profile data or behavior of data. So comparison takes place between current profile and stored profiles[24]. Where ever anomaly score is more profile updated accordingly[25]. In this if the anomaly value score crosses the threshold value to that particular statistical line then it generates an alert. Statistical Packet Anomaly Detection Engine (SPADE) is used for the statistical anomaly detection system[26]. This is present in SNORT as a plug-in.

The limitation of this technique is that it purely depends uponthe stored profile data and to model all the behavior is not possible using statistical methods. So we have to update profiles periodically[27].

1.3.2 Machine Learning based Anomaly Detection

Machine learning system havecapability to learn and improve their performance on the basis of certain tasks. This technique is having the capability to change their execution process on the basis of newly learned information. Machine learning focus on increasing the performance on the basis of previous results, but not on understanding the process unlike statistical approaches[24].

Hofmeyr et al. [30] proposed a strategy in which a database is created on the premise of basic table-lookup calculation by taking in the profiles. After that, conduct is weighed in the database and anomalies are detected.

1.3.3 Data mining based Anomaly Detection [24] [27-29]

In this technique the main concerns are with detecting uncovered patterns, anomalies, changes, associations and statistically significant structures. To eliminate the manual process of data profiles or updating of database data mining techniques is widely used nowadays for detecting the anomalies. Data mining has the ability to detect deviation

from normal behavior by creating a boundary value of network activity between normal and abnormal behavior. Data mining technique is categorized into two types:

1.3.3.1 Supervised Learning

It classifies audit data as a normal or abnormal traffic using different classification algorithms. This classification algorithm defines a set of rules and patterns to be used while detecting the intrusion. In this classification method a particular process is followed:

- Identify class and classes, attribute from training data
- Identify attributes for classification
- Learn about a model or algorithm used to train the data.

There are so many classification algorithms are:

- Decision Tree
- Naive Bayes
- Random Forest
- SVM
- Artificial Neural Networks
- Naive Bayes
- Fuzzy Logic
- SOM
- Genetic Algorithm

1.3.3.2 Unsupervised Learning

This is a clustering technique for finding patterns in many directions from unlabeled data using different clustering algorithms. Clustering has ability to detect intrusion in the audit data without an explicit description about various attack classes. So in this training time is automatically reduced. There are two approaches to detect intrusion based on clustering algorithm. First one model is trained using unlabeled data which consist of both types of data i.e normal and abnormal .Second one is the model in which only normal data activities are trained and the rest of them are considered as abnormal.

In the clustering technique distance between the two data points plays an important role. Mostly Euclidean distance metric is used to calculate the distance

between the two data points.So there are so many data mining techniques to classify the normal or abnormal behavior of the user. But these data mining techniques have some limitation so my objective is to reduce these limitations and to improve the accuracy.

1.4 Motivation

Due to this dependence technology is defending against a number of threats to maintain the integrity, confidentiality and availability of the computer system. Hackers, corporate rival, the terrorists and even foreign governments aim to carry out the attack against computer system for their own purposes. The rapid growth of electronic technology and the field of information security plays a very important role in the safety of computer devices. We have to detect intruder by applying proper techniques. But completely prevention from the security breaches using existing technologies is completely unrealistic. IDS plays a major role in network security.

Every day a new type of attack has been introduced to a misuse based detection method. It is not possible to detect new attacks. So anomaly detection plays a vital role while detecting the unknown attack on the basis of their behavior. There are so many different kinds of techniques to detect the anomalies. Data mining techniques, detecting the anomalies with more accuracy and easy to build.

The main motivation of my thesis is to detect anomalies using different data mining techniques. After that the aim is to build a model which is accurate while detecting the unknown attacks. So we have designed a hybrid model based on data mining models. Due to this hybrid model the accuracy of detecting the intrusion has improved automatically.

1.5 Organization of Thesis

There are 5 chapters in this thesis. It is organized as follows:

Chapter 2- This chapter contains a literature survey of existing data mining techniques for anomaly detection. The extensive survey on limitation and features of various data mining techniques.

Chapter 3- This chapter presents the problem statement along with the objectives of thisResearch work.

Chapter 4- This chapter describes a proposed algorithm based on a hybrid algorithm to solve the stated problem.

Chapter 5- This chapter focuses on implementing proposed algorithm and measure the experimental results.

Chapter 6- In this chapter conclusion and future research work is discussed.

Chapter 2

Literature Review

This chapter presents a literature survey of various models and techniques used to detect the intrusion. How IDS developed and various kinds of changes take place in existing and new models.

The IDS notion has been introduced or born in the beginning of 1980 with James Anderson's seminal paper, Computer Security Threat Monitoring and Surveillance. This idea has been approximately for 20 years, but due to the rise of the security infrastructure leads to a dramatic popularity and progress in IDS.

In terms of security various kinds of intrusion detection model have been introduced in the past decades. In which various different kinds of techniques have been introduced which having its own advantage and limitation.

Anderson's *et al.*[17] introduced a term audit trail which includes information for tracking down the misuse and user behavior. Basically, with the release of this paper misuse detection concept gets introduced. His opinion provides a base for IDS design and development.

After that in 1983,SRI International's[18] Computer Science Lab, start working on a government project for IDS development. Their research goal was to explore misuse detection techniques to analyze audit trails. The 1st generation misuses components, analyze system management facility (SMF) records from the IBM mainframe system. Later on it started working on a rule based expert system to detect known intrusion. This early research developed the very first prototype intrusion detection expert system (IDES) on the bases of audit trails.

DE.Denning *et al.*[19] bring a model based on real-time IDES ability to detect intrusion, break-ins and any kind of computer abuse. The model has been focused on detecting abnormal pattern or some kind of security violation of the system of monitoring, audit records. On the basis of detection model profiles have been created to represent the anomalous behavior in the form of metrics and statistical models which helps to obtain information about the behavior of Audit records in future.

This approach has become very useful for detecting intrusions that exploit a system. This approach provides an understanding that how intrusion detected by

practically monitored it. But the major problem was to provide prevention from any kind of exploiting which was caused by a vulnerability in a system. So basically concept was to improve the system security by providing an extra layer of protection of IDES.

U.Lindqvist *et al.* [20] this paper introduces a tool set known as production based expert system tool set (P-BEST). This tool set works for misuse detection and developed a new signature corresponding to the attacks. Basically, this was the advanced version of the IDS that was used for research purpose.

P-BEST provides a better mechanism and performance to detect intrusion in real time environment. This was integrated with the c-programming which make it easy to use, powerful and flexible. But this tool was some limitation. This was less capable of detecting attacks or intrusion where data were incomplete, inaccurate and uncertain. P-BEST was more feasible in the known environment. Tool set was not able to generate new form of attacks in this never know how that particular attack was performed by an attacker.

C.Taylor *et al.* [21] introduce a low cost approach based on clustering and multivariate analysis known as NATE - Network Analysis of Anomalous Traffic Events. This resolves the problem of those IDS who was not able to handle high volume traffic and real time detection constraint. A purpose solution was like any other light weighted approach with the quality feature of minimal network traffic measurement, limited attack scope and anomaly detection. NATE model performed on MIT Lincoln lab's data.

It consists of two phases of operation. In Phase-I data collection and a database creation were performed. In this phase collected data were closely analyzed for possible attack and imagine that only normal data was captured. But in reality if in Phase-I intrusion was treated as a normal than it was a big problem for further detection. While in Phase-II detects intrusion in real-time environment. This classification of normal and abnormal data was performed on the basis of cluster algorithm.

In this paper provide an idea about clustering, which was performed on the real time traffic so that easily and quickly updating of the new features of the attacks in the database.

D.Barbard *et al.* [22] rather than using the traditional method of detecting intrusion a new approach was introduced based on a data mining technique known as Bayesian analysis. This paper creates a test bed with ADAM (Audit data and analysis and mining) which helps to describe and study about various different types of data mining techniques in IDS.

M. Shyu *et al.* [23] introduce a statistical based IDS in which anomaly was treated as outlier. The intrusion predictive model was proposed known as PCA (Principal Component Analysis). This model was constructed on the basis of distance between the occurrences of a normal or anomaly instance. This proposed method was implemented on KDD Cup dataset 1999.

M.Qin *et al.* [24] introduce a new technique of generating frequent episode rules (FER) of categorized internet traffic. This rule was helping to differentiate between various abnormal sequences (like TCP, UDP and ICMP) to normal traffic. Due to increase in rule pruning the search space, time was reduced automatically. Therefore, it was performed with better efficiency and higher detection rate. This proposed approach was also shown better performance when implemented on the real time traffic.

A.Patch *et al.* [25] provide a survey on anomaly and hybrid (misuse and anomaly) detection system which introduce in the past. In this paper also discussed about the recent trends in anomaly detection and determine various problem and challenges faced while detection. This provides a complete description about how novel or zero-day attack detected by anomaly detection and on the basis of rule or knowledge base known attack detected by misuse detection. Paper include a complete explanation about how traditional models different from the existing models.

So to overcome the limitation of existing models, while detecting intrusion data mining techniques was introduced. There were so many different kinds of data mining techniques introduced to secure the network.

W.Lee proposed so many approaches to detect anomalies. The main idea of his proposed methods to use data mining techniques to find out an accurate and robust model using system audit trail data. In 1998W.Lee [26] provides data mining classifiers features help to recognize between anomalies and known intrusion. This was a kind of general and systematic method to detect anomalies. This paper

introduces the first time that how user behavior was detected using data mining techniques. The paper discussed two data mining algorithm:

- Associative rule algorithm
- Frequent episode algorithm

The audit record pattern was used to describe about normal or abnormal behavior. This paper purpose an agent based architecture for IDS that provides real time detection and efficiency.

After that, various other types of techniques have been introduced by using which anomalies were detected. Therefore, there were so many techniques by using them classified network traffic into normal or abnormal traffic.

W.Lee *et al.* **[**27-33] designed a model based on meta-learning mechanism which provides a more adaptive and effective approach of data mining classifiers. In 1999 introduced a distributed architecture for constructing a model which was cost-sensitive while real time detection. The adaptive learning algorithm was used to increase the usability of the system. They also introduced an unsupervised anomaly detection algorithm to reduce the dependency on labeled data. They also introduced a JAM distributed model. In 2001introduced a novel framework model MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection). This model helps to compute patterns from audit data and extracts predictive features using patterns. It was implemented in DARPA dataset 1998. They provide a model based on JAM and MADAM ID which helps to detect fraud in commercial systems. This was a significant and interesting research to prevent our system from any kind of fraud.

V.Chandola *et al.* [34] paper introduces a complete survey of anomaly detection and data mining techniques to prevent computer system from security issues like fraud. In this paper give a complete description about anomalies detection type and their categorization on the basis of their behavior their features and limitations.

Dokas *et al.* [35] in this paper overview about the data mining techniques in which rare class prediction models build for detecting known and unknown attacks. Experimental result performs on KDD Cup dataset 1999 and shows that rare class prediction models more effective and efficient for detecting intrusion rather than standard classification techniques. The rare class prediction model also implemented on the real time traffic on the University of Minnesota where it also performs better. It

also detects novel intrusion which could not be detected using intrusion detection tool SNORT.

An SVM is a supervised learning method. In this classification of data is performed by constructing an N-dimensional hyper plane for creating boundary between the dataset [36] [38]. Classifiers divide the data into two categories. In the training set of instances, labeled pairs {(x, y)}, where y is the label of instance x, SVM performs best classification when they obtain margin is maximum as shown in the below figure [37] [39].

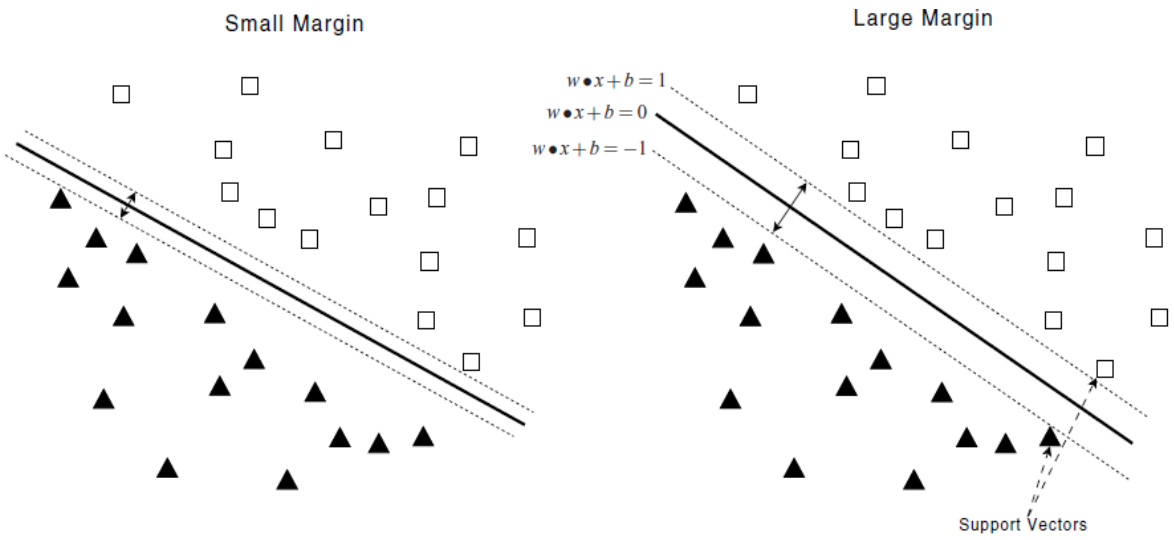

Figure 3.3: Boundary decisions to separate two classes.

In SVM those data point values are 0 then they are non-support vectors. Those whose value is greater than 0 they are support vector [24]. In SVM training complexity depends upon the size of the dataset. So bigger datasets means more training time complexity.

H.Yu *et al.* [36] introduce a new method called Cluster-based SVM (CB-SVM). SVM has several features which were better than other data mining techniques. But in spite of all these features SVM not provide favorable result for large scale dataset because of its training time complexity depend upon the size of the dataset. But in our daily life real-time traffic was around millions or billions of data. So to overcome the complexity of the training data, hierarchical micro-clustering algorithm was used. A hierarchal algorithm provides the best boundary to reduce the training time of SVM.

This approach was only limited to the usage of linear kernels Because of this input space was not maintained in the high dimensional feature space. So this was the limitation of CB-SVM model. So to improve this limitation new form of techniques was introduces.

L.Khan *et al.* [37] in this paper a new clustering approach was used to reduce the complexity while training phase in SVM. They implemented a new clustering technique known as DGSOT(Dynamically Growing Self-Organizing Tree) with SVM. By using this approach it remove the limitation of the traditional hierarchical clustering method. In DGSOT help to find the most qualified data points come as a boundary point while train SVM. This approach was classified dataset into two classes normal or abnormal on the basis of clustering algorithm.

SJ.Horng *et al.* [38] proposed SVM-based IDS, which combines a hierarchical clustering algorithm (BIRCH) with SVM technique. Hierarchical clustering provides a simple feature selection procedure. It helps to reduce the training time for dataset and also improve the performance of SVM.

S.Lin *et al*. [39] the main idea in this approach was to combine the features of three classifiers SVM, DT and SA (Simulated Annealing). In this SVM and SA was used to perform to find out the subset of feature selection for anomaly detection. On another side SA and DT was used to find new decision rules to detect new attacks in a dataset.

W.Lin *et al.*[40] this paper proposed a new novel approach called CANN (cluster center and nearest neighbor). In this approach two distances were calculated and then summed up, first one distance was between its cluster center and each data sample. Second was K-NN distance known as nearest neighbor distance. It is a distance between the data and its nearest neighbor in the same cluster. Experimental results show that CANN not performed better than K-NN.

The limitation of this research method is that CANN didn't give better performance in terms of detecting U2L and R2L attacks. Means that 1-dimensional distance based feature selection performed less efficiently. So to remove this limitation, need to filter out the bad and noisy data which creates a problem while feature selection of a dataset.

ZS.Pan *et al.* [43] in this paper introduce a hybrid approach based on neural network and C4.5.They combined the classification ability of both classifiers for

detecting different attacks. Because neural network is more accurate while detecting DOS and probe attack, but on another side C4.5 is better while detecting R2L and 2R attacks. So both complement each other. This hybrid approach shows better results rather than individual performance of both this classifiers.

G.Stein *et al.* [44] this paper performed a comparison between two classifiers native Bayes networks and decision tree using KDD Cup dataset 1999. Native Bayes and decision tree having their own decision capable to detect the intrusion. Both were performed equally. While detecting U2R and probe native bayes and in normal, DOS and R2L decision tree performs better.

The Bayes network provides a great decision and reasoning under uncertainty. It has strong features independent assumptions due to that it creates an effect on later probability estimation while detecting the intrusion. Whereas decision tree builds models which are easily interpreted. It follows divide and conquer approach to build models, that's why it performs better.

O.Depren *et al.* [45] in this paper a novel or hybrid IDS architecture has been proposed. In which both approaches anomaly and misuse detection utilized. This hybrid architecture consists of both misuse, anomaly detection module and decision support system which combining the result of these two detection modules. The proposed architecture uses self-organizing map (SOM) in anomaly detection module to classify between the normal or abnormal behavior. In misuse detection module J48 decision tree algorithm used to classify between various types of attacks. A rule based decision support system developed to understand the result of both misuse and anomaly detection module. Basically main focus of this hybrid approach to improve the performance as compared to using individual model.

In the future work an idea was introduced in which rather than analyzing the whole dataset classified it into two detection method anomaly and misuse based on network rules and services. Due to that classification of dataset provides better accuracy and less number of the service type attributes.

G.Stein *et al.* [46] introduce the two machine learning technique such as genetic algorithm and decision tree. Genetic algorithm was used to select the appropriate classifier. The non-deterministic approach and initial filtering of input features helps the decision tree classifiers to improve the detection abilities.

The limitation of this process was that it takes more time than decision tree classifier. The training process was done only once, which was creating problems in the future like if a new type of attack behavior and then need to train our dataset accurately.

S.Peddabachigari*et al.*[47] paper provides a new hybrid approach called DT-SVM (Decision trees - SVM) in which two classifiers decision tree and SVM were used as an individual base classifier. The motive of this hybrid technique was to increase the detection accuracy and reduce the computational complexity. This hybrid approach was provided better accuracy than the individual classifier. The paper gives a great idea or a new concept of using multiple classifiers to improve the detection accuracy and reduce the computational complexity.

Wu *et al.* [48] in this paper, research focus on comparing the efficiency of data mining learning techniques like classification tree and SVM on the IDS. It compared various factors like the true positive ratio, false positive ratio and accuracy for four types of attacks like DOS, U2R type, R2L type and probe. Experimental result shows that C4.5 shows better results for detecting DOS, U2R, R2L and probe attack, but SVM was better in reducing the false alarm rate.

M.Ektefa *et al.* [49] in this they works on various different types of data mining techniques like SVM and C4.5 (classification tree) to detect anomaly detection using KDD cup dataset 1999.In this result indicates that classification tree algorithm is better than SVM in detecting the false alarm ratio.

AP.Muniyandi *et al.*[50] a semi-supervised learning technique was used in this paper. Firstly unsupervised learning using K-means clustering. In which part of training instances, was trained using the Euclidean distance method. After that supervised learning performed using C4.5 algorithm algorithm. By applying the clustering the boundary was refined, it helps the C4.5 algorithm to detect anomalies with more accuracy.

This semi-supervised learning technique performs better than unsupervised or supervised learning technique. But limitation was that it takes more time than simple classification or clustering. It was a disadvantage while detecting the real time traffic.

VD.Katkar *et al.* [51] in these multi-category classifiers were used to achieve high accuracy of intrusion detection. The main motive of the research was to use the best features of all classifiers. The paper used Naive Bayesian, Bayesian Network,

SMO, C4.5 and Reduced Error Pruning Tree were used. To reduce error and bad or noisy data, pruning tree classifiers and Bayesian networks were used respectively. By combining all these classifiers together improve the detection ratio.

G.Kim *et al.* [52] introduced a hybrid detection method which was hierarchical integrates the misuse and anomaly detection model. Firstly C4.5 decision tree was implemented to train the dataset after that decomposed into various subsets. Next SVM applied to build the profiles of normal and abnormal behavior. Experimentation was performed on NSL-KDD dataset which was modified version of KDD Cup 1999.This hybrid approach shows the better performance than the conventional models. But in this there was a limitation that C4.5 was degraded while decomposition of data into subsets in misuse detection.

TG.Dietterich *et al.* [53] paper introduces bagging; boosting and randomization technique was used to improve the effectiveness of the decision tree algorithm. Bagging and boosting generate a different range of classifiers by manipulating the training data which provided to the learning algorithm as a base. Bagging shows that if there was a substantial classification noise than randomization technique performs better. This paper describes that bagging performed better than another classifier if there was a noise dataset which is a great advantage of detecting the noise traffic in a real-time environment.

A.Lazarevic*et al.* [54] the aim of this paper that how bagging technique shows a successful improve the accuracy of the classifier in artificial or real-world datasets for anomaly detection. Bagging provides predictive analysis, which improves bad or noisy data improve probabilistic with no-pruning and reduce the variance of unstable method.

M.Galar *et al.* [55] this paper aim was to solve the problem of imbalanced class distribution in the data mining community. This was not even solved by the basic classifiers techniques. So by combining with ensemble technique imbalanced of class distribution was reduced. In this paper bagging or boosting ensemble technique was used. A hybrid approach shows the much better performance and resolve the problem due to imbalanced class attributes.

S.Mukkamala *et al.* [58] in this paper neural network and SVM classifiers was used to detect the anomalies. The main objective of this paper was to create robust, effective and efficient classifiers which detect the intrusion in the real-time .The idea

was to discover patterns or features that describe the user behavior. In this approach both neural network and SVM perform better rather than another technique of classifiers.

J. Zhang *et al.* [59] in this paper a hybrid approach based on both misuse and anomaly detection was implemented on the NIDS. The main objective of this paper was to reduce the limitation of both detection techniques by combining with each other. In this paper for detecting the intrusion random forest data mining techniques were implemented. Firstly random forest implemented for misuse detection in real time. After that it was implemented on the anomaly detection for detecting unknown attacks in an off-line mode. So by combining the two approaches the false alarm ratio and overall performance has been improved.

M. Panda *et al.* [60] in these paper anomalies were detected using native Bayes' techniques. The paper compared the accuracy of the detection rate with the neural network technique. The proposed technique based on the native Bayes approach was performed better rather than neural network approach in terms of false positive rate, Time complexity and cost.

Chapter 3

Problem Statement

3.1 Problem Statement

This chapter consists of discussions about problems that were analyzed during literature survey. It was found that there are some limitations in the existing model so there is a need to improve that limitations.

3.1.1 Problem Formulation

With the increase in science and technology, people are getting more dependent on computer networks for news, stock prices, email and online shopping. Due to this increasing dependency, technology is defending against a number of threats to maintain the integrity, confidentiality and availability of computer system. Now-a-days it has become important to maintain a safe computer system and devices. So IDS plays a major role while detecting intrusion. In misuse detection method only known attacks are detected and on other side anomaly detection method detects unknown attacks also. Anomaly detection can prevent the system from any kind of new or Zero-day attack. Anomalies are detected on the basis of user behavior. There are so many data mining techniques to classify the normal or abnormal behavior of the user. But these data mining techniques have some limitation so the main objective of the thesis is to reduce these limitations and improve the accuracy. Existing techniques having following problems:

- Noisy, meaningless, unrelated or corrupt data is present in a dataset while data collection. Due to the presence of this kind of data classification is never done accurately.
- Overfitting of model due to its excessive parameter comparison to its number of observation creates excessively complex behavior to classify.
- Missing values or missing data have a significant effect on the result while classification. These missing values are extremely important for feature selection while classification. Because of this missing data, conclusions that could be drawn from the data are not accurate.
- Real-time traffic analysis is one of the major concerns that are being faced now-a-days. If detection of real-time traffic is not accurate, then it leads to

compromising the system. So there is a need for the classifiers which are more accurate in detecting intrusion.

- Due to big dataset it becomes difficult to train the classifier which has a direct effect on the accuracy of classifiers.

3.2 Objectives

The objectives of the thesis are as follows:

- To analyze and compare different data mining techniques for the above stated problem.
- To propose a new technique to improve the accuracy and reduce the false alarm rate.
- To improve the detection rate of anomalies.
- To validate the new proposed technique on the dataset.

Chapter 4

Simulation Environment

This chapter discusses about the simulator used in the proposed hybrid model.

4.1 Weka (Waikato Environment for Knowledge Analysis)

Weka is a popular machine learning workbench written in Java [61]. Weka is open source software available under the GNU General Public License (GPL) form last two decades [62]. This was developed at the University of Waikato, New Zealand [63]. Weka allows user to perform and compare different kinds of data mining techniques.

Weka is considered as a landmark of data mining and machine learning. Due to its Graphical User Interface (GUI) and easy access it has achieved a wide acceptance in every field [64] .Weka contains classes which can be accessed by other classes of weka. The most important classes in weka are attribute and instance.An attribute is represented by an object of class attributes which contains attribute types, name,type,nominal values of attributes. [62].

Weka software is not a single program, but it is a collection of algorithm and GUI tool for data analysis and predictive modeling. Older version of weka was in the C language utilities. But now-a-days it is Java-based version [65]. Some of its advantages are:

- GPL: Freely available
- GUI: Ease of use
- Portability: It can run on any modern platform like Windows, Linux, Unix, Mac OS
- It consists of various data processes and modeling techniques.

4.1.1 User Interface

Weka consists of several user interfaces.But the functionality can be performed by any one of them as they give the same result. In weka user interface is classified into four categories[66].

4.1.2 Explorer

This is the main interface of the weka where data mining techniques are implemented. This interface consists of preprocessing, classification,clustering, associative, select attributes and visualizing features[67].

- **Preprocessing:** Data is loaded and converted using Preprocessing algorithm of the weka. This is also known as Filters.
- **Classification:** A second panel of the explorer provides a classification and regression algorithms called as classify.The panel provides a cross-validation of selecting learning algorithm which is by default to estimate predictive performance.
- **Clustering:** A third panel of explorer provides a statistical clustering algorithm. Clustering is performed on the basis of membership or degree of closeness between the data.
- **Associative:** This helps to find out the interrelationship between the attributes using a different associative rules algorithm.
- **Select attributes:** This is used to identify most important attributes in a dataset by accessing and evaluating a wide range of algorithms.
- **Visualize:** The last panel of the explorer. It provides the information about individual data points in a form of color coded scatter plot matrix for visualizing the data.

4.1.3 Knowledge Flow

The Explorer is used only for batch-based data processing where a complete training dataset is loaded into the memory and is then further processed. But it creates problems for the incremental model because of its large dataset.

4.1.4 Experimenter

This interface provides an experimental comparison of different data mining algorithms.Experimenter distributes the multiple dataset to reduce the computational load due to repeated cross-validation for an individual node.

4.1.5 Command Line Interface(CLI)

This is the last and most important interface of weka in perspective of a developer. In this any module can be added manually.

4.2 Netbeans[68][69]

Netbeans is a software tool which provides a platform for developing different applications.The netbeans is written in Java. Application based Netbeans platform includes an IDE extended by third party. Netbeans platforms provide a modular software component known as module.

Netbeans is mostly used for developing Java applications and language like PHP, C++ and HTML5. Netbeans is highly interactive because its GUI provides menus, toolbar and many more options. Netbeans provides a complete tool for developing latest JavaEE6 standards like servlets, Java API,springs,struts and many more features.

Chapter 5

Proposed Work and Implementation Details

This chapter discusses about the proposed work and implementation of the proposed hybrid model. Here the discussion is about the architecture of the proposed model and different components used during the implemention.

5.1 KDD Cup 1999 Dataset

KDD Cup 1999 dataset is the most popular dataset that is used for evaluating the anomaly type intrusion. In 1998 DARPA conducted an evaluation program for intrusion detection in the MIT Lincoln Labs. The main objective of this program was to evaluate the intrusion [70] [71] . In this a standard dataset was provided which simulated a wide variety of intrusions on the perspective of a military network environment. This dataset was prepared by Stotfo *et al.* [71].

DARPA 1998 dataset consist of 4 gigabytes of compressed tcp dump data which was collected over the period of 7 weeks . In this dataset, around 5 million connections and each one connection having 100 bytes were records. KDD Cup 1999 having 4,90,000 single vector connection [72]. This dataset contains 41 features which were labeled as normal or as abnormal. In this dataset attack lies in four categories they are [73-75]:

- **Denial of Service Attack (DOS):** It is a malicious attempt in which the attacker makes server or network resource unavailable or too busy to handle requests. Due to the unavailability of the resources,any other kind of request made by legitimate users gets unprocessed.
- **User to Root Attack (U2R):** It is a type of exploit in which attacker has a local normal user account access on the system. But an attacker takes the advantage of present vulnerabilities in the system like sniffing passwords, a dictionary attack or social engineering and gets the super user privilege access.
- **Remote to Local Attack (R2L):** is a type of attack in which an attacker machine is able to send a packet remotely but does not have an account on

the victim machine. So by taking advantage of any vulnerabilties on the victim machine like clock, guest, xnsnoop, phf, sendmail, the attacker gets access to the victim machine.

- **Probing Attack:** It is a very basic and initial step of exploiting any system.The attacker scans a machine to find out the weakness or vulnerabilities in the network using saint, portsweep, mscan, nmap etc. To exploit the victim machine.

In real time traffic, data are not distributed with the same probability as in KDD Cup 1999.But KDD Cup dataset helps to know about the type of attacks. Every day a new type of attack is introduced so to detect that particular attack known detection methodsare not feasible. Hence the researchers concluded out an anomaly detection method to be useful for detecting abnormal data. KDD Cup1999 consists of 24 training attack types,with the 14 test data additional. The detailed description about various training attack types is[68]:

- **Basic features:** It consists of all attributes which are extracted from a TCP/IP connection.This type of connection takes more time to detect because in this classification distinguishing between the normal or abnormal TCP/IP is bit complicated.
- **Traffic features:** In this technique features are computed with respect to the window interval of the packet. It categorizes the traffic on the basis of two features namely:
  - **"Same host" features:** It examines the time period of connection between server host and destination host.After that calculates the behavior of that particular destination host.
  - **"Same service" features:** It examines the services utilized by the destination host and keep tracking that all resources can never get utilized by a single host.
- **Content features:** There are no frequent sequential patterns of intrusion so as to know about the frequency of the different attacks like DOS, Probing, R2L and U2R attacks.To measure out DOS and probe attack resource utilization should be taken care of. However, on the other side in U2R and R2L attack

malicious data is embedded in the data traffic.So in this type of attack of content received from the destination host is taken care.

5.2 Architecture of Proposed Model

In the proposed method for preprocessing of data an unsupervised filter is used and for the classification of data C4.5 and Bagging algorithm is used. In the proposed scheme consists of various phases through which data processed to classify the anomalies in a network traffic. These are described as follows:

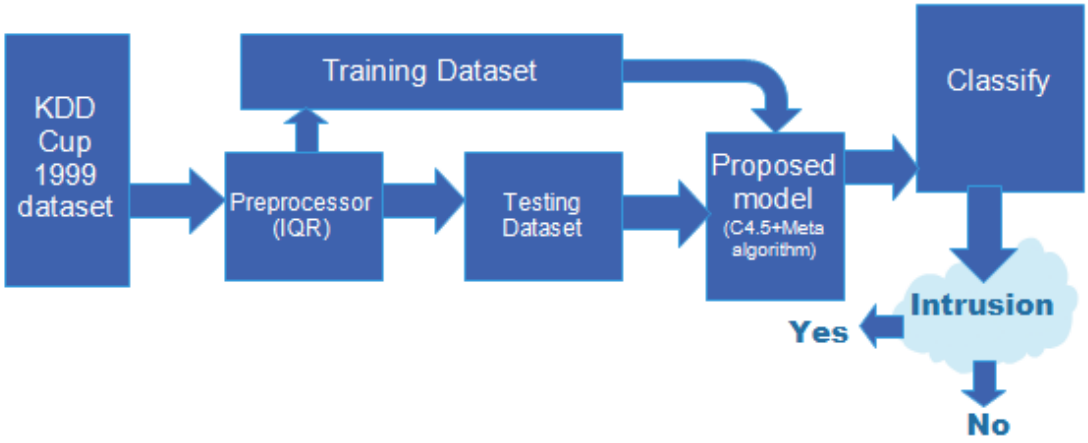

Figure 4.1 Architecture of Proposed model

Preprocessor: While collecting real-time traffic it includes some irrelevant , redundant unreliable data ,which is unrelated for knowledge discovery known as noisy data. Due to this noisy data there is effect on the result of data during the classification of normal and abnormal traffic. Therefore, to improve the detection capabilities of classifiers, there is a need to preprocess the data. Preprocessing provides a feature to clean and normalize the data so that any kind of irrelevant data never affects the accuracy of classifiers. To remove the redundant data there are various kinds of preprocessing filters which are available.

In the proposing method INTERQUARTILE RANGE (IQR) filter is used for preprocessing thedata. This is also known as mid spread or middle fifty because in

this statistical dispersion is measured equal to the difference between lower and upper quartiles [75].

Outlier calculation using IQR:

1. Arrange dataset in order.
2. Calculate the value of first quartile (Q1)
3. Calculate the value of the third quartile (Q3)
4. Calculate the value of inter quartile range (IQR)=Q3-Q1
5. Calculate the value of lower boundary= Q1-(1.5*IQR)
6. Calculate the value of upper boundary= Q3+(1.5*IQR)
7. Anything outside from the lower and upper boundary values is an outlier

After applying the preprocessor filter noisy data was normalized and two new attributes outlier and extreme. Now dataset contains 44 attributes with outlier and extreme were created attributes. It helped the training set to provide more accuracy. In the next phase test dataset, go under classification.

C4.5 Algorithm: C4.5 is a typical decision tree algorithm. It is basically an extension of ID3 algorithm. This was developed by Ross Quinlan [41]. C4.5 is also known as statistical or supervised learning classifier. In this algorithm, divide and conquer approach is used to recursively construct a decision tree of database/dataset attributes [26]. Decision tree is built by using two methods i.e top down manner or bottom up manner. A decision tree is constructed with empty root node and then decision tree node corresponding to the algorithm is built [40] [52].

To construct the decision tree following steps is followed recursively. Following are the steps [49]:

1. Information gain (IG) is computed for each and every attribute.
2. A tree like structure gets started constructing from its root node by the selection of attributes which have a maximum value of IG.
3. If the attributes are more discrete than its branch with all possible values and if they are continuous than maximum value is selected on the basis of G or cut point.
4. After splitting, consider whether or not these new nodes are leaves (their data belongs to the same type); otherwise, new nodes are the root of the sub-trees.
5. Repeating all the above steps, until all new nodes are leaves.

This partition of the training dataset is performed and a tree like structure is created. Due to selection of highest IG value it is best suited for decision making tree. Let D is a training set which contains m distinct classes, Ci(i=1,2….m) Ci,D set of tuples in class Ci in D.|Ci,D| and |D| represent the number of tuples in D and Ci,D respectively and Pi is the probability that any tuple present in D belongs to class Ci, which is calculated by |Ci,D| / |D|. When an attribute A is selected, it can be used to split D into v partitions {D1,D2…..Dv}where Dj contains tuples in D that have outcome aj of A.There are some method for selecting best or splitting attributes as follows:

$$Info(D) = -\sum_{i=1}^{m} P_i \log_2 Pi \quad (1)$$

$$Gain(a,T)=Entropy(a)-Info(a,T) \quad (2)$$

$$Info\ (a,\ T)\ +=\ \frac{T_{a,\,v}}{T_a}\ Entropy(a_v) \quad (3)$$

$$GainRatio(a,T)=\frac{Gain(a,T)}{SplitInfo\,(a,T)} \quad (4)$$

$$SpiltInfo\ (a,\ T)\ +=\ -\frac{|\,T_{a,\,v}\,|}{|\,T_a\,|}\log\frac{|\,T_{a,\,v}\,|}{|\,T_a\,|} \quad (5)$$

Features of C4.5 algorithm

- This is available as open source in a WEKA interface for java as J48.[45][51].
- This algorithm builds the model which is easily interpreted.
- This algorithm uses both categorical and continuous values as a numerical form.
- This provides a method called imputation, which deals, with the missing values. If in a dataset an important feature is missing than by using an imputation method the problem of missing values can be resolved. In this method missing values are filled by estimating from the available data .
- This algorithm is known for its classification of training data which helps tocorrectly classify the test data because of to its tree pruning method. Tree pruning method helps to build small tree and avoid over fitting of data.
- This reduces the classification error by replacing a sub-tree with leaf performing sub-tree replacement.

Limitation of C4.5 algorithm

- A small change in the dataset leads to complete variation in decision tree formation.
- Performance is not good where the dataset is small.

Bagging or Bootstrap Aggregating [41] [43] [44]: Bagging is based on statistical classification and regression technique in the data mining algorithm by Leo Breiman.This is a machine learning ensemble meta-algorithm used to improve the stability and accuracy.It creates different samples for training dataset and creates a classifier for each sample dataset .The result is the combination of these multiple classifiers. It also reduces variance, classification noise and helps to avoid overfitting by introducing randomization while its construction procedure and after that ensemble meta-algorithm. Mostly it applies to decision tree methods. But can also be used with any type of classifier. Bagging provides a predictive probabilistic model which improves bad or noisy data. In bagging to construct the Random samples and features sets following steps have to follow recursively [56][57]:

- Train the training set by replacing the N sample using randomly sample.
- Test for each trained base on the dataset to check for more accuracy.

Advantage

- Improved accuracy for over fitting dataset.

Disadvantage

- Computational complexity.
- Loss of interpretability.

Proposed Algorithm:A proposed model is based on the C4.5 algorithm and meta algorithm.As later discussed that C4.5 algorithm is more accurate making any kind of decision.But problem is while any kind of change occur in the dataset affects on the decision making and needs to train the dataset .It leads to variance in the classifiaction of data.So to remove this variance meta algorithm introduce with the C4.5 algorithm.In this proposed algorithm it provides a predictive feature using randomization which improve the variance while decision tree formation in C4.5 algorithm.The pseudo code of proposed algorithm is:

Input:

- Data partition D_i which is a set of training tuples and their associated class labels.
- Attribute_list ,the set of candidate attributes
- Attribute_selection_method, a procedure to determine the splitting criterion the "best" partitions the data tuples into individual classes .This criterion consists of a splitting_attribute and possible either a split_point or splitting subset.
- K_i the no. of models in the ensemble;

Output: A meta_decision tree:

Method:

1. Create a node N;
2. **If** tuples in Dare all the same class C_i ,**then**
3. **Return** N as a leaf node labeled with the class C;
4. **If** attribute_list is empty **then**
5. **Return** N as a leaf node labeled with the majority class in D; //majority voting
6. Apply attribute_selection_method(D_i attribute_list) to find the "best" splitting_criterion;
7. Label node N with splitting_criterion;
8. Create bootstrap sample D_i by sampling D with replacement;
9. Use D_i and the learning schema to derive a splitting_attribute model;
10. **If** splitting_attribute is discrete_valued and multiway splits allowed then // not restricted to binary trees
11. Attribute_list ←attribute_list_splitting attribute;//remove_splitting_attribute;
12. **For** each outcomes j of splitting_criterion; //partition the tuples and grow subtrees for each partition
13. Let D_j the set of data tuples in D satisfying outcome j;//a partition
14. **If** D_j is empty **then**
15. Attach a leaf labeled with the majority class in D to node N;
16. **Else** attach the node return by generate_decision_tree(D_j ,attribute_list) to node N;

End for

17. **Return** N;

Chapter 6
Experimental Results

The proposed approach is based upon Decision tree and meta algorithms with a selection of 41 features in a dataset. In this hybrid approach, application of the meta algorithm at the building time of decision tree helps to reduce the variance while decision tree improves the accuracy of anomaly detection. To evaluate the proposed model it compares different supervised machine learning models and measure their accuracy, false alarm rate and true positive rate, etc. the proposed approach compares popular machine learning models, i.e. decision tree, native bayes, support vector machine (SVM), decision table and random tree. A brief description about all these models already has been given in the literature survey. To measure the accuracy of these predictive models, *k*-fold cross validation is used, where *k* is an integer. Here 10 is used as value for k. We randomly partition, our samples into 10 subsamples. Out of these 10 subsamples, 4 samples used for testing and remaining 6 subsamples used for training. The purpose of doing this is that, every subsample is used for both the training and testing. The robustness or correctness of our proposed model are measured using different parameters.

Accuracy: This refers to the ability of the classifiers to correctly measure the intrusions from the training dataset. This is defined as the ratio of correctly classified data to the total classified data.

$$\text{Accuracy}= \frac{TP+TN}{TP+FP+TN+FN} \tag{6}$$

Experimentation result shows that the proposed model is more accurate as compare to other data mining techniques. The accuracy of a proposed model is near about 100% as shown in the graph. This proposed method performs better than individual performance of the J48 (C4.5).

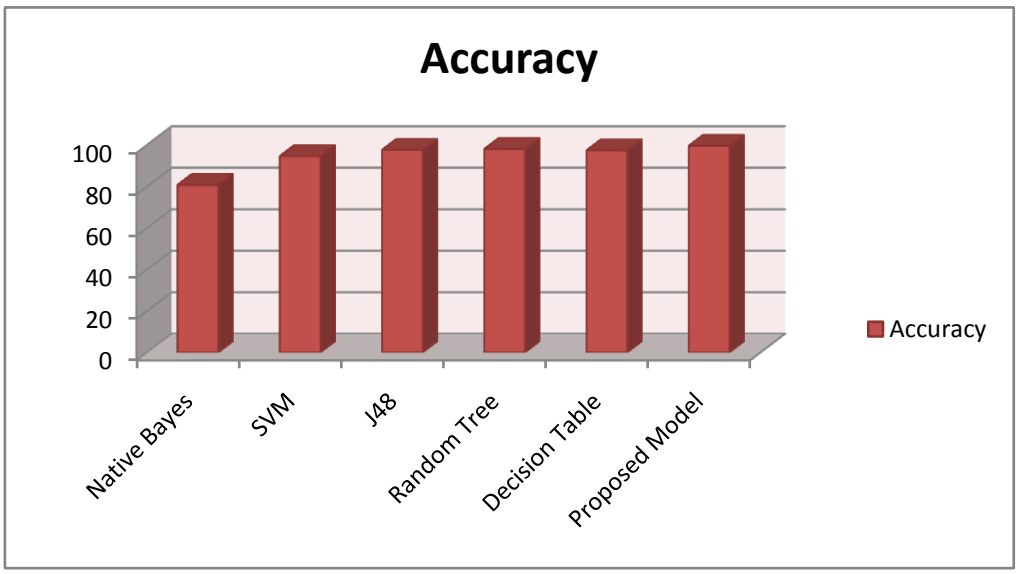

Figure 6.1 Accuracy analyses of different models

Detection Ratio: It is defined as the ratio of detecting attacks to total no of attacks. This is the best parameter to measure the performance of the model.

$$\text{Detection Ratio} = \frac{TP}{TP+FN} \qquad (7)$$

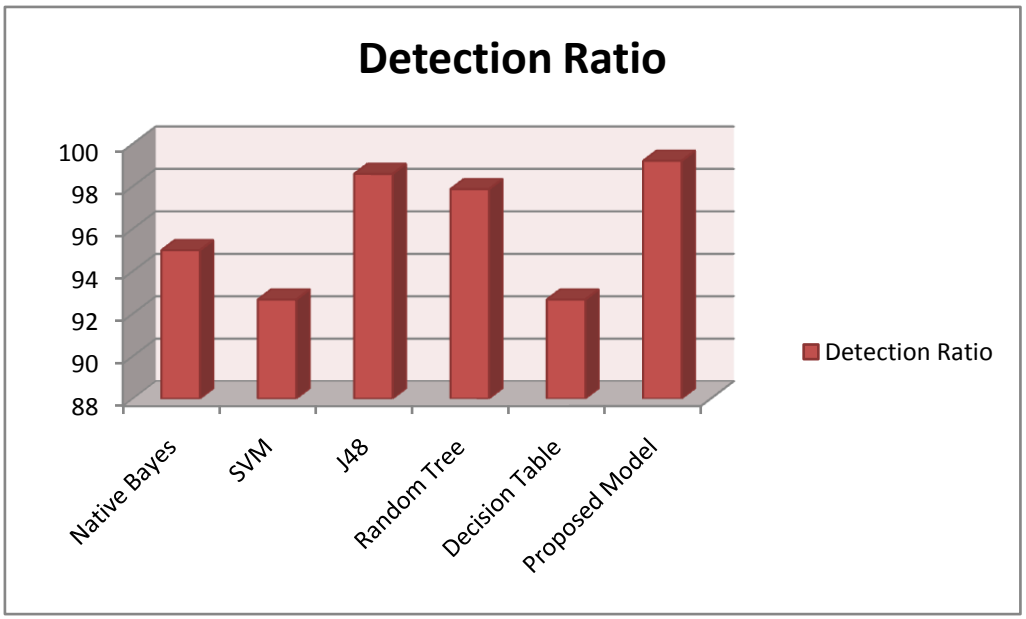

Figure 6.2 Detection Rate analysis of different models

Detection ratio means correctness in a model for detecting intrusion. Here also experimental result shows that the proposed algorithm performs better in term of correctness in detecting intrusion.

In above experimentation, the result shows the average performance of 10-Cross validation. These models are compared on the basis of each individual fold or rounds. To measure the robustness and effectiveness of any model, comparison of different parameters like False Positive Rate, True Positive Rate, F-measure, Precision and Recall is computed and the performance of different models on the above parameters is evaluated.

True Positive Ratio: This is one in which correct classification of data has been performed. Means correctness in a system to detect normal or abnormal data. It is defined as:

$$\text{TPR} = \frac{TP}{TP + FN} \tag{8}$$

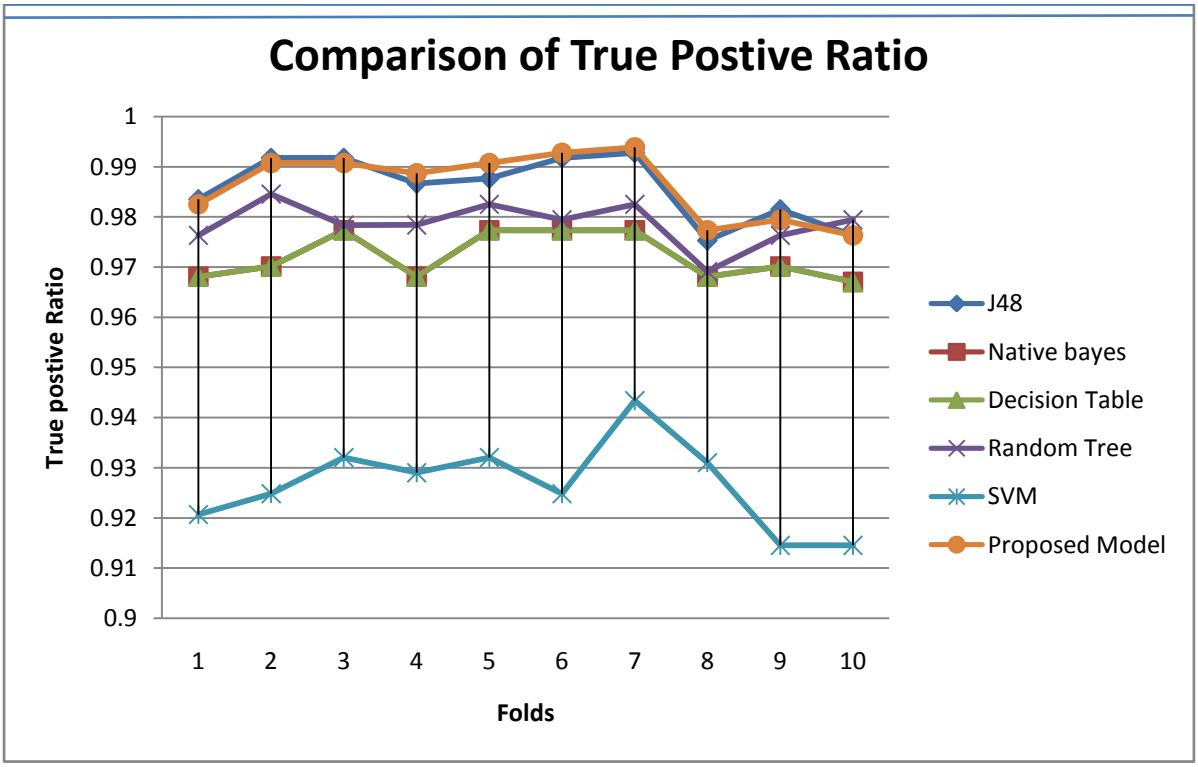

Figure 6.3 True Positive Ratio analyses of different models under 10 folds

False Positive Ratio: This is one of the main parameters to find out the effectiveness of various models and also the major concern while network setup. A normal data is considered as abnormal or attack type data. It is defined as:

$$FPR = \frac{FP}{TN+FP} \qquad (9)$$

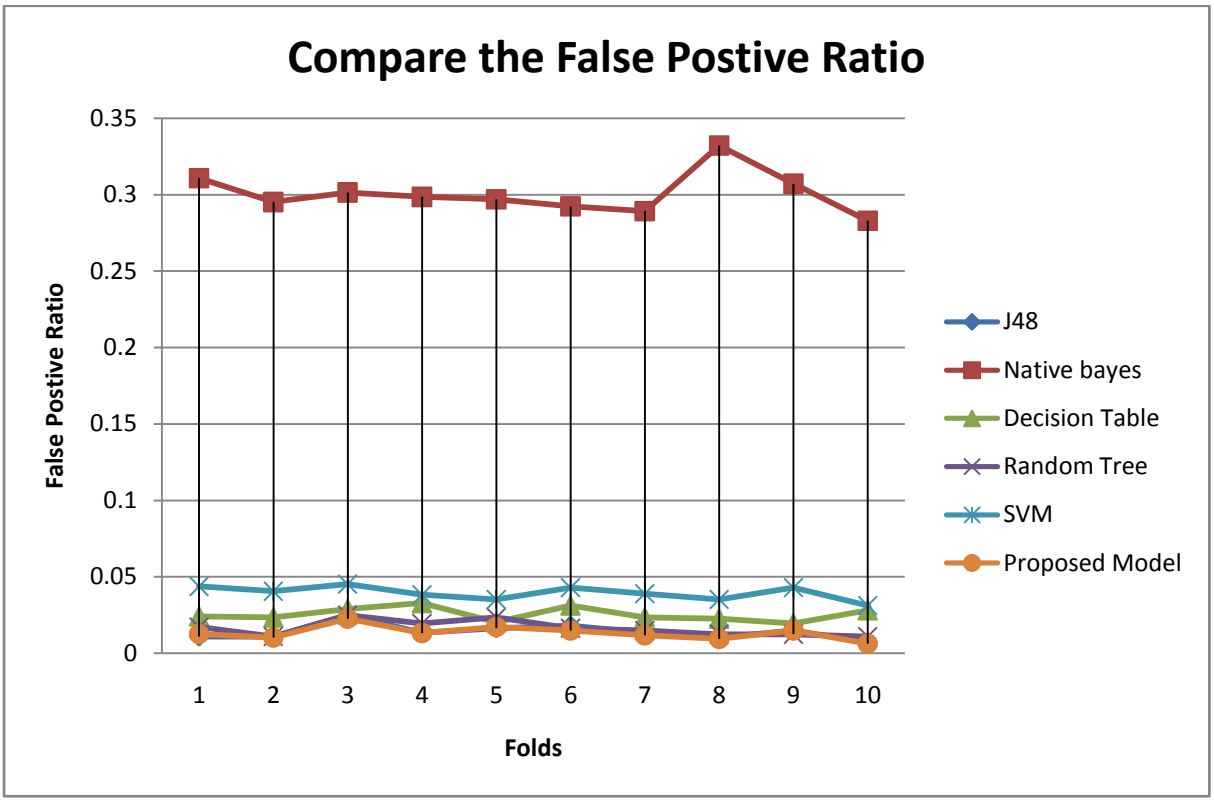

Figure 6.4 False Positive Ratio analyses of different models under 10 folds

As shown in the comparison, both true positive rate and false positive rate proposed model perform better as compared to other models. These two parameters are very important measure to evaluate the performance of a model. Hence results show that the proposed model performs better than other models.

Precision Ratio: It is also known as Positive Predictive Value (PPV). It measures the relevant instance that is retrieved after classification. A classifier that has high precision means that classifiers or algorithm returns more relevant results.

$$PPV = \frac{TP}{TP+FP} \qquad (10)$$

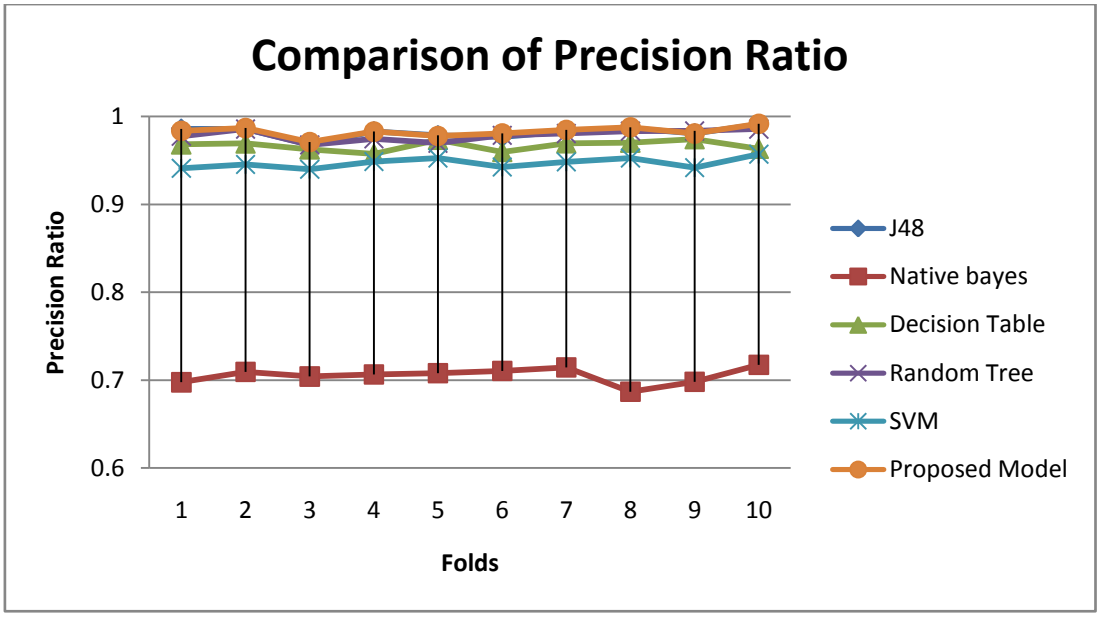

Figure 6.5 Precision Ratio analyses of different models under 10 folds

As shown in the figure, precision ratio of the proposed model is high as compared to other models. Proved that proposed model provides the less relevant results.

Recall: It is also known as sensitivity. This is also used to measure the relevant instance, that is selected. The higher value of recall more the relevant data is selected for classification. It is defined as:

$$Recall = \frac{TP}{TP+FN} \qquad (11)$$

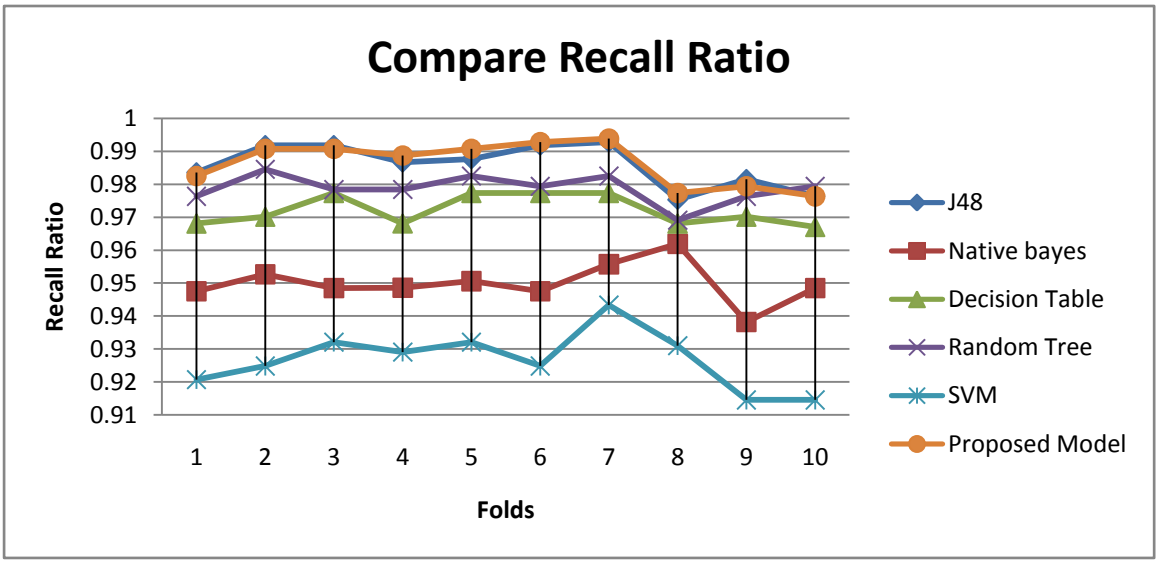

Figure 6.6 Recall Ratio analyses of different models under 10 folds

The above figure shows that the proposed model is having a high recall or sensitivity. Hence, the most relevant data is selected as compared to other classifiers.

F-Measure: It is basically used to measure the effectiveness of the classifiers. This is harmonic mean of precision and recall. It is also known as traditional F-measure or balanced F-score. It is defined as:

$$\text{F-measure} = 2\frac{precison.*recall}{\text{Pr}ecision + recall}$$

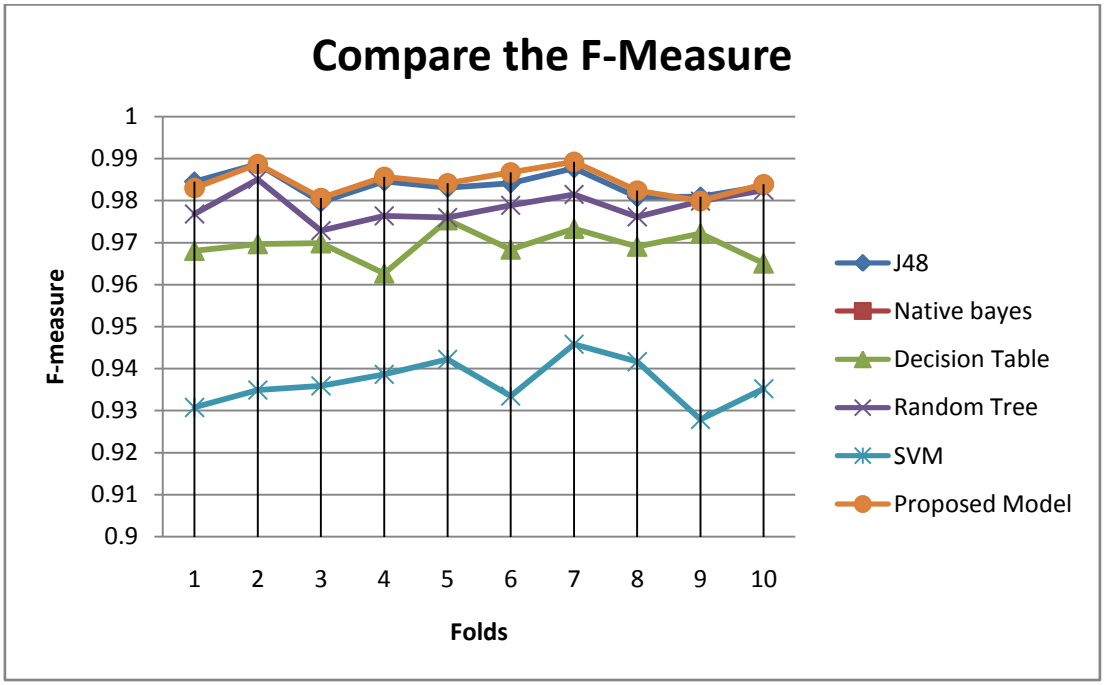

Figure 6.7 F-Measure analyses of different models under 10 folds

The above figure shows that the proposed model shows better results in every fold or round of evaluation. A proposed technique performs much better in all aspects of the evaluation parameter of anomaly detection.

Area under ROC curve: It defined the correctness of the classifier that how a normal or abnormal dataset is separated by using training dataset. More the area under the ROC curve the more accurate the classifier is.

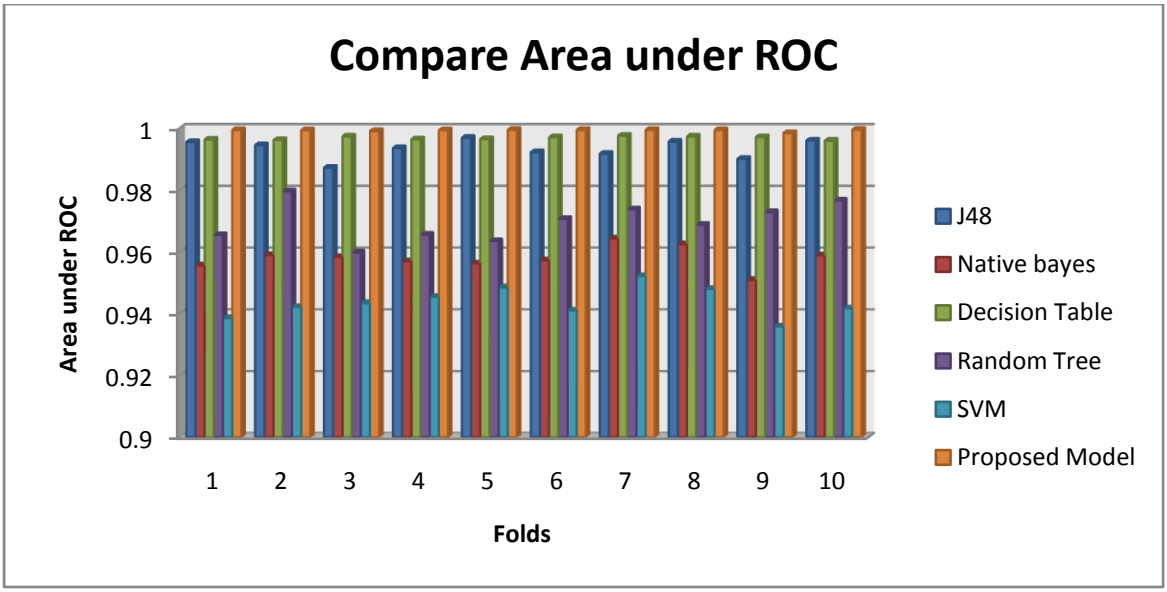

Figure 6.8 Comparisons of different models under Area under ROC in 10 folds

As above figure shows that the proposed algorithm covers the maximum area means the maximum accuracy in the result while classification.

Area Under PRC: This is a graph between the recall and precision value. In this, it defined the correctness. Area under PRC means that classifiers are providing more correctness while classification.

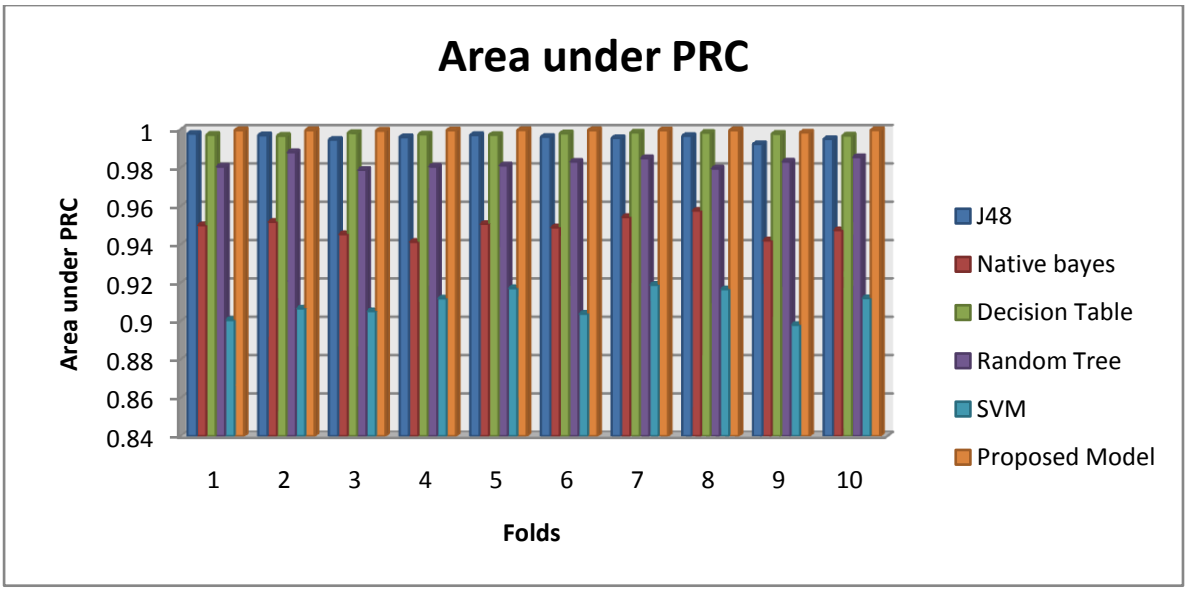

Figure 6.9 Comparisons of different models under Area under PRC in 10 folds

The above figure shows that the proposed algorithm covers the maximum area means the maximum correctness in the result while classification.

From all the above experimentation results, it is shown that after applying all the evaluation parameters, proposed model found to be the best model in all scenarios. By applying the hybrid approach of data mining models on the dataset, the detection rate is improved for anomaly detection. So the main objective to improve the detection rate in anomaly detection has been met.

Chapter 7

Conclusion and Future Scope

7.1 Conclusion

Nowadays prevention of security breaches using the existing security technologies is unrealistic. As a result, intrusion detection is an important component in network security. Also, misuse detection technique cannot detect unknown attacks so the anomaly detection technique is used to identify these attacks. To improve the accuracy rate of intrusion detection in anomaly based detection data mining technique is used.

In this thesis a hybrid approach using data mining is implemented which is an amalgam of two different techniques namely decision trees and Meta algorithm. The results of the proposed approach are compared with the results of other data mining techniques and it outperforms them. The new approach is effective during detection of attacks. The detection ratio of the proposed algorithm is better than other techniques.

The proposed approach properly classifies the data either as normal or abnormal. The detection ratio with proposed technique is 98.70% ,which is better than all other techniques. Accuracy is also improved by the use of filters which removed the meaningless, noisy and unrelated data from the original data. It can be concluded that this approach is simple and efficient in terms of reducing the false alarm ratio.

7.2 Future Scope

In the proposed approach data are classified into normal and abnormal data. The approach can be improved by classifying data further into the subclasses such as DOS, Probe, U2R and R2L. Better results were obtained while performing it on KDD Cup 1999 Dataset. It can also be used for real time traffic analysis for obtaining better results.

While achieving the detection accuracy of data traffic in real time anomaly detection, better results can be obtained by combining it with supervised and unsupervised techniques.While analyzing the real time traffic for supervised learning method is a bit complicated due to the data size. To overcome this problem unsupervised learning is used to define the boundaries between normal and abnormal

data. So after that when supervised learning model is applied to real time traffic, then it gets easily classified without any considerable delay.

References

[1] H. Debar, M. Dacier, and A. Wespi, "A revised taxonomy for intrusion-detection systems," In *Annales des telecommunications, Springer-Verlag*, vol. 55, no. 7-8, pp. 361-378, 2000.

[2] A.S. Ashoor and S. Gore, "Importance of Intrusion Detection system (IDS)," *International Journal of Scientific and Engineering Research*, vol. 2, no. 1, pp. 1-4, 2011.

[3] P.G. Majeed and S. Kumar, "Genetic algorithms in intrusion detection systems: A survey," *International Journal of Innovation and Applied Studies*, vol. 5, no. 3, pp. 233-236, 2014.

[4] F. Alserhani, M. Akhlaq, I. U. Awan, A. J. Cullen, J. Mellor, and P. Mirchandani, "Snort Performance Evaluation," In *Proceedings of Twenty Fifth UK Performance Engineering Workshop (UKPEW 2009), Leeds, UK*, 2009.

[5] B. Khosravifar and J. Bentahar, "An experience improving intrusion detection systems false alarm ratio by using honeypot," In *22nd International Conference on Advanced Information Networking and Applications, AINA*, IEEE, 2008, pp. 997-1004.

[6] J. Kim and P.J. Bentley, "Towards an artificial immune system for network intrusion detection: an investigation of dynamic clonal selection," In *Proceedings of the IEEE 2002 Congress onEvolutionary Computation, CEC*, IEEE, 2002, vol. 2, pp. 1015-1020.

[7] H.T. Elshoush and I.M. Osman, "Reducing false positives through fuzzy alert correlation in collaborative intelligent intrusion detection systems—A review," In *IEEE International Conference on Fuzzy Systems (FUZZY), IEEE*, 2010, pp. 1-8.

[8] H. Njogu, L. Jiawei, J. Kiere and D. Hanyurwimfura, "A comprehensive vulnerability based alert management approach for large networks", *Future Generation Computer Systems*, vol. 29, no. 1, pp. 27-45, 2013.

[9] A. Patel, M. Taghavi, K. Bakhtiyari, and J.C. Júnior, "An intrusion detection and prevention system in cloud computing: A systematic review," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 25-41, 2013.

[10] T. Pietraszek and A. Tanner, "Data mining and machine learning—Towards reducing false positives in intrusion detection," *Information Security Technical Report*, vol. 10, no. 3, pp. 169-183, 2005.

[11] H. Highland, "A software architecture to support misuse intrusion detection", *Computers & Security*, vol. 14, no. 7, pp. 607, 1995.

[12] K. Ilgun, R. Kemmerer, and P. Porras, "State transition analysis: A rule-based intrusion detection approach," *IEEE Transactions on Software Engineering*, vol.21, no. 3, pp. 181-199, 1995.

[13] S. Kumar, "Classification and detection of computer intrusions," *PhD dissertation*, Purdue University, 1995.

[14] K. Kent and P. Mell, *Guide to intrusion detection and prevention systems (IDPS)*. Gaithersburg, MD: U.S. Dept. of Commerce, Technology Administration, National Institute of Standards and Technology, 2007.

[15] A.K. Jackson, "Intrusion detection system (IDS) product survey," *Los Alamos National Laboratory, Los Alamos, NM, LA-UR-99-3883 Ver* 2, pp. 1-103, 2002.

[16] P. Kabiri and A.A. Ghorbani, "Research on Intrusion Detection and Response: A Survey," *IJ Network Security*, vol. 1, no. 2, pp. 84-102, 2005.

[17] J.P. Anderson, "Computer security threat monitoring and surveillance," *Technical report*, James P. Anderson Company, Fort Washington, Pennsylvania, vol. 17, 1980.

[18] C. Endorf, E. Schultz and J. Mellander, *Intrusion detection & prevention*. New York: McGraw-Hill/Osborne, 2004.

[19] D.E. Denning, "An intrusion-detection model," *IEEE Transactions on Software Engineering*, vol. 2, pp. 222-232, 1987.

[20] U. Lindqvist and P.Porras, "Detecting computer and network misuse through the production-based expert system toolset (P-BEST)," In *Proceedings of the 1999 IEEE Symposium on Security and Privacy,IEEE,* 1999, pp. 146-161..

[21] C.Taylor and J. Alves-Foss, "Nate: Network analysis of anomalous traffic events, a low-cost approach," In *Proceedings of the 2001 workshop on New security paradigms*, ACM, 2001, pp. 89-96.

[22] D. Barbará, J. Couto, S. Jajodia and N. Wu, "Audit Data Analysis and Mining",*ACM SIGMOD Record*, vol. 30, no. 4, p. 15, 2001.

[23] M.L. Shyu, S.C. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier," *Miami University Coral Gables FL Deptarment of Electrical and Computer Engineering*, 2003.

[24] M. Qin and K. Hwang, "Frequent episode rules for internet anomaly detection," In *Third IEEE International Symposium on Network Computing and Applications (NCA)*, IEEE, 2004, pp. 161-168.

[25] A. Patcha and J. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448-3470, 2007.

[26] W. Lee and S.J. Stolfo, "Data mining approaches for intrusion detection," In *Usenix Security*, 1998.

[27] W. Lee, S.J. Stolfo, and K.W. Mok, "A data mining framework for building intrusion detection models," In *Proceedings of the 1999 IEEE Symposium on Security and Privacy, IEEE*, 1999, pp. 120-132.

[28] W. Lee, S.J. Stolfo, and K.W. Mok, "Adaptive intrusion detection: A data mining approach," *Artificial Intelligence Review*, vol. 14, no. 6, pp. 533-567, 2000.

[29] W. Lee, S.J. Stolfo, and K.W. Mok,, "Mining Audit Data to Build Intrusion Detection Models," 1998, pp. 66-72.

[30] W. Lee and S.J. Stolfo, "A framework for constructing features and models for intrusion detection systems," *ACM transactions on Information and system security (TiSSEC)*, vol. 3, no. 4, pp. 227-261, 2000.

[31] W. Lee, S.J. Stolfo, and K.W. Mok, "Mining in a data-flow environment: Experience in network intrusion detection," In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 1999, pp. 114-124.

[32] S.J. Stolfo, W. Lee, P.K. Chan, W. Fan, and E. Eskin, "Data mining-based intrusion detectors: an overview of the columbia IDS project," *ACM SIGMOD Record*, vol. 30, no. 4, pp. 5-14, 2001.

[33] S.J. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P.K. Chan, "Cost-based modeling for fraud and intrusion detection: Results from the JAM project," In *Proceedings DARPA Information Survivability Conference and Exposition, DISCEX'00*, IEEE, vol. 2, 2000, pp. 130-144.

[34] V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection", *ACM computing surveys*, vol. 41, no. 3, pp. 1-58, 2009.

[35] P. Dokas, L. Ertoz, V. Kumar, A. Lazarevic, J. Srivastava, and P.N Tan, "Data mining for network intrusion detection," In *Proceeding NSF Workshop on Next Generation Data Mining*, pp. 21-30, 2002.

[36] H. Yu, J. Yang, and J. Han, "Classifying large datasets using SVMs with hierarchical clusters," In *international conference on Knowledge discovery and data mining Proceedings of the ninth ACM SIGKDD*, ACM, 2003, pp. 306-315.

[37] L. Khan, M. Awad, and B. Thuraisingham, "A new intrusion detection system using support vector machines and hierarchical clustering," *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 16, no. 4, pp. 507-521, 2007.

[38] Horng, M. Su, Y. Chen, T. Kao, R. Chen, J. Lai and C. Perkasa, "A novel intrusion detection system based on hierarchical clustering and support vector machines," *Expert Systems with Applications*, vol. 38, no. 1, pp. 306-313, 2011.

[39] S. Lin, K. Ying, C. Lee and Z. Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection," *Applied Soft Computing*, vol. 12, no. 10, pp. 3285-3290, 2012.

[40] W. Lin, S. Ke and C. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors",*Knowledge-Based Systems*, vol. 78, pp. 13-21, 2015.

[41] J. Quinlan, *Bagging, Boosting, and C4. 5*. 1996.

[42] J. Quinlan, *C4.5*. San Mateo, Calif.: Morgan Kaufmann Publishers, 1993.

[43] K. Wang and J. Wu, "Machine Learning and Cybernetics," In *International Conference on Springer*, Springer, vol. 3, 2003, pp. 1583-1585.

[44] N.B. Amor, S. Benferhat, and Z. Elouedi, "Naive bayes vs decision trees in intrusion detection systems," In *Proceedings symposium on Applied computing of the ACM*, ACM, 2004, pp. 420-424.

[45] O. Depren, M. Topallar, E. Anarim and M. Ciliz, "An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks," *Expert Systems with Applications*, vol. 29, no. 4, pp. 713-722, 2005.

[46] G. Stein, Gary, B. Chen, A.S. Wu, and K.A. Hua, "Decision tree classifier for network intrusion detection with GA-based feature selection," In *Regional*

Conference Proceedings of the 43rd Annual Southeast, ACM, 2005, vol.2, pp. 136-141.

[47] S. Peddabachigari, A. Abraham, C. Grosan and J. Thomas, "Modeling intrusion detection system using hybrid intelligent systems," *Journal of Network and Computer Applications*, vol. 30, no. 1, pp. 114-132, 2007.

[48] S. Wu and E. Yen, "Data mining-based intrusion detectors", *Expert Systems with Applications*, vol. 36, no. 3, pp. 5605-5612, 2009.

[49] M. Ektefa, S. Memar, F. Sidi, and L.S. Affendey, "Intrusion detection using data mining techniques," In *International Conference on Information Retrieval and Knowledge Management,(CAMP),* IEEE, 2010, pp. 200-203.

[50] A. Muniyandi, R. Rajeswari and R. Rajaram, "Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm," *Procedia Engineering*, vol. 30, pp. 174-182, 2012.

[51] V.D. Katkar and S.V. Kulkarni, "Experiments on detection of Denial of Service attacks using ensemble of classifiers," In *International Conference on Green Computing, Communication and Conservation of Energy (ICGCE),* IEEE, 2013, pp. 837-842.

[52] G. Kim, S. Lee and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1690-1700, 2014.

[53] T.G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine learning*, vol. 40, no. 2, pp.139-157, 2000.

[54] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," In *International Conference on Knowledge Ddiscovery in Data Mining Proceedings of the eleventh ACM SIGKDD*, ACM, 2005, pp. 157-166.

[55] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463-484, 2012.

[56] W. Dai and W. Ji, "A MapReduce Implementation of C4.5 Decision Tree Algorithm," *International Journal of Database Theory and Application*, vol. 7, no. 1, pp. 49-60, 2014.

[57] S. Li and A. Jain, *Encyclopedia of biometrics*. New York: Springer, 2009..

[58] C. Manikopoulos and S. Papavassiliou, "Network intrusion and fault detection: a statistical anomaly approach," *IEEE Communications Magazine*, vol. 40, no. 10, pp. 76-82, 2002.

[59] J. Zhang and M. Zulkernine, "A hybrid network intrusion detection technique using random forests," In *First International Conference on Availability, Reliability and Security, ARES*, IEEE, *2006*, pp. 8-15.

[60] M. Panda and M.R. Patra, "Network intrusion detection using naive bayes," *International journal of computer science and network security*, vol. 7, no. 12, pp. 258-263, 2007.

[61] I. Witten and E. Frank, *Data mining*. San Francisco, Calif.: Morgan Kaufmann, 2000.

[62] R.R. Bouckaert, E. Frank, M.A. Hall, G. Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, "WEKA Experiences with a Java Open-Source Project," *The Journal of Machine Learning Research*, vol. 11, pp.2533-2541, 2013.

[63] G.Holmes, A. Donkin, and I.H. Witten, "Weka: A machine learning workbench," In *Proceedings of the Second Australian and New Zealand Conference on Intelligent Information Systems*, IEEE, 1994, pp. 357-361.

[64] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, "The WEKA data mining software", *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, p. 10, 2009.

[65] S.R. Garner, "Weka: The Waikato environment for knowledge analysis," In *Computer Science Research Students Conference Proceedings of the New Zealand*, 1995, pp. 57-64.

[66] E.Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I.H. Witten, and L. Trigg, "Weka, " In *Data Mining and Knowledge Discovery Handbook*, Springer, 2005, pp. 1305-1314.

[67] M.Tavallaee, E. Bagheri, W. Lu, and A.A. Ghorbani, “A detailed analysis of the KDD CUP 99 dataset, “ In *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications,* IEEE, *2009.*

[68] H. Bock, *The definitive guide to NetBeans Platform*. Berkeley, CA: Apress, 2009.

[69] M. Sabhnani and G. Serpen, “In *MLMTA* Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context,” MLMTA, 2003, pp. 209-215.

[70] I. Levin, “KDD-99 classifier learning contest LLSoft's results overview,” *ACM SIGKDD Explorations Newsletter*, vol. 1, no. 2, p. 67, 2000.

[71] S.J. Stolfo, A.L. Prodromidis, S. Tselepis, W. Lee, D.W. Fan, and P.K. Chan, “In *KDD* JAM: Java Agents for Meta-Learning over Distributed Databases,” vol. 97, pp. 74-81, 1997.

[72] I.Levin, “KDD-99 classifier learning contest: LLSoft's results overview,” *SIGKDD explorations,* vol.1, no. 2, pp. 67-75, 2002.

[73] P.G. Jeya, M. Ravichandran, and C. S. Ravichandran, “Efficient Classifier for R 2 L and U 2 R Attacks,” *International Journal of Computer Applications,* vol. 45, no. 21, 2012.

[74] S.Paliwal and Ravindra Gupta, “Denial-of-Service, Probing & Remote to User (R2L) Attack Detection using Genetic Algorithm,” *International Journal of Computer Applications,* vol. 60, no. 19, 2012.

[75] L.Sunitha, M. BalRaju, J. Sasikiran, and E.V. Ramana. "Automatic outlier identification in data mining using IQR in real-time data." *International Journal of Advanced Research in Computer and Communication Engineering,*vol. 3, no. 6, pp. 7255-7257, 2014.

Publication

Accepted

P.Bansal and D.Garg, "A Hybrid Honeyd approach to reduce false alarm ratio in intrusion detection," *In International Symposium on Advanced Computing and Communication*, IEEE, 2015.

P.Bansal and D.Garg, "A Hybrid Approach to Improve the Anomaly Detection Rate Using Data Mining Techniques," *In 4th International Conference on Reliablity and Infocom Technologies and Optimazation*, IEEE, 2015

Video Link

This is link to my YouTube video where i have presented brief summary of my thesis topic:https://www.youtube.com/watch?v=RnJawjIuiCE&feature=youtu.be

Plagiarism Report

Turnitin Originality Report

A Hybrid Approach to Improve the Anomaly Detection Rate Using Data Mining Techniques by Priya Bansal

From Thesis (ME 2013-2015 Batch)

- Processed on 15-Jul-2015 05:34 IST
- ID: 555014109
- Word Count: 10301

Similarity Index
6%
Similarity by Source

Internet Sources:
1%
Publications:
5%
Student Papers:
0%

sources:

1 1% match (publications)
Patcha, A.. "An overview of anomaly detection techniques: Existing solutions and latest technological trends", Computer Networks, 20070822

2 < 1% match (publications)
Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software : an update", ACM SIGKDD Explorations Newsletter, 2009.