Data Mining: Evaluation for Intrusion Detection System

[1]Jucy Job M, [2]Hinduja Hariharan
[1]Computer Programmer Kerala Agricultural University, Kerala, India
[2]Programmer Kerala Forest Research Institute, Thrissur, Kerala, India

Abstract
Data mining has been gaining popularity in knowledge discovery field, particularity with the increasing availability of digital documents in various languages from all around the world. Network intrusion detection is the process of monitoring the events occurring in a computing system or network and analyzing them for signs of intrusions. In this paper, intrusion detection & several areas of intrusion detection in which data mining technology applied are discussed. Data mining techniques are used to discover consistent and useful patterns of system features that describe program and user behavior. Data mining can improve variant detection rate, control false alarm rate and reduce false dismissals. By using these set of relevant system features to compute classifiers that recognize anomalies & known intrusion.

Keywords
Intrusion Detection, Data Mining, Misuse Detection, Anomaly Detection.

I. Introduction

In recent year computer technology have been utilized by many people all over the world in several areas. With the development of internet technology, network security has become a global focus in the world. Traditional security such as firewall, VPN and data encryption is insufficient to detect against attacks by crackers. However, intrusion detection is a dynamic one, which can give dynamic protection to the network security in monitoring, attack and counter attack [1]. For collecting the data set,

1. **Anomaly Detection:** Anomaly detection refers to storing features of user's usual behaviors into database, then comparing user's current behavior with those in database. If the deviation is huge enough, we can say that there is something abnormal.
2. **Misuse Detection:** Misuse Detection refers to confirming attack incidents by matching features through the attacking feature library.

We decided to use data mining for solving the problem of network intrusion because of following reasons [1, 4, 5,6]

- Data mining can process huge amount of data.
- It is more useful to find out the ignored and hidden information.

Data mining algorithms are used to perform data summarization and visualization that help the security analysis in various areas. [7].

Now a days, users are facing the new challenges of electronic attacks. In this context, Intrusion detection is the important technology which gives us remedial solution to this problem. An intrusion is defined as any set of actions that threat the integrity, confidentiality, or availability of a network resource such as user accounts, file systems, system kernels, and so on.

According to Webster's an intrusion as the act of thrusting in, or of entering into a place without invitation, right, or welcome [1]. Intrusion is defined as the act of wrongfully entering upon, grasping, or taking possession of the property of another [2]. Intrusion is coming into place without permission.

In Intrusion Detection (ID), collects the information and analyzing it for uncommon or unexpected events. Intrusion detection is the process of monitoring and analyzing the events which occurred in a computer system in order to detect signs of security problems [3]. Over the past few years, intrusion detection and other security technologies such as cryptography, authentication, and firewalls have increasingly gained importance in digital data [4]. Intrusion detection is data analysis process. The main theme of our approach is to apply data mining techniques to intrusion detection. Data mining is the process of extracting patterns from large amount of stored data.

Now days the main reason of applying Data Mining for intrusion detection systems is the enormous volume of existing and newly appearing network data that requires processing [6].

Traditional intrusion detection systems face many limitations. So this has led to an increased interest in data mining for intrusion detection. Data mining can improve variant detection rate, control false alarm rate and reduce false dismissals [2]. This paper has been divided into six sections. Section I defines the overview of data mining approaches for network intrusion detection system. Section II portrays the basic idea of intrusion detection. Section III Mining patterns and Architecture support are discussed. Section IV highlights the various components of intrusion detection system. Section V highlights some areas of intrusion detection in which data mining are applied and section VI finally conclude by discussing the outcome of study.

II. Intrusion Detection

Intrusion detection is the process of monitoring the events occurring in a digital network and analyzing them for signs of possible incidents [7].

The security of our digital network and data is at continual risk. Due to the extensive growth of the Internet and increasing availability of tools and tricks for intruding and attacking networks have prompted that intrusion detection is become a critical component for network administrator. The purpose of intrusion detection is to detect security violations in information systems. Intrusion detection is a passive approach to security as it monitors information systems and raises alarms when security violations are detected. Examples of security violations include the abuse of privileges or the use of attacks to exploit software or protocol vulnerabilities. So there is need of one of the tool which automatically detects the intrusions in digital network. Hence an intrusion detection system is software that automatically detects the intrusions occurred in system

Intrusion detection system uses many methodologies to detect incidents. Incidents have many reasons, such as malware e.g., worms, spyware, attackers gaining unauthorized access to systems from the Internet, and authorized users of systems who misuse their privileges or attempt to gain additional privileges for which they are not authorized. Traditionally, intrusion detection techniques are classified into two broad categories: misuse detection and anomaly detection [8].

Intrusion detection systems are also categorized according to the kind of input information they analyze. So this is classified

into host-based, network-based, wireless and Network Behavior Analysis (NBA) intrusion detection system [7].

A. Anomaly Intrusion Detection

Misuse detection system unable to detect new or previously unknown intrusions occurred in computer system or digital network. Novel intrusions can be found by anomaly detection. Anomaly detection uses a model of normal user or system behavior and flags significant deviations from this model as potentially malicious [9-10]. This model of normal user or system behavior is commonly known as the user or system profile. Strength of anomaly detection is its ability to detect previously unknown attack. A typical anomaly detection system is as shown fig. 1.

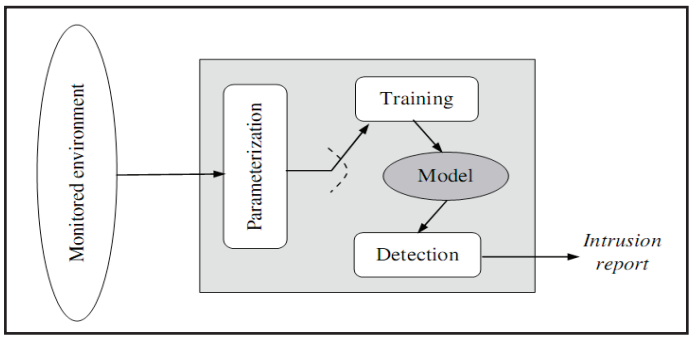

Fig. 1: Anomaly Detection

The anomaly detection system is effective against novel or unknown attacks. There is no need of prior knowledge about specific intrusions in anomaly detection technique. One of the drawbacks of anomaly detection is the high percentage of false positives [10].

B. Misuse Intrusion Detection

Misuse detection searches for the traces or patterns of well-known attacks which are stored as signatures [9-10]. These signatures are provided by human expert based on their extensive knowledge of intrusion techniques. In this process if a pattern matched is found, this signals an event for which an alarm raised. After that security analyst evaluate the alarms to decide what action to take for e.g. shutting down part of the system, alerting the relevant internet service provider of suspicious traffic, or simply nothing unusual traffic for future reference. In misuse detection, each instance in data set is labeled as normal or intrusion and a learning algorithm is trained over labeled data [10]. From this discussion we can say that only known attacks that leave characteristic traces can be detected. This is one of the drawbacks of misuse detection. Also there is need to update the signature whenever new software version arrive or changes in network configuration occur because the systems are dynamic. A key advantage of misuse detection technique is their high degree of accuracy in detecting known attacks and their variations [9-10]. A typical misuse detection system is as shown in fig. 2.

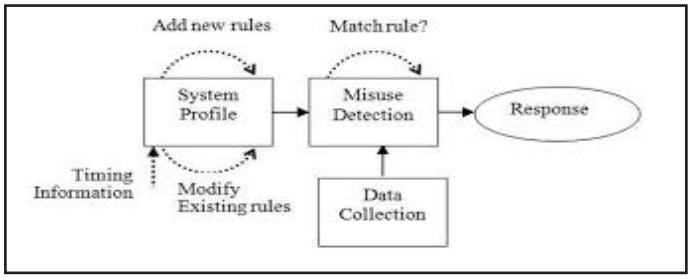

Fig. 2: Misuse Intrusion Detection

C. Host-based Intrusion Detection

Host-based intrusion detection system analyzes host-bound audit sources such as operating system audit trails, system logs, or application logs. These types of systems actually run on the system being monitored. These data come from the records of different host system activities, including appraisal record of OS, system logs, application program information, and so on.

D. Network-based Intrusion Detection

Network-based intrusion detection system analyzes network packets that are captured on a network. Network packet is the data source for network intrusion detection system. In the past few years, a growing number of research projects have applied data mining to intrusion detection in network data [10, 14, and 15]. These types of system are placed on the network, near the system or system being monitored. They examine the network traffic and determine whether it falls within acceptable boundaries. these data come through network segments, such as :Internet packets.

A "network intrusion detection system (NIDS)" monitors traffic on a network looking for suspicious activity, which could be an attack or unauthorized activity.A large NIDS server can be set up on a backbone network, to monitor all traffic; or smaller systems can be set up to monitor traffic for a particular server, switch, gateway, or router. In addition to monitoring incoming and outgoing network traffic, a NIDS server can also scan system files looking for unauthorized activity and to maintain data and file integrity. The NIDS server can also detect changes in the server core components.

A NIDS server can also scan server log files and look for suspicious traffic or usage patterns that match a typical network compromise or a remote hacking attempt. Possible uses include scanning local firewalls or network servers for potential exploits, or for scanning live traffic to see what is actually going on.A NIDS server does not replace primary security such as firewalls, encryption, and other authentication methods. The NIDS server is a backup network integrity device. Neither system (primary or security and NIDS server) should replace common precaution (building physical security, corporate security policy, etc.)

E. Wireless Intrusion Detection

Wireless intrusion detection system monitors wireless network traffic and analyzes its wireless networking protocols to identify suspicious activity involving the protocols themselves. It cannot identify suspicious activity in the application or higher-layer network protocols such as TCP, UDP that the wireless network traffic is transferring.

F. Network Behavior Analysis

Network Behavior Analysis which examines network traffic to identify threats that generate unusual traffic flows, such as distributed denial of service (DDoS) attacks [6], certain forms of malware such as worms, backdoors, and policy violations e.g., a client system providing network services to other systems. Network behavior analysis systems are also deployed to monitor flows on an organization's internal Networks, and are also sometimes deployed where they can monitor flows between an organization's Networks and external networks such as the Internet.

Table 1: Misuse Vs. Anomaly Detection

	Misuse Detection	Anomaly Detection
Definition	mathching the sequence of "Signature actions" of known intrusion scenarios	using statistical measure on system features
Shortcoming	1. Has to hand coded known pattern 2. Unable to detect any future intrusion	1. Rely upon in selecting the system features 2. Has to study sequential interrelation between transactions

III. Mining Patterns and Architecture Support

A. Mining Patterns from Audit Data

In order to construct an accurate (effective) base classifier, we need to gather a sufficient amount of training data and identify a set of meaningful features. Both of these tasks require insight into the nature of the audit data, and can be very difficult without proper tools and guidelines. In this section we describe some algorithms that can address these needs. Here we use the term "audit data" to refer to general data streams that have been properly processed for detection purposes. An example of such data streams is the connection record data extracted from the raw tcpdump output.

1. Association Rules

The goal of mining association rules is to derive multi-feature (attribute) correlations from a database table. A simple yet interesting commercial application of the association rules algorithm is to determine what items are often purchased together by customers, and use that information to arrange store layout. Formally, given a set of records, where each record is a set of items, an association rule is an expression X --> Y, confidence, support. X and Y are subsets of the items in a record, support is the percentage of records that contain X+Y, whereas confidence is support(X+Y)/ support(X). For example, an association rule from the shell command history file (which is a stream of commands and their arguments) of a user is

trn --> rec.humor; [0.3, 0.1],

which indicates that 30% of the time when the user invokes trn, he or she is reading the news in rec.humor, and reading this newsgroup accounts for 10% of the activities recorded in his or her command history file. Here 0.3 is theconfidence and 0.1 is the support.

The motivation for applying the association rules algorithm to audit data are:

- Audit data can be formatted into a database table where each row is an audit record and each column is a field (system feature) of the audit records;
- There is evidence that program executions and user activities exhibit frequent correlations among system features. For example, one of the reasons that "program policies", which codify the access rights of privileged programs, are concise and capable to detect known attacks [KFL94] is that the intended behavior of a program, e.g., read and write files from certain directories with specific permissions, is very consistent. These consistent behaviors can be captured in association rules;
- We can continuously merge the rules from a new run to the aggregate rule set (of all previous runs).

Our implementation follows the general association rules algorithm, as described in [Sri96].

2. Frequent Episodes

While the association rules algorithm seeks to find intra- audit record patterns, the frequent episodes algorithm, as described in [MTV95], can be used to discover inter- audit record patterns. A frequent episode is a set of events that occur frequently within a time window (of a specified length). The events must occur (together) in at least a specified minimum frequency, min_ fr, sliding time window. Events in a serial episode must occur in partial order in time; whereas for a parallel episode there is no such constraint. For X and Y where X+Y is a frequent episode, X --> Y with confidence=frequency(X+Y)/ frequency(X) and support=frequency(X+Y) is called a frequent episode rule. An example frequent serial episode rule from the log file of a department's Web site is

home, research --> theory; [0.2, 0.05], [30s]

which indicates that when the home page and the research guide are visited (in that order), in 20% of the cases the theory group's page is visited subsequently within the same 30s time window, and this sequence of visits occurs 5% of the total (the 30s) time windows in the log file (that is, approximately 5% of all the records).

We seek to apply the frequent episodes algorithm to analyze audit trails since there is evidence that the sequence information in program executions and user commands can be used to build profiles for anomaly detection [FHSL96,LB97]. Our implementation followed the description in [MTV95].

3. Using the Discovered Patterns

The association rules and frequent episodes can be used to guide the audit process. We run a program many times and under different settings. For each new run, we compute its rule set (that consists of both the association rules and the frequent episodes) from the audit trail, and update the (existing) aggregate rule sets using the following merge process:

- For each rule in the new rule set: find a match in the aggregate rule set. A match is defined as the exact matches on both the LHS and RHS of the rules, plus epsilon matches (using ranges), on the support (or frequency) andconfidence values
- If a match is found, increment the match_count of the matched rule in the aggregate rule set. Otherwise, add the new rule and initialize its match_count to be 1.

When the rule set stabilizes (there are no new rules added), we can stop the data gathering process since we have produced a near complete set of audit data for the normal runs. We then prune the rule set by eliminating the rules with low match_count, according to a user-defined threshold on the ratio of match_count over the total number of audit trails. The system builders can then use the correlation information in this final profile rule set to select a subset of the relevant features for the classification tasks. We plan to build a support environment to integrate the process of user selection of features, computing a classifier (according to the feature set), and presenting the performance of the classifier. Such a support system can speed up the iterative feature selection process, and help ensure the accuracy of a detection model.

We believe that the discovered patterns from (the extensively gathered) audit data can be used directly for anomaly detection. We compute a set of association rules and frequent episodes from a new audit trail, and compare it with the established profile rule set.

Scoring functions can be used to evaluate the deviation scores for: missing rules with high support, violation (same antecedent but different consequent) of rules with high support and confidence, new (unseen) rules, and significant changes in support of rules.

B. Architecture Support

The biggest challenge of using data mining approaches in intrusion detection is that it requires a large amount of audit data in order to compute the profile rule sets. And the fact that we may need to compute a detection model for each resource in a target system makes the data mining task daunting. Moreover, this learning (mining) process is an integral and continuous part of an intrusion detection system because the rule sets used by the detection module may not be static over a long period of time. For example, as a new version of a system software arrives, we need to update the "normal" profile rules. Given that data mining is an expensive process (in time and storage), and real-time detection needs to be lightweight to be practical, we can't afford to have a monolithic intrusion detection system.

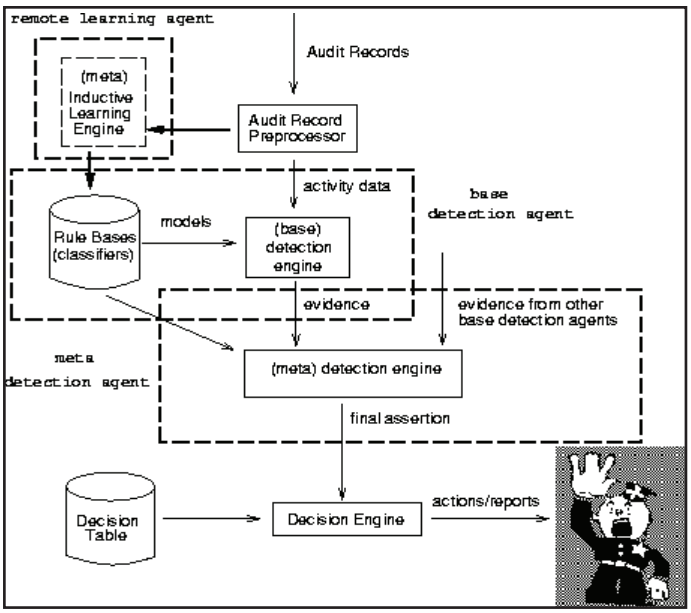

Fig. 3: An Architecture for Agent-Based Intrusion Detection System

We propose a system architecture, as shown in fig. 3, that includes two kinds of intelligent agents: the learning agents and the detection agents. A learning agent, which may reside in a server machine for its computing power, is responsible for computing and maintaining the rule sets for programs and users. It produces both the base detection models and the meta detection models. The task of a learning agent, to compute accurate models from very large amount of audit data, is an example of the "scale-up" problem in machine learning. We expect that our research in agent-based meta-learning systems [SPT+97] will contribute significantly to the implementation of the learning agents. Briefly, we are studying how to partition and dispatch data to a host of machines to compute classifiers in parallel, and re-import the remotely learned classifiers and combine an accurate (final) meta-classifier, a hierarchy of classifiers [CS93].

A detection agent is generic and extensible. It is equipped with a (learned and periodically updated) rule set (i.e., a classifier) from the remote learning agent. Its detection engine "executes" the classifier on the input audit data, and outputs evidence of intrusions. The main difference between a base detection agent and the meta detection agent is: the former uses preprocessed audit data as input while the later uses the evidence from all the base detection agents. The base detection agents and the meta detection agent need not be running on the same host. For example, in a network environment, a meta agent can combine reports from (base) detection agents running on each host, and make the final assertion on the state of the network.

The main advantages of such a system architecture are:

- It is easy to construct an intrusion detection system as a compositional hierarchy of generic detection agents.
- The detection agents are lightweight since they can function independently from the heavyweight learning agents, in time and locale, so long as it is already equipped with the rule sets.
- A detection agent can report new instances of intrusions by transmitting the audit records to the learning agent, which can in turn compute an updated classifier to detect such intrusions, and dispatch them to all detection agents. Interestingly, the capability to derive and disseminate anti-virus codes faster than the virus can spread is also considered a key requirement for anti-virus systems [KSSW97].

IV. Components of Intrusion Detection System

From the above discussion, intrusion detection is the monitoring and analyzing digital data. So typical components used in an intrusion detection system are :

A. Sensor or Agent

Sensors and agents monitor and analyze activity. The term sensor is typically used for intrusion detection systems that monitor networks, including network-based, wireless, and network behavior analysis technologies. The term agent is typically used for host-based intrusion detection system technologies [7,14].

B. Management Server

A management server is a centralized device that receives information from the sensors or agents and manages them. Some of the management servers perform analysis on the event information that the sensors or agents provide and can identify events that the individual sensors or agents cannot. Matching event information from multiple sensors or agents, such as finding events triggered by the same IP address, is known as correlation. Some small intrusion detection system deployments do not use any management servers, but most intrusion detection system deployments management server. In larger intrusion detection system deployments, there are often multiple management servers [7].

C. Database Server

A database server is a repository for event information recorded by sensors, agents, or management servers. Many intrusion detection systems have database servers.

D. Console

A console is a program that provides an interface for the intrusion detection system's users and administrators. Console software is typically installed onto standard desktop or laptop computers. Some consoles are used for intrusion detection system administration only, such as configuring sensors or agents and applying software updates, while other consoles are used only for monitoring and analysis. Some intrusion detection system consoles provide both administration and monitoring capabilities.

V. Application of Data Mining in Intrusion Detection

After discussing the various components in intrusion detection system in this section various areas of intrusion detection in which data mining technology are applied are studied. The following are areas in which data mining technology applied or further developed for intrusion detection.

A. Data Mining Algorithms for Intrusion Detection

Data mining algorithms can be used for misuse detection and anomaly detection. In misuse detection, training data are labeled as either "normal" or "intrusion." A classifier can then be derived to detect anomalies & known intrusions [10, 12]. Research in this area has included the application of classification algorithms, association rule mining, and cost-sensitive modeling. Anomaly detection builds models of normal behavior and automatically detects significant deviations from it [9]. Supervised or unsupervised learning can be used. In a supervised approach, the model is developed based on training data that are known to be "normal." In an unsupervised approach, no information is given about the training data [15]. Anomaly detection research has included the application of classification algorithms, statistical approaches, clustering, and outlier analysis [2,8,12,13, and 15]. The techniques used must be efficient and scalable, and capable of handling network data of high volume, dimensionality, and heterogeneity [11].

Classification algorithm about Data Mining can be used to construct classifier, after the invasion of a large number of data sets being trained [12]. Classifier can be used for intrusion detection. Clustering analysis algorithm can also be used to construct the network model of normal behavior, or intrusion behavior model [2, 13]. Association analysis algorithm can be used to describe the invasion of behavior patterns of association rules, through these rules intrusion detection can come [12].

B. Association and Correlation Analysis Helps to Select And Build Discriminating Attributes

Association and correlation mining can be applied to find relationships between system attributes describing the network data [12]. Such information can provide insight regarding the selection of useful attributes for intrusion detection. New attributes derived from aggregated data may also be helpful, such as summary counts of traffic matching a particular pattern [11].

C. Analysis of Stream Data

Due to the transient and dynamic nature of intrusions and malicious attacks, it is difficult to perform intrusion detection in the data stream environment. However, an event may be normal on its own, but considered malicious if viewed as part of a sequence of events. Thus it is necessary to study what sequences of events are frequently encountered together, find sequential patterns, and identify outliers [15]. Other data mining methods for finding evolving clusters and building dynamic classification models in data streams are also necessary for real-time intrusion detection.

D. Distributed Data Mining

Intrusions can be launched from several different locations and targeted to many different destinations. Distributed data mining methods may be used to analyze network data from several network locations in order to detect these distributed attacks [6, 14].

E. Visualization and Querying Tools

Visualization tools should be available for viewing any anomalous patterns detected. Such tools may include features for viewing associations, clusters, and outliers. Intrusion detection systems should also have a graphical user interface that allows security analysts to pose queries regarding the network data or intrusion detection results [11].

These are the areas in which data mining technologies are applied and developed for intrusion detection.

VI. Conclusion

Intrusion detection system has tremendous demand in this digital era which enables us to detect security violation in information system. Intrusion detection systems based on data mining are generally more precise and require far less manual processing and input from human experts. Different data mining approaches like classification, association rule, clustering, and outlier detection are the few techniques frequently used to analyze network traffic or data to gain knowledge that helps in controlling intrusion. Today the main reason of using Data Mining for intrusion detection systems is the enormous volume of existing and newly appearing network data that will be useful for future pattern generation and recognition in the digital forensics research.

References

[1] [Online] Available: www.syngress.com,chapter 1,"Intrusion Detection System".

[2] Meng Jianliang, Shang Haikun BianLing,"The Application on Intrusion Detection Based on K-means Cluster Algorithm", International Forum on Information Technology and Applications IEEE, pp. 150-152, 2009.

[3] Bace R,"Intrusion Detection", MacMillan Technical Publishing, 2000.

[4] Allen, J., Christie, A., Fithen, W., McHugh, J., Pickel, J., Stoner, E.,"State of the practice of Intrusion Detection Technologies", Technical report. Camegie Mellon University. [Online] Available: http://www.cert.org/archive/pdf/99tr028.pdf, 2000.

[5] U.Fayyad, G.Piatetsky-Shapiro, P.Smyth,"From Data Mining to Knowledge Discovery in Databases", Articles in AI Magazine,1996.

[6] Kanwal Garg, Rshma Chawla,"Detection of DDOS Attacks Using Data Mining", International Journal of Computing and Business Research (IJCBR), Vol. 2, Issue 1, 2011.

[7] Karen Scarfone, Peter Mell,"Guide to Intrusion Detection and Prevention Systems (IDPS)", National Institute of Standards and Technology Special Publication, pp. 800-94, 2007.

[8] Mannila, H.,"Data Mining: Machine Learning, Statistics, and Databases", In Proceedings of the 8th International Conference on Scientific and Statistical Database Management, pp. 1–8,1996.

[9] Foong Heng Wai, Yin Nwe Aye, Ng Hian James,"Intrusion Detection in Wireless Ad-Hoc Networks", CS4274 Introduction to mobile computing, 2004.

[10] Paul Dokas, levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, Pang-Ning Tan,"Data Mining For Network Intrusion Detection", 2003.

[11] Han Jiawei, Kamber Micheline,"Data Mining: Concepts and Techniques", (Second Edition) San Francisco, Morgan Kaufmann Publishers, 2006.

[12] Wenke Lee, Salvatore J. Stolfo, "Data Mining Approaches for Intrusion Detection", 1998.

[13] Li Bo, Jiang Dong-Dong,"The Research of Intrusion Detection Model Based on Clustering Analysis" International Conference

on Computer and Communications SecurityIEEE,2009.
[14] Imen Brahmi, Sadok Ben Yahia, Pascal Poncelet", MAD-IDS: Novel Intrusion Detection System using Mobile Agents and Data Mining Approaches", 2010.
[15] Jiong Zhang, Mohammad Zulkernine,"Anomaly Based Network Intrusion Detection With Unsupervised Outlier Detection", 2006.
[16] D. E. Denning,"An intrusion detection model," IEEE Transaction on Software Engineering, 1987.

Jucy Job M received her B.Sc. degree in Computer Science from University of Calicut, Kerala, India, in 2011, the MCA. degree from University of Calicut, Kerala, India, in 2014. She was a Computer Programmer in Kerala Agricultural University since December 2014. Her research interests include Data Mining and in the Google's Project LOON.