IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. IX (May-Jun. 2014), PP 73-79
www.iosrjournals.org

Improved Intrusion Detection System Using Discriminative learning Approach (A Review)

[1]Charanjeet Kaur, [2] Dr.Vinay Gautam(Asst Professor)
Desh Bhagat University,India *Desh Bhagat University,India*

Abstract: *With the advent of anomaly-based intrusion detection systems, many approaches and techniques have been developed to track novel attacks on the systems. High detection rate of 98% at alow alarm rate of 1% can be achieved by using these techniques. Though anomaly-based approaches are efficient, signature-based detection is preferred for mainstream implementation of intrusion detection systems. As a variety of anomaly detection techniques were suggested, it is difficult tocompare the strengths, weaknesses of these methods. The reason why industries don"t favor the anomaly-based intrusion detection methods can be well understood by validating the efficiencies of the all the methods. To investigate this issue, the current state of the experiment practice in the field of anomaly-based intrusion detection is reviewed and survey recent studies in this. This paper contains summarization study and identification of the drawbacks of formerly surveyed works..*

Keywords: *Intrusion Detection, Anomaly-based Detection, Signature-based detection*

I. Introduction

As the growing need of internet in our daily life and our dependence on the world wide system of computer networks, the network security is becoming a necessary requirement of our world to secure the confidential information available on the networks. The precious information is always prone to maximum attacks over the network. Intrusion may occur due to system vulnerabilities or security breaches, such as system misconfiguration, user misuse or program defects. Attackers can also combine multiple security vulnerabilities into an intelligent intrusion. Intrusion detection plays an important role over the large network system. In a big network system there are large number of servers and on-line services running in the system while such networks may lure more attackers. Efficient intrusion detection model is needed as a defence of the network systems.

Intrusion detection systems are the `burglar alarms' (or rather `intrusion alarms') of the computer security field. The aim is to defend a system by using a combination of an alarm that sounds whenever the site's security has been compromised, and an entity most often a site security officer (SSO)that can respond to the alarm and take the appropriate action, for instance by ousting the intruder, calling on the proper external authorities, and so on. This method should be contrasted with those that aim to strengthen the perimeter surrounding the computer system. We believe that both of these methods should be used, along with others, to increase the chances of mounting a successful defense, relying on the age old principle of defense in depth. It should be noted that the intrusion can be one of a number of different types. For example, a user might steal a password and hence the means by which to prove his identity to the computer. We call such a user a masquerader, and the detection of such intruders is an important problem for the field. Other important classes of intruders are people who are legitimate users of the system but who abuse their privileges, and people who use pre-packed exploit scripts, often found on the Internet, to attack the system through a network. This is by no means an exhaustive list, and the classification of threats to computer installations is an active area of research. Early in the research into such systems two major principles known as anomaly detection and signature detection were arrived at, the former relying on fagging all behavior that is abnormal for an entity, the latter fagging behavior that is close to some previously defined pattern signature of a known intrusion. The problems with the first approach rest in the fact that it does not necessarily detect undesirable behavior, and that the false alarm rates can be high. The problems with the latter approach include its reliance on a well-defined security policy, which may be absent, and its inability to detect Intrusions that have not yet been made known to the intrusion detection system. It should be noted that to try to bring more stringency to these terms, we use them in a slightly different fashion than previous researchers in the field. An intrusion detection system consists of an audit data collection agent that collects information about the system being observed. This data is then either stored or processed directly by the detector proper, the output of which is presented to the SSO, who then can take further action, normally beginning with further investigation into the causes of the alarm

II. Related Work

Sharmila Kishore Wagh, et.al[1] have proposed the different machine approaches for Intrusion detection system, and also presents the system design of an Intrusion detection system to reduce false alarm rate and improve accuracy to detect intrusion. An Intrusion Detection System is designed to detect system attacks and classify system activities into normal and abnormal form. Machine learning techniques have been applied to intrusion detection systems which have an important role in detecting Intrusions. Machine learning is concerned with the design and development of algorithms and methods that allow computer systems to autonomously acquire and integrate knowledge to continuously improve them to finish their tasks efficiently and effectively.

Mahdi Zamani, et.al[2] have proposed efficient adaptive methods like various techniques of machine learning can result in higher detection rates, lower false alarm rates and reasonable computation and communication costs. Because most techniques used in today's IDS are not able to deal with the dynamic and complex nature of cyber-attacks on computer networks. So the author reviewed several influential algorithms for intrusion detection based on various machine learning techniques, and divided these algorithms into two types of ML-based schemes: Artificial Intelligence (AI) and Computational Intelligence (CI).These algorithms has many features such as adaptation, fault tolerance, high computational speed and error resilience in the face of noisy information.

Mahesh Kumar Sabhnani, et.al[3] have proposed a small subset of machine learning algorithms, mostly inductive learning based, applied to the KDD 1999 Cup intrusion detection dataset resulted in dismal performance for user-to-root and remote-to-local attacks. The uncertainty to explore if other machine learning algorithms can demonstrate better performance compared to the ones already employed .Specifically, exploration of if certain algorithms perform better for certain attack classes and consequently, if a multi-expert classifier design can deliver desired performance measure is of high interest. The proposed model evaluated performance of a comprehensive set of pattern recognition and machine learning algorithms on four attack categories as found in the KDD 1999 Cup intrusion detection dataset. Results of simulation study implemented to that effect indicated that certain classification algorithms perform better for certain attack categories: a specific algorithm specialized for a given attack category . Consequently, a multi-classifier model, where a specific detection algorithm is associated with an attack category for which it is the most promising, was built. Empirical results obtained through simulation indicate that noticeable performance improvement was achieved for probing, denial of service, and user-to-root attacks.

Hua TANG, Zhuolin CAO, et.al[4] have proposed a new approach to detect network attacks, which aims to study the efficiency of the method based on machine learning in intrusion detection, including artificial neural networks and support vector machine. The research work compares accuracy, detection rate, false alarm rate and accuracy of other attacks under different proportion of normal information. KDD CUP'99 dataset is benchmark dataset in intrusion detection. However, the data is not distributed evenly error may occur if only one set is used. The proposed approach performs high performance, especially to U2R and U2L type attacks.

The Applied Research Laboratories of the University of Texas at Austin(ARL:UT),Chris Sinclair, Lyn Pierce et.al [5] has proposed genetic algorithms and decision trees to automatically generate rules for classifying network connections. Differentiating anomalous network activity from normal network is difficult and tedious. A human analyst must connections search through vast amounts of data to find anomalous sequences of network connections. To support the analyst's job, the researcher built an application which enhances domain knowledge with machine learning techniques to create rules for an intrusion detection expert system. This paper describes the machine learning methodology and the applications employing this methodology. The main result is the creation of rules to detect complex network intrusions to maximize the utility of the expert system, and to produce a dynamic rule base capable of detecting new attack signatures. `

Deepika P Vinchhukar, Alpa Reshamwal et.al [6] analyzed the neural network approach and the machine learning approach in overcoming the challenges of the IDS and discussed the Support Vector Machine to deal with the classifier construction problem. Intrusion Detection materializes the high network security. In this paper thus tries to be the most perfect system to deal with the network security and the intrusions attacks. Monitoring activity of the network and that of threats is the feature of the ideal Intrusion Detection System. Intrusion Detection System is classified on the basis of the source of Data and Model of Intrusion. The challenges faced by the IDS can be overwhelmed by Neural Network and Machine Learning Approaches.

III. Machine Learning Approach

Machine learning is a system capable of acquiring and integrating the knowledge automatically. The capability of the systems to learn from experience, training, analytical observation, and other means, results in a system that can continuously self-improve and thereby exhibit efficiency and effectiveness. A machine

learning system usually starts with some knowledge and a corresponding knowledge organization so that it can interpret, analysis, and test the knowledge acquired.

Machine learning techniques are based on establishing an explicit or implicit model. A singular characteristic of these schemes is the need for labeled data to train the behavioral model, a procedure that places severe demands on resources. In many cases, the applicability of machine learning principles coincides with that for the statistical techniques, although the former is focused on building a model that improves its performance on the basis of previous results. Hence, machine learning for IDS has the ability to change its execution strategy as it acquires new information. This feature could make it desirable to use such schemes for all situations.

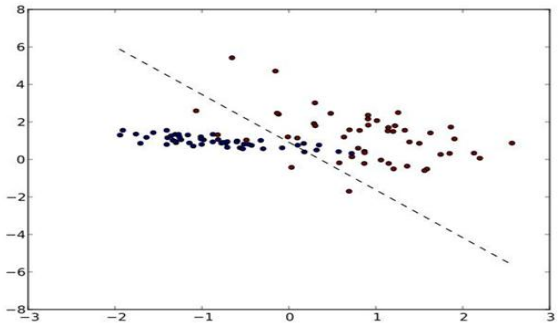

Fig.1: Machine Learning Approach

3.1 **Intrusion Detection System**

Intrusion detection system (IDS) is an active process or device that analyzes system and network activity for unauthorized and nasty activity. Intrusion Detection System (IDS) is any hardware, software, or a combination of both that monitors a system or network of systems against any malicious activity. The ultimate goal of any IDS is to catch perpetrators in the act before they do real damage to resources. An IDS protects a system from attack, misuse, and compromise. It also monitor network activity, audit network and system configurations for vulnerabilities, analyze data integrity, and more. IDS, these days, have become vital component in the security toolbox An IDS provides three functions: monitoring, detecting and generating an alert.

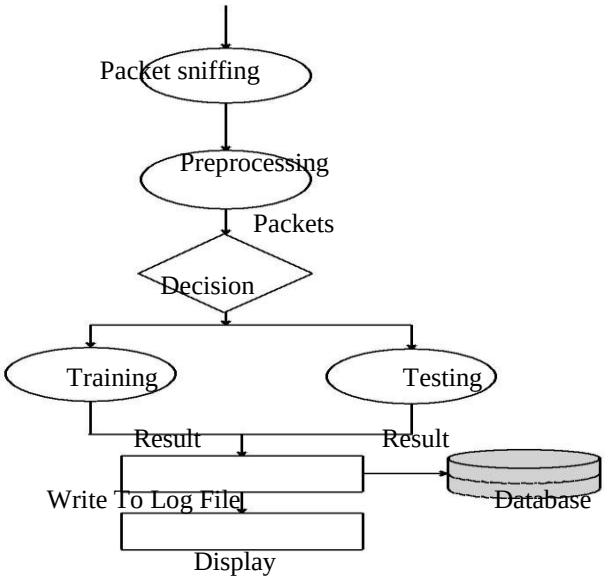

Fig 2: System Design for IDS

3.2 Classification Of Anomaly Detection

Currently the two basic methods of detection (analytical method) are signature-based and anomaly-based .The signature-based method, also known as misuse detection, seems for a specific signature to match, signalling an intrusion. They can detect many or all known attack patterns, but the weakness of signature based intrusion detection systems is the incapability of identifying new types of attacks or variations of known attacks.

Another useful method for intrusion detection is called anomaly detection. Anomaly detection

applied to intrusion detection and computer security has been an active area of research.. In anomaly based IDSs, the normal behaviour of the system or network traffic are represented and, for any behaviour that varies over a pre-defined threshold, an anomalous activity is identified. By the other side, in anomaly based IDSs, the number of false positives generated are higher than on those based on signatures. An important issue in anomaly based IDSs is how these systems should be trained, i.e., how to define what is a normal behaviour of a system or network environment (which features are relevant) and how to represent this behaviour computationally.

Accordingto the type of processing related to the ‘‘behavioural’’ model of the target system, anomaly detection techniques can be classified into three main categories statistical based, knowledge-based, and machine learning-based. In the well-known intrusion detection approaches and Comparison of various approaches reviewed with the strength and weakness of those approaches.

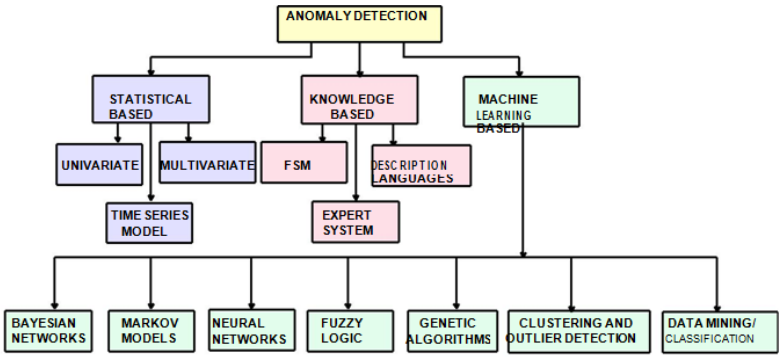

Fig 3: Classification of Anomaly detection

3.2.1 Statistical anomaly-based IDS

A statistical anomaly-based IDS find out normal network activity like what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other- and aware the administrator or user when traffic is detected which is anomalous (not normal). It is again categorized into univariate, multivariate and time series model. Univariate model parameters are modelled as independent Gaussian random variables thus defining an acceptable range of values for every variable. The multivariate model considers the correlation between two or more variables. The time series models uses an interval timer, together with an event counter or resource measure and take into account the order and inter arrival times of observations and their values which are labelled as anomaly if its probability of occurrence is too low at a given time.

3.2.2 Knowledge-based techniques

Knowledge based stores information about subject domain. Information in knowledge based contains symbolic representations of expert’s rules of judgment in a format that allow the inference engine to perform deduction upon it. The expert system approach is one of the most widely used knowledge-based IDS schemes. Knowledge based techniques are divided into frame based model, rule based model and expert system. Rule based is modified form of the grammar based production rules. Frame based model localizes an entire body of expected knowledge and actions into a single structure. Expert systems are intended to classify the audit data according to a set of rules, involving three steps. First, different attributes and classes are identified from the training data. Second, a set of classification rules, parameters or procedures are deduced. Third, the audit data classified.

3.2.3 Machine learning-based IDS

Machine learning techniques are based on establishing an explicit or implicit model. A singular characteristic of these schemes is the need for labeled data to train the behavioral model, a procedure that places severe demands on resources. In many cases, the applicability of machine learning principles coincides with that for the statistical techniques, although the former is focused on building a model that improves its performance on the basis of previous results. Hence, machine learning for IDS has the ability to change its execution strategy as it acquires new information. This feature could make it desirable to use such schemes for all situations.

Several machine learning-based schemes have been applied to IDS. Some of the most important techniques are explained in following subsections.

3.2.4 Bayesian Network

A Bayesian network is a model that encodes probabilistic relationships among important variables. This technique is generally used for intrusion detection in combination with statistical schemes, including the capability of encoding interdependencies between variables and of predicting events, as well as the ability to incorporate both prior knowledge and data.

Conditional probability P (A|B) is used for calculating the probability of at once the condition B is present. However, in the real world applications.one needs to know about the conditional probability P (B|A) for B once its evidence A is present. In this Bayes theory, the goal is to calculate the probability of a given hypothesis H considering its sign or evidence E already exists. The H can be assumed to be a sampled column feature vector and noted as x = {x1, x2, . . .}. In the following text the E (Evidence) and the C (Class) sign can be replaced (where C = {c1 c2, . . .}), if it makes it easier for the reader to understand the concept. The formula to calculate this probability is presented below

$$P\begin{bmatrix} H \\ E \end{bmatrix} = \frac{P(H)*P(E/H)}{P(E)} \quad (1)$$

3.2.5 Markov models

There are two subtypes of Markov models: Markov chains and hidden Markov models. A Markov chain is a set of states that are interconnected through certain transition probabilities, which determine the topology and the capabilities of the model. During a first training phase, the probabilities associated with the transitions are estimated from the normal behaviour of the target system. The detection of anomalies is then carried out by comparing the anomaly score (associated probability) obtained for the observed sequences with a fixed threshold. In the case of a hidden Markov model, the system of interest is assumed to be a Markov process in which states and transitions are hidden. Only the so-called productions are observable. Markov-based techniques have been extensively used in the context of host IDS, normally applied to system calls.

3.2.6 Neural networks

The Multilayer Perceptions (MLP) neural networks have been very successful in a variety of applications, producing results, which are at least competitive and often exceed other existing computational learning models. They are capable of approximating to arbitrary accuracy, any continuous function as long as they contain enough hidden units. This means that such models can form any classification decision boundary in feature space and thus act as non-linear discriminate function.

When the NN is used for pattern classification, there is one input node for each element of the feature vector. There is usually one output node for each class to which a feature may be assigned (show in Fig. 4). The hidden nodes enable internal representation of the data to be developed by the NN during learning.

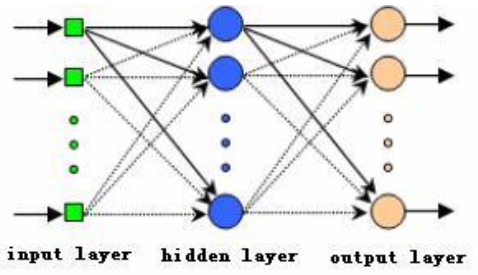

Fig.4 MPL NN Structure

One learning algorithm used for MLP is called back-propagation rule. This is a gradient descent method and based on an error function that represents the difference between the network's calculated output and the desired output. This error function is defined, based on the Mean Squared Error (MSE). Thus the error

for pattern ***i*** is written as:

$$E' = \frac{1}{2}\sum_{p=1}^{k} (y_p^i - o_p^i)^2 \quad (1)$$

Where y^i_p is the true output of the *p*th output node of the network when the *i*th feature vector is fed to the network and ***k*** represents the number of neurons of the output layer. Similarly the o^i_p is the desired output of the *p*th output node. Consequently the MSE can be summed over the entire training set.

In order to successfully learn, the network's true output should approach the desired output by continuously reducing the value of this error. The back-propagation rule calculates the error for a particular input and then back-propagates the error from one layer to the previous one. The connection weights, between the nodes, are adjusted according to the back-propagated error so that the error is reduced and the network learns.

3.2.6 Support Vector Machines

Support Vector Machines have become an increasingly popular tool for machine learning tasks involving classification and regression. The SVMs demonstrate various attractive features such as good generalization ability compared to other classifiers. Indeed, there are relatively few free parameters to adjust and the architecture does not require to be found experimentally.

Given a training set of instance-label pairs (x_i, y_i), $i=1, \ldots, l$ where $x_i \in R^n$ and $y \in \{1, -1\}^l$, the SVMs require the solution of the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} w^t w + C\sum_{i=1}^{l}\xi_i \qquad (2)$$

$$\text{Subject to } y_i (w^T \varphi(x_i) + b) \geq 1 - \xi_i , \quad \xi_i \geq 0 .$$

Here training vectors x_i is mapped into a higher (maybe infinite) dimensional space by the function $\boldsymbol{\varphi}$. Then SVM finds a linear separating hyper plane with the maximal margin in this higher dimensional space. ***b*** determines an offset of the discrimination hyper plane from origin. Slack variables $\boldsymbol{\xi_i}$ are introduced to measure the amount of violation of the constraints. The penalty ***C*** is a user defined positive regularization parameter (setting $\boldsymbol{C} = \infty$ leads back to the linearly separable case) which controls a trade-off between the wide margin and a small number of margin failures.

Furthermore, $\boldsymbol{K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)}$ is called the kernel function. There are many kernels that can be used such as Gaussian radial basis functions (RBF):

$$K(xi, x_j) = \exp - \frac{\left| xi - x j \right|^2}{2\sigma^2} \qquad (3)$$

Where $\sigma > 0$ is a constant that defines the kernel width .Another kernel function is the polynomial (of degree ***d***):

$$K(xi, x_j) = (1 + xi \cdot x_j)^d \qquad (4)$$

where $d > 0$ is a constant that defines the kernel order. The associated parameters, order d or Gaussian **σ** are determined within the training phase.

.

3.2.7 Fuzzy logic techniques

Fuzzy logic is derived from fuzzy set theory under which reasoning is approximate rather than precisely deduced from classical predicate logic. Fuzzy techniques are thus used in the field of anomaly detection mainly because the features to be considered can be seen as fuzzy variables. The fuzzy logic part of the system is responsible for both handling the large number of input parameters and dealing with the inaccuracy of the input data. Three fuzzy characteristics used in this work are COUNT, UNIQUENESS and VARIANCE. The implemented fuzzy inference engine uses five fuzzy sets for each data element (HIGH, MEDIUM-HIGH, MEDIUM LOW and MEDIUM-LOW) and suitable fuzzy rules to detect the intrusion. In their report authors have not specified how they have derived their fuzzy set. The fuzzy set is a very important issue for the fuzzy inference engine\ and in some cases genetic approach can be implemented to select the best combination. The proposed system is tested using data collected from the local area network in the college of Engineering at Iowa State University and the results are reported in this paper. The reported results are descriptive and not numerical; therefore, it is difficult to evaluate the performance of the reported work.

3.2.8 Genetic algorithms

Genetic algorithms are classified as global search heuristics, and evolutionary computation that uses techniques inspired by evolutionary biology such as recombination, selection, inheritance and mutation. Thus,

genetic algorithms represent another type of machine learning-based technique, capable of deriving classification rules and/or selecting appropriate features or optimal parameters for the detection process.

In rule evolution approach based on Genetic Programming (GP) for detecting novel attacks on networks is proposed. In their framework, four genetic operators, namely reproduction, mutation, crossover and dropping condition operators, are used to evolve new rules. New rules are used to detect novel or known network attacks. Experimental results show that rules generated by GPs with part of KDD 1999 Cup data set has a low false positive rate (FPR), a low false negative rate (FNR) and a high rate of detecting unknown attacks.

3.2.9 Clustering and outlier detection

Clustering techniques work by grouping the observed data into clusters, according to a given similarity or distance measure. The procedure most commonly used for this consists in selecting a representative point for each cluster. Clustering techniques to determine the occurrence of intrusion events only from the raw audit data, and so the effort required to tune the IDS is reduced. One of the most popular and most widely used clustering algorithms is K-Means, which is a non-hierarchical Centroid-based approach.

3.2.10 Decision Trees

Decision trees are structures used to classify data with common attributes. Each decision tree represents a rule which categorizes data according to these attributes. A decision tree consists of *nodes*, *leaves*, and *edges*. A node of a decision tree specifies an attribute by which the data is to be partitioned. Each node has a number of edges which are labelled according to a possible value of the attribute in the parent node. An edge connects either two nodes or a node and a leaf. Leaves are labelled with a decision value for categorization of the data.

3.2.10.1 C4.5 decision tree (C4.5).

The C4.5 algorithm [C4.5 Simulator], developed by Quinlan, generates decision trees using an information theoretic methodology. The goal is to construct a decision tree with minimum number of nodes that gives least number of misclassifications on training data. The C4.5 algorithm uses divide and conquer strategy.

3.2.11 K-means clustering (K-M) K-means clustering algorithm positions K centres in the pattern space such that the total squared error distance between each training pattern and the nearest centre is minimized.

References

[1]. Sharmila Kishore Wagh, Vinod K Pachghare, "Survey on Intrusion Detection System using Machine Learning Techniques", International Journal of Computer Applications (0975 – 8887**)** Volume 78 – No.16, September 2013.

[2]. Mahdi Zamani , "Machine Learning Techniques For Intrusion Detection " , arXiv:1312.2177vl [cs.CR] Dec 2013

[3]. Mahesh kumar sabhnani, Gursel Serpen , "Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context",

[4]. Hua Tang, Zhuolin CAO, "Machine Learning based Intrusion Detection Algorithms", Journal of Computational Information Systems, june2009, Available at http://www.JofCI.org.

[5]. Chris Sinclair, Lyn Pierce, Sara Matzner, "An Application of Machine Learning To Network Intrusion Detection",

[6]. Deepika P Vinchurkar, Alpa Reshamwala, "A Review of Intrusion Detection System Using Neural Network and Machine Learning Technique", International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 1, Issue 2, November 2012

[7]. Chia-Mei Chen Ya Lin Chen, Hsiao-Chung Lin, "An efficient network intrusion detection", Elsevier, Vol. 33, No. 4, 2010, pp. 477- 484.

[8]. "Nsl-kdd" data set for network-based intrusion detection systems," Available on: http://nsl.cs.unb.ca/NSL-KDD/, March 2009.

[9]. G. Meera Gandhi, Kumaravel Appavoo , S.K. Srivatsa, "Effiective Network Intrusion Detection using Classifiers Decision Trees and Decision rules", Int. J. Advanced Networking and Applications, Vol. 2, No. 3, 2010, pp. 686.

[10]. R. Aarthy and P. Marikkannu, "Extended security for intrusion detection system using data cleaning in large database" International Journal of Communications and Engineering, Vol. 2, No. 2, 2012, pp. 56-60.

[11]. Guy Helmer, Johnny S.K. Wong, Vasant Honvar, Les Miller, Yanxin Wang, "Lightweight agents for intrusion detection", Journal of systems and Soft-ware. Elsevier, Vol. 67, No. 2, 2010, pp.109-122.

[12]. Anna Sperotto, Gregor Scha_rath, Ramin Sadre, Cristian Morariu, Aiko Pras and Burkhard Stiller, "An Overview of IP Flow-Based Intrusion Detection" ,IEEE communications surveys & tutorials, Vol. 12, No. 3, 2010, pp. 343.

[13]. M. Tavallaee, E. Bagheri, W. Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", IEEE Symposium on computational intelligence in security and defense application, 2009.

[14]. Jiawei Han and Micheline kamber,"Data Mining Concepts and Techniques", Publisher Elsevier, pp.67-69, 296-301, 2001.