CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

The computer network systems and their technology require prevention from the threats, enemies and scandalous activities. In order to promise the integrity of computer network systems, security techniques are being brought in to a greater extent like antivirus software, firewalls and intrusion detection systems, etc. Intrusion detection system is a distinctive protection method which can find out unauthorized intrusion in a computer network or server. The IDS approaches are usually categorized into misuse and anomaly detection approaches in the literature. Misuse detection approach can unfailingly classify intrusion attacks in relation to the well-known signatures of discovered vulnerabilities. However, developing of a defense expert is required to characterize perfect rules or signatures that enhance the purpose of misuse detection approach. Conversely, the anomaly detection technique generally accords with statistical analysis and pattern recognition problems. It is capable of identifying new attacks without a priori knowledge about attacks. Many methods and frameworks have been developed to detect intrusions. Various techniques are also employed such as association rules, clustering, naïve Bayes classifier, support vector machines, genetic algorithms, artificial neural networks, fuzzy logic and boosting algorithm, and others have been applied to detect intrusions. The ample summary of IDS is presented in this chapter.
2.2 MACHINE LEARNING BASED INTRUSION DETECTION SYSTEM

Chi-Ho Tsang et al (2007) have given observable number of KDD cup99 feature details that was used by various classifiers like C4.5, Naïve Bayes, KNN, SVM and the proposed method Multi Objective Genetic Fuzzy Intrusion Detection System (MOGFIDS). Han-Ching Wu et al (2010) proposed an approach that makes use of the analytical strengths of neural networks to detect stepping-stone intrusion with two schemes. Among the two schemes, one uses eight packet variables and the other clusters a sequence of consecutive packet in round-trip times. Various transfer functions and learning rules were analyzed and the authors concluded that using sigmoid transfer function and delta learning rule generally provides better prediction.

Andrew Sung and Srinivas Mukkanama (2003) explained CI type learning machines for intrusion detection to protect the networked computers responsible for the operational control. The SVMs are better than ANNs in the important aspects of scalability, training time, running time and prediction accuracy. Both methods have been applied to identify the important features in the DARPA data. The real time intrusion detection system was developed by Byung-joo Kim and Il-kon Kim (2005), which combines on-line feature extraction method with Least Squares Support Vector Machine classifier. Applying the system to KDD Cup99 data, experimental results showed that it has significant performance compared to off-line intrusion detection system.

Rachid Beghdad (2007) presented a NN based model and experimented by training with two different types of dataset. The small data subset consists of 260 records, whereas large data subset consists of 65536 records. The experimentation outcome indicated that detection and classification performance was better for small data subset compared with large data subset. Chunlin Zhang et al (2005) proposed two kinds of Neural
Network (NN) based hierarchical IDS frameworks using Radial Basis Functions (RBF). A Serial Hierarchical IDS (SHIDS) is used to identify misuse attack accurately and anomaly attacks adaptively. A Parallel Hierarchical IDS (PHIDS) is used to enhance the SHIDS functionalities and performance. The experiments show that the combination of SHIDS and PHIDS detect attacks in real-time efficiently, train new classifiers for new intrusions automatically, and modify their structures adaptively after new classifiers are trained.

Neural methods are used for dealing with very huge data in an effective manner to distinguish and classify abnormal data in log files and lightening the administrator’s load (Shi-Jinn Horng et al 2008). The results of intrusion detection on Internet Information Services utilizes a hybrid IDS, and the feasibility of the hybrid IDS is validated based on the Internet Scanner System. In order to reduce the void time for updating the detector, the downgrade percentage should be limited. The Multilayer Perceptron (MLP) method was used for attack recognition by Vladimir Golovko and Pavel Kochurko (2005) and the important feature of MLP is the ability to generalize. The experimentation has been conducted with the bench mark KDD Cup99 data for training and testing of neural networks.

Taeshik Shon and Jongsub Moon (2007) proposed an approach, named Enhanced SVM, similar to that of a supervised SVM approach, which provides unsupervised learning and low false alarm capability. The additional techniques named Anomaly Detector using Enhanced SVM are used to improve the performance of the proposed approach. Initially, a profile of normal packets is created using Self-Organized Feature Map (SOFM), for SVM learning without preexisting knowledge. Secondly, a packet filtering scheme based on Passive TCP/IP Fingerprinting (PTF) is installed in order to reject incomplete network traffic that either violates the TCP/IP standard or
generation policy inside of well-known platforms. Thirdly, genetic algorithm based feature selection technique is used for extracting optimized information from raw internet packets. Fourthly, the flow of packets based on temporal relationships during data preprocessing is analyzed for considering the temporal relationships among the inputs used in SVM learning. Finally, the efficiency of the approach using the techniques of SOFM, PTF, and genetic algorithm was experimented on DARPA data.

The three-tier architecture of intrusion detection system was designed and developed by Tsong Song Hwang et al (2007). It consists of a blacklist, a whitelist and a multi-class support vector machine classifier based on the KDD Cup99 benchmark data. The known attacks from the traffic would be filtered by the blacklist and the whitelist discovers the normal traffics. The rest of the traffics and the anomalies detected by the white list were classified by a multi-class SVM classifier into any one of the attack categories. The network administrator can add the new patterns into the blacklist and allows fine tuning of the whitelist according to the environment of their network system and security policy. Shi-Jinn Horng et al (2010) proposed a Support Vector Machine (SVM) based intrusion detection system on KDD Cup99 training set, which combined the hierarchical clustering algorithm, a simple feature selection procedure and the SVM technique. The hierarchical clustering algorithm reduced the training time. The simple feature selection procedure is applied to eliminate unimportant features from the training set and is used to classify the network traffic data more accurately.

In the IDS models, the normal behaviors with HMM and attempts to detect intrusions by observing significant deviations from the models are explained (Sung-Bae Cho 2002). The NN and fuzzy logic are integrated to attain more robustness and flexibility. Self-organizing Map is used for optimal measures of audit data and HMM reduces them into appropriate size
for efficient modeling and finally, the fuzzy logic makes the decision of whether current behavior is anomalous or not. Shelly Xiaonan Wu and Wolfgang Banzhaf (2010) reviewed IDS with Computational Intelligence (CI) systems that included (ANN), fuzzy systems, evolutionary computation, artificial immune systems and swarm intelligence. The Least Squares Support Vector Machines (LSSVM) model was suggested by Haihua Gao et al (2006), using kernel space approximation through greedy searching and thus constructed a subspace basis of original space populated by training data. By means of this approximation, the training data was downsized and consequently, the numbers of support vectors of LSSVM model were reduced. So, the response time of intrusion detection was improved. The model has been evaluated using KDD Cup99 data and the results demonstrate that the method can be an effective way for fast intrusion detection.

The Next Generation Proactive Identification Model (NeGPAIM) was suggested by Martin Botha et al (2002), which was an enhanced method of Proactive Identification Model. NeGPAIM follows a three tier architecture used to improve the performance of the intrusion detection system. The model is nonspecific and can be implemented in different computing environments based on the prerequisites and consequences of different types of attacks. Srinivas Mukkamala et al (2004) examined the appropriateness of Linear Genetic Programming approach to model efficient intrusion detection systems and compared their performance with ANN and SVM. Based on the variety of comparative experiments, it is found that with appropriately chosen population size, program size, crossover rate and mutation rate, linear genetic programs could do better than support vector machines and neural networks in terms of detection accuracy. The experiments are tested with DARPA data.
The Multiple Classifiers (MCS) approach was suggested based on pattern recognition distinct feature representation and tested with different fusion rules (Giorgio Giacinto et al 2003). The reported results proved that the MCS approach provides a better false alarm generation than that provided by an individual classifier trained on the overall feature set. Among the fusion rules, the dynamic classifier selection technique provided the best results. The fusion of multiple classifiers achieves a better trade off than that provided by individual classifiers between generalization abilities and false alarm generation. Dae-Ki Kang et al (2005) proposed a method ‘bag of system calls’ and experimented misuse and anomaly detection results with other machine learning techniques. With the feature representation as input, the performance has been compared with several machine learning techniques for misuse detection. The results showed that simple ‘bag of system calls’ combining standard machine learning and clustering techniques is effective and often performs better than other approaches.

John Kros et al (2005) used a Bayesian influence diagram in conjunction with a decision tree to compute the cost of network intrusion. After analyzing this cost, system managers compare the loss associated with network security violation with the cost of utilizing appropriate security technology. This integrated model has adequate flexibility to accommodate the different threats and associated costs faced by different organizations. Srinivas Mukkamala et al (2005) addressed an ensemble approach of different soft computing and hard computing techniques for intrusion detection. The performances of ANN, SVM and Multivariate Adaptive Regression Splines (MARS) have been studied and an ensemble of ANNs, SVMs and MARS is found to be superior to individual approaches for intrusion detection in terms of classification accuracy.
Thomas Gamer et al (2007) suggested a collaboration of neighboring detection systems that the receiving systems search specifically for the attack which might have been missed by using local knowledge only. After receiving such attack information, a decision process has to determine if a search for this attack should be started. The design of the system is based on several principles which guide the decision process and finally the attack information will be forwarded to the next neighbors to increase the area of collaborating systems. The string matching algorithm (Lin Tan et al 2006) allowed the system to make decisions based not just on the headers, but the actual content flowing through the network. The approach showed how to convert the large database of strings into many tiny state machines, each of which searches for a portion of the rules and a portion of the bits of each rule and is 10 times more efficient than the other approaches. String matching plays a vital part in the execution of many spam detection algorithms and a faster string matching algorithm could boost optimization speed for embedded systems.

An Artificial Immune System-inspired Multi objective Evolutionary Algorithm (Charles Haag et al 2007) was suggested to reduce the size of the signature database. The approach is intended to measure the vector of tradeoff solutions among detectors with regard to two independent objectives of best classification fitness and optimal hyper volume size. The detectors monitor network traffic for exact and variant abnormal system events based on only the detector’s own data structure and the application domain truth set and tested with KDD Cup99 data. Wu Yang et al (2007) introduced an algorithm called Fast Inductive Learning Method for Intrusion Detection (FILMID). The FILMID could meet the requirement of detection accuracy and speed in handling huge network traffic data in high-speed network.
Terran Lane and Carla Brodley (2007) suggested Instance Selection Techniques on the basis of Instance Based Learning (IBL) techniques. The approach transformed temporal sequences of discrete into a metric space by means of a similarity measure that encodes intra-attribute dependencies. The classification boundaries are elected from an a posteriori characterization of normal behaviors, coupled with a domain heuristic. The clustering technique is based on sequential and greedy selection of clusters and it is able to produce a large saving in storage requirements with small loss in accuracy. Furthermore, the approach detects anomalous conditions quickly before the potential damage by a malicious one. Alexander Widder et al (2007) described an approach to detect unknown event patterns where an event not only happens, but also can be analyzed. The Complex Event Processing engine method is processed for detecting known pattern of events and correlating them to complex events in real time IDS. In the work, the discriminate analysis was used for recognizing unknown patterns for the use of credit card transactions.

Abdun Naser Mahmood et al (2008) presented a hierarchical clustering and summarization algorithm called Echidna for identifying structure of traffic attributes such as IP addresses, in combination with mixed type attributes including numerical, categorical and hierarchical attributes for a one-pass hierarchical clustering algorithm. The approach addresses the problems in previous approaches of network traffic analysis and has the advantage of improving accuracy and considerably reduces the computation time on a benchmark data. The artificial immune systems, genetic algorithms, genetic programming, and neural networks for a stronger, more robust security system have been discussed and analyzed by Winard Britt et al (2007). Different artificial intelligence techniques for specific applications with algorithm are described. Both the positive and negative aspects of the techniques are synthesized and the necessity of hybridization was recommended for defense systems.
John Felix Charles Joseph et al (2010) proposed a method, namely Cross layer based Adaptive Real-time routing Attack Detection System for adapting the intrusion detection model at real time with two major stages. In the first stage, the occurrence of new patterns in the routing control traffic was identified and prioritized based on their information content. The second stage of adaptation was to incrementally update the detection model using the new patterns with minimum computational overhead. Zanero (2008) presented a tool for network anomaly detection and network intelligence called ULISSE (Unsupervised Learning IDS with 2Stage Engine). The two tier architecture with unsupervised learning algorithms performs the network intrusion and anomaly detection. ULISSE utilizes a combination of clustering of packet payloads and correlation of anomalies in the packet stream. Furthermore, the ULISSE is combined with host based system of Syscall Sequence and Arguments Anomaly Detection Engine and is used to filter out false positives automatically and to improve correlation and alert quality.

Peng Ning et al (2004) proposed a method that correlates alerts by matching the partial consequences of some prior alerts with the prerequisites of some later ones. Furthermore, a set of interactive analysis utilities are used for the investigation of large sets of intrusion alerts. A toolkit named Toolkit for Intrusion Alert Analysis is explained which provides system support for interactive intrusion analysis and the techniques are experimented with the DARPA data. Kapil Kumar Gupta et al (2010) suggested Conditional Random Fields and Layered approach for accuracy and efficient IDS. The experimental results on the KDD Cup99 data showed that the attack detection accuracy is improved very high for the User to Root attacks (34.8 %) and the Remote to User attacks (34.5 %).
The fuzzy IF–THEN rules method (Chi-Ho Tsang et al. 2007) was an enhancement perception of agent based evolutionary framework and multi-objective optimization. In addition, it acts as a genetic feature selection wrapper to search for an optimal feature subset for dimensionality reduction. The classification evolution and feature selection performance have been compared with some well-known classifiers and feature selection filters and wrappers. Due to the classification accuracy, the data mining approaches have been adopted for automatic generation of rule-based knowledge. The experimental results on the KDD Cup99 benchmark data shows that the technique using interpretable fuzzy systems is found to be better than other classifiers and wrappers by providing the highest detection rate for attacks and low false alarm rate for normal network traffic with minimized number of features. Gautam Singaraju et al. (2004) presented a method, namely TIDE (Testbed for evaluating Intrusion Detection Systems) which provided a quantitative analysis using fuzzy logic under varying network loads and robust metrics to evaluate IDS. The TIDE has been built with plug and play architecture and the user can easily introduce or remove scripts depending on the environment and for the new attacks more illegitimate traffic scripts need to be added into the system. The false alarm rate IDS activities have been reduced with fuzzy logic and data mining algorithm (Jian Guan et al. 2004). A set of fuzzy rules can be used to classify the normal and abnormal activities in a computer network, and fuzzy data mining algorithms can be applied over fuzzy rules to determine when an intrusion is in progress. The mine fuzzy association rules and fuzzy frequent episodes are explained for anomaly detection from audit data and the experimental results indicate that fuzzy data mining can provide effective approximate anomaly detection.

Ajith Abraham et al. (2007) evaluated three fuzzy rule based classifiers to detect intrusions in a network and compared the results with other machine learning techniques like decision trees, SVM. Further, the
Distributed Soft Computing-based IDS (D-SCIDS) model has been suggested as a combination of different classifiers to model light weight and heavy weight IDS. The lightweight Soft Computing-based IDS (SCIDS) would be useful for Mobile Ad hoc Network or distributed systems and the heavy weight SCIDS would be ideal for conventional static networks or wireless base stations. The experimental results indicated that soft computing approach could play a key role for intrusion detection.

The port scanning attacks are detected by combining and analyzing various traffic parameters (Wassim El-Hajj et al 2008). The parameters could not be simply combined using a mathematical formula. The fuzzy logic is used to combine and reduce the number of false alarms. A fuzzy logic controller is designed and integrated to enhance the functionality of port scanning detection in wired and wireless networks. Bharanidharan Shanmugam and Norbik Bashah Idris (2007) proposed a dynamic model of intelligent IDS based on AI. The methods that are being examined include fuzzy logic with network profiling, which uses simple data mining techniques to process the network data. The proposed hybrid system combines both anomaly and misuse detection with simple fuzzy rules to describe security attacks experimented with DARPA data.

Yihua Liao and Rao Vemuri (2002) proposed an approach based on the k-Nearest Neighbor (kNN) classifier which is used to classify the program behavior as normal or intrusive. In the kNN classifier, the frequencies of system calls have been used to describe the program behavior. Also, text categorization techniques were adopted to convert each process to a vector and calculate the similarity between two program activities. Hence, there is no need to learn individual program profiles separately and so more calculations have been mainly reduced. The approach has been experimented with 1998 DARPA data. Gang Wang et al (2010) proposed an IDS approach on the KDD Cup99 data, called FC-ANN, based on ANN and fuzzy flustering, to
achieve higher detection rate, less false positive rate and stronger stability. Fuzzy clustering technique was used to generate different training subsets, and different ANN models have been trained to formulate different base models. A meta-learner and fuzzy aggregation module are used to aggregate the results.

Liwei Kuang and Mohammad Zulkernine (2008) suggested an anomaly intrusion detection method named Combined Strangeness and Isolation measure K-Nearest Neighbors (CSI-KNN). The CSI-KNN algorithm analyzed different characteristics of network data by employing strangeness and isolation measures. An association unit was designed for intrusion alerts with two measures and their associated confidence estimates. Both the strangeness and isolation models provided a confidence score that estimates the reliability of the prediction and the value of the confidence indicated the unlikelihood of a predicted intrusion being normal. A higher confidence specifies that the prediction is more likely to be correct. The hybrid learning model based on the Triangle Area based Nearest Neighbors (TANN) method was proposed by Chih-Fong Tsai and Chia-Ying Lin (2010). In TANN, the k-means clustering was used to obtain cluster centers corresponding to the attack classes respectively, and triangle area by two cluster centers with one data from the given data has been calculated and a new feature signature of the data is formed. Finally, the kNN classifier was used to classify similar attacks based on the new feature represented by triangle areas. The approach has been experimented on KDD Cup99 data.

2.3 BOOSTING BASED INTRUSION DETECTION SYSTEM

Boosting is a method to develop the accuracy of several given learning algorithms by combining weak classifiers to get one strong classifier.
Weiming Hu et al (2008) suggested an Adaboost based intrusion detection algorithm, in which the decision stumps are used as weak classifiers and the decision rules are provided for both categorical and continuous features. By combining the weak classifiers for continuous features and the categorical features into a strong classifier, the relations between these two different types of features are handled without any unnatural conversions. The performance of the algorithm is improved with adaptable initial weights and a simple strategy for avoiding over fitting. The algorithm has low computational complexity and error rates have been tested on the benchmark sample data. Devi Parikh and Tsuhan Chen (2008) have used an ensemble of classifiers approach from multiple sources which is tuned to minimize the cost of the errors as opposed to the error rate itself. The method dLEARNIN, which utilizes an ensemble of classifiers approach that combines information from different sources of information, and a cost minimization strategy dCMS minimizes the cost of the errors and not the error rate itself.

Jun Gao et al (2009) presented a distributed IDS framework based on the Model based online Adaboost algorithm and Particle Swarm Optimization-SVM algorithm. Experimental results demonstrated that the approach could achieve a good detection performance. The detection of distributed attacks across cooperating enterprises was suggested by Deborah and Frincke (2000). The relationships between cooperative hosts are defined and then take-grant model has been used to identify the attack. Also, a brief explanation of both a simple prolog model incorporating data sharing policies and a prototype cooperative intrusion detection system was included.

Youchan Zhu and Ying Zheng (2008) explained that the combination of Adaboost classification algorithm and nonlinear dimensional reduction Isomap algorithm have a very low false alarm and high detection rate than the use of Adaboost alone. Xuchun Li et al (2005) proposed the
Adaboost SVM algorithm, in which SVM is used as weak learners for Adaboost algorithm. It adaptively adjusts the kernel parameter in SVM instead of using a fixed one and provides a way to handle the over-fitting problem of Adaboost. It has the advantages of easier model selection and better generalization performance when compared with the existing Adaboost methods. Also, an enhanced method called Diverse Adaboost SVM deals with some parameter adjusting strategies. The distributions of accuracy and diversity over these SVM weak learners are tuned to achieve a good balance and saves computational cost.

2.4 DATA MINING AND RULE BASED INTRUSION DETECTION SYSTEM

In data mining approaches, IDS comprises association rules and recurrent episodes, which are based on constructing classifiers by discovering appropriate patterns of program and user performance. The association rules and frequent episodes are used to study the patterns that illustrate user behavior. These methods can deal with symbolic data, and the features can be defined from a group of packets and connection information. However, mining requires the huge number of records and requires a large number of rules that increase the complexity of the system.

In general, interpretability of the decision tree method (Kabiri and Ghorbani 2005) is used for classification in different domains that handling huge data sets and can be represented in terms of a rule set. In the tree, each leaf node signifies a class and is interpreted by the path from the root node to the leaf node in the form rule. For example “If attribute1 and attribute 2 and attribute 3, then class1” where attribute 1, attribute 2, and attribute 3 are the elements concerning the attributes and class1 is the one of the class label. Thus, a set of rules are used to express each class. The decision tree technique (Li and Ye 2001) can be transformed into IF-THEN rules by outlining the
path from root node to each leaf node in the tree. Also, logical functions have been implied between each of the extracted rules because the rules directly obtained from the tree and no similar two rules would be generated.

Distributed multiagent intrusion detection system architecture (Mei-Ling Shyu et al 2007) is proposed to provide an accurate and lightweight solution to network intrusion detection. The Adaptive Sub-Eigenspace Modeling based anomaly and misuse detection schemes are also employed for the intrusion detection task and give high detection rates and low false alarm rates. The host layer in the architecture with lightweight host agents executes anomaly detection in network connections to their respective hosts. The classification layer performs misuse detection for the host agents, detects distributed attacks, and publishes the network security information to the entire network. Zhenwei Yu et al (2007) proposed an Automatically Tuning IDS (ATIDS) system which automatically tunes the detection model on-the-fly according to the feedback provided by the system operator when false predictions are encountered. The system is assessed using the KDD Cup99 intrusion detection data and the system achieves up to 35% improvement in terms of misclassification cost when compared with a system lacking the tuning feature. The detection model is in the form of rule sets, which are easily understood and controlled. A hybrid IDS was suggested by Duanyang Zhao et al (2010), in which data mining programs are used to learn rules that accurately capture the behavior of intrusions and normal activities.

Jingmin Zhou et al (2007) explained a logical model that abstracts the logical relation between the alerts in order to support automatic correlation of those alerts involved in the same intrusion. The logical formula called capability is used to abstract consistently and accurately all levels of illegal access obtained by the malicious node in all levels of a multistage intrusion. Also, inference rules to define logical relations between different
capabilities are derived. Based on the model and the inference rules, several new alert correlation algorithms have been developed and implemented with a prototype alert correlator. The alert correlator effectively correlates more alerts involved in massive scanning incidents. Based on the distributed analysis employed by agent-based Distributed intrusion detection system (DIDS), Yu-Fang Zhang et al (2005) proposed the clustering based intrusion detection technique which overcomes the drawbacks of relying on labeled training data. For effectively choosing the attacks, twice clustering scheme has been employed: the first clustering is to choose the candidate anomalies at agent IDS and the second clustering is to choose the true attack at central IDS. Also, the KDD Cup99 data records were verified and the authors suggested the merging of clustering technique with real homogeneous distributed system.

Young Ryu and Hyeun-Suk Rhee (2008) presented intrusion detection method under an inspection constraint in which the information security officer’s limitations in incident inspection could be analyzed. The analysis could provide the system administrators and information security officers with a better understanding of operational and managerial characteristics of intrusion detection and prevention systems. Yun Huang et al (2007) discussed two components of solutions to Distributed Denial of Service attacks: cooperative filtering and cooperative traffic smoothing by caching and analyzing the broken incentive chain in each of the solutions. Furthermore, the usage based pricing and Capacity Provision Networks facilitate victims to disseminate enough incentive along attack paths to stimulate cooperation against attacks. The Intrusion System Preventions (ISPs) and the solutions initiated usage based pricing mainly provide for better Quality of Service in transmitting legitimate.
Alexander Hofmann et al (2007) proposed absolutely a distributed intrusion detection system based on Distributed Hash Tables, which efficiently exchanges and aggregates alerts and meta-alerts in a cooperative, self-organizing and load balanced way. The self-regulating intrusion detection instrument publishes the alerts based on a new measure for alerts which forbids the distribution of already known and hence, worthless knowledge. The benefits of the approach are assessed for the probing attack. The Distributed Security Operation Center (DSOC) architecture was developed by Abdoul Karim Ganame et al (2008) and it is able to detect attacks occurring simultaneously on several sites in a network and it gives a global view of the security of that network. The architecture has Global Analyzer (GA) and a number of Local Analyzers (LAs). The DSOC approach used around 443 times less bandwidth than a centralized SOC because the intrusions were detected by the LA. The alerts were merged and correlated before being transmitted to the GA. The DSOC alerts from heterogeneous platforms on several sites to detect more complex malicious activities.

Wenke Lee and Salvatore Stolfo (2000) described the framework called Mining Audit Data for Automated Models for Intrusion Detection (MADAM ID) using data mining algorithms to work out the action of patterns from audit data and extracting the predictive features from the patterns. The machine learning algorithms were applied to the audit records and processed according to the feature definitions to generate intrusion detection rules. The authors discussed the practice of converting the detection models produced by off-line data mining programs into real-time modules of existing IDSs. Also, a combined detection model is constructed by learning the correlation of intrusion evidence from multiple detection models called Meta-learning technique and the framework algorithms can incorporate the user-defined cost factors and policies to compute cost-sensitive ID models in the future work.
Costantina Caruso and Donato Malerba (2007) proposed the methodology based on the application of several data mining methods. The real monitoring units of the network connections are transformed into symbolic objects and the normal model itself is given by a particular set of symbolic objects. A new symbolic object is considered as attack, if it is dissimilar from those belonging to the model. It can be added to the model, if it is ranked as a changing point. The obtained model of network connections can be used by a network administrator to identify deviations in network traffic patterns. Bruce D. Caulkins et al (2005) formed a decision tree using the Chi-Square splitting principle and modeled with the data from the Lincoln Laboratory data set off-line evaluation (IDEVAL). They concluded that the number of leaves required to maximize the profit of the model was around four or five leaves. In their work, the first tier would be a signature based filter and the remaining packets would routed to proposed anomaly based filter.

Ramana Rao Kompella et al (2007) examined many of DoS and scanning attacks and proposed a data structure called partial completion filters (PCFs) that can detect claim-and-hold attacks scalably in the network. Also, the authors have been analyzed the PCFs both analytically and by means of experiments on real network traces to reveal how the PCFs to attain extremely low false positive and false negative probabilities. The approach has been compared with network telescopes approach for scalable monitoring, followed by detection of scanning activity. The combined techniques of data mining and expert systems were used to design effective anomaly based IDS (Sodiya Longe and Akinwale 2004). They give better coverage with more detection rate. The data mining audit data keeps the normal behaviors in profiles and the expert system performed the detection of anomalies with alarm. The periodic profiling makes the design dynamic and updates the users’ activities. Satnam Singh et al (2009) proposed the transaction-based probabilistic model which is a combination of Hidden Markov Models
(HMMs) potential and feature-aided tracking for the detection of asymmetric threats. This method is able to detect the modeled pattern of an asymmetric threat with a high performance capability as compared to other data mining techniques.

The data mining techniques are useful in IDS to detect the attack especially in anomaly detection and the decision tree method is an excellent method for the data mining approach (Joong-Hee Leet et al 2008). They used the Defense Advanced Research Projects Agency (DARPA) data as the training data and the testing data for the decision tree. Zhan Jiuhua (2008) proposed the IDS based on data mining technique. The system demonstrates diversity, adaptability, robustness and light weight. The method can detect intrusions of known and unknown attacks with reduced omissions and misstatements. Also, the method has improved accuracy, more speed of intrusion detection with good adaptive capacity and scalability.

Suratna Budalakoti et al (2009) proposed Sequence Miner algorithm that detects and characterizes anomalies in large sets of high dimensional symbol sequences that arise from recordings of switch sensors in the cockpits of commercial airliners. The anomaly detection approach uses unsupervised clustering of sequences using the normalized length of the longest common subsequence as a similarity measure, followed by detailed outlier analysis. An outlier sequence was defined as a sequence that is far away from the cluster center and the algorithms for outlier analysis provide comprehensible indicators as to why a particular sequence is deemed to be an outlier.

Nor Badrul Anuar et al (2008) suggested a strategy of hybrid statistical approach which uses data mining and decision tree classification to focus on statistical analysis of both attack and normal traffics based on the training data of KDD Cup99. The statistical analysis is used to reduce
misclassification of false positives. The post processing filter was used to reduce false positives in network based IDS (Georgios Spathoulas and Sokratis Katsikas 2010). The filter components are based on statistical properties of the input alert set and special characteristics of alerts corresponding to true attacks. The alerts were observed in batches, which contain similarities in the source or destination IPs or any abnormalities in the distribution of alerts of the same signature. The filter limited false positives up to 75%.

Stefan Axelsson (2000) defined that rule modeling is the system itself observes the traffic and frames various rules that describe the normal operation of the system. The rules would be applied by the system in the detection stage and hoists the alarm if the observed traffic forms a poor match with the framed rule. Likewise, the system has collected the required statistics, the user can program predefined thresholds which may be in the form of simple ranges that describe whether to hoist the alarm or not.

### 2.5 OTHER METHODS USED IN INTRUSION DETECTION SYSTEM

The Multiple intrusion detection system (MIDS) approach was developed by Tung Le and Christoforos Hadjicostis (2008). It operated by running belief propagation on an appropriately constructed weighted bipartite graph. The different types of possible intrusions are represented by one set of nodes and the other set of nodes represents the set of available significant measures. The weighted connections represent the dependence of a certain measure on a particular type of intrusion. The key factor of MIDS is the development of a modified belief propagation Max-product Algorithm (MPA) that avoids the exponential complexity of the original MPA by limiting, during the iteration process, the number of active intrusions that are connected to a particular measure. The simulation results show that the MIDS performs
well in detecting both single and multiple intrusions with a very low false alarm rate.

Anna Sperotto et al (2010) prepared a survey of current research in the area of flow-based intrusion detection. The theory of flows was explained and relevant standards have been identified. The authors have also presented a classification of attacks and defense techniques in detection of attacks. Chenfeng Vincent Zhou et al (2010) reviewed challenges in Collaborative IDSs architectures and alert correlation algorithms. Nong Ye et al (2006) classified the attacks in a comprehensive and sensible format and explained the System-Fault-Risk framework for cyber attack classification. The base of the framework is on a scientific foundation, combining theories from system engineering, fault modeling, and risk-assessment. They suggested that an organization should understand the attacks by assessing them using the metrics like attack characteristics, activities, state and performance impact that helped in choosing effective barriers and should decide the deployment position of such barriers.

Zhichun Li et al (2010) suggested a High-speed Flow level Intrusion Detection system (HiFIND). The HiFIND was scalable to flow level detection on high-speed networks which allowed aggregate detection over multiple routers / gateways and separated anomalies to limit false positives in detection. A Lightweight Network Intrusion Detection system (LNID) was proposed by Chia-Mei Chen et al (2010) for detecting attacks on Telnet traffic. It distinguished normal traffic behavior and computed the anomaly score of a packet based on the deviation from the normal behavior. Instead of handling out all traffic packets, an efficient filtering scheme was used to reduce system workload.
Jacob Ulvila and Gaffney Jr (2003) presented a comprehensive method for evaluating IDS, which incorporates and extends the Receiver Operating Characteristics curve and cost analysis methods to provide the expected cost metric. The method illustrated to find the optimum operating point for a single and double IDS, based on the costs of mistakes, and the enmity of the operating atmosphere is represented in the prior probability of intrusion. Also, the expected cost of single and double IDSs has been computed. Further, the representation of compound IDS as single IDS has been described. Mizuki Oka et al (2004) proposed a method, called Eigen Co-occurrence Matrix (ECM) that models sequences of UNIX commands, and extracts their principal features. The ECM method models a sequence by correlating an event by the distance between events and the frequency of their occurrence.

The PCA was used to extract the principal features (Steven Scott 2004) which have dealt with the latent variable hierarchical models created using Bayesian methods and lead to coherent systems that can handle various complex distributions involved with network traffic. The intrusion detection systems on stochastic models combined the user and intruder behavior using Bayes theorem. Bayesian methods minimize the complexity by smashing the complicated user and intruder models into manageable pieces. The network complexity was handled using hierarchical models that allow designers to concentrate on individual model components, and graphical models are a good choice for modeling the distribution of individual network transactions.

Azzedine Boukerche et al (2004) suggested that the security system technique for detection of intruders and improper use of both computer system and mobile telecommunication operations is based upon data analysis inspired by the natural immune human system. The scheme took significant features of the immune human system and mapped them with predefined
security system and unusual activities according to the usage log files. The system also reduces the size of the log files and the size of the report maintained by the system manager significantly and efficiently.

Marek Ostaszewski et al (2008) have discussed an intrusion detection architecture based on Idiotypic Network Theory (INIDS) that deals with large scale network attacks featuring variable properties. The dynamic and adaptive clustering of the network traffic was performed for effective countermeasures against DoS attacks. INIDS have been experimented on the bench mark of KDD Cup99 data. Moses Garuba et al (2008) analyzed several organizational security objectives in order to determine effective network intrusion detection system. The heuristic based systems provided better organizational objectives than signature based systems. The analysis is based on the systems which provide definitive security objectives and offer flexibility, adaptability, and reduced vulnerability that an organization requires.

Zhuowei Li et al (2005) proposed a theoretical method to analyze the existing problems for intrusion detection in a quantified manner. The root causes of the problems are identified as model inaccuracy and model incompleteness as well as the lack of distinguishing capability in the features utilized. Chin-Tser Huang et al (2008) applied signal processing techniques in intrusion detection systems. A framework called Waveman for real time wavelet-based analysis of network traffic anomalies is implemented. The two metrics, namely percentage deviation and entropy are used to evaluate the performance of various wavelet functions on detecting different types of anomalies like DoS attacks and portscans. The evaluation results indicate that Coiflet and Paul wavelets performed better than other wavelets in detecting anomalies.
Mahbod Tavallaee et al (2010) investigated various issues in IDS and reviewed the present state of the experimental observation in the area of anomaly based intrusion detection. The authors surveyed 276 publications during the period of 2000 – 2008 and summarized the observations and also identified the common drawbacks among surveyed works. Zhongqiang Chen et al (2010) proposed an Intrusion Prevention System (IPS) evaluator which has a trace-driven inline simulator engine and mechanisms for generating and manipulating test cases and a comprehensive series of test procedures. Also, algorithms have been proposed to partition attacker / victim-emanated packets. Alvaro Cardenas et al (2006) analyzed the advantages and disadvantages of the metrics, the detection rate, the expected cost, the sensitivity and the intrusion detection capability have been used to evaluate the performance and the best configuration of IDS. In addition to these metrics, the Intrusion Detection Operating Characteristic (IDOC) curves metric has been introduced as an IDS performance tradeoff. Also, a formal framework for the performance of IDS using the above metrics against adaptive adversaries would be experimented. The authors suggested that the IDOC curves method was not only applicable to IDSs but also to any classification algorithm whose classes are heavily imbalanced.

Otey et al (2003) suggested the approach of network intrusion detection by means of Network Interface Cards (NICs) and its feasibility. The intrusion detection using the NIC makes the system highly protective and can be extended to a distributed environment. The anomaly detection and signature detection based models have been implemented in NIC firmware and the preliminary results show that NIC based security schemes could work very well for the network security systems. The authors also suggested the hybrid models that use resources at both the host and NIC level. Due to memory or speed constraints of the NICs, operations can be performed at the host level, and their results can be sent to the NIC to aid in detection.
Furthermore, both signature detection algorithms and anomaly detection algorithms are embedded for evaluation.

In order to protect high speed Ethernet network from malicious attacks, Young Cho and William Mangione-Smith (2008) designed a high-performance scalable pattern matching architecture for Field-Programmable Gate Arrays (FPGAs) of less cost. Furthermore, architecture enables high-performance NIDS, and the pattern matcher can be used in the field of networking. Also, FPGAs have the ability to reprogram the filter when there are any changes to the signature set. Nitesh Guinde and Sotirios Ziavras (2010) suggested Field-Programmable Gate Array and Application-Specific Integrated Circuit based solution which performed high-speed pattern matching while permitting pattern updates without resource reconfiguration. The approach reduced significantly the required on-chip storage as well as the complexity of pattern matching.

An intrusion response decision making model based on Hierarchical Task Network (HTN) planning has been presented by Chengpo Mu and Yingjiu Li (2010). The self adaptive response decision making model has the ability of tolerating false positive IDS alerts. The proposed model has been used in the intrusion detection alert management and intrusion response system. Zubair Fadlullah et al (2010) proposed Detection and TRAceBack to fight against attacks on encrypted protocols using strategically distributed Monitoring Stubs (MSs). Various attacks have been categorized against cryptographic protocols. The MSs, by sniffing the encrypted traffic, extract features for detecting attacks and create normal usage behavior profiles. In detection of any activities due to the deviations from these normal profiles, the MSs notify the victim servers, which take necessary actions. The detection and traceback methods are verified through internet data.
2.6  OVERVIEW OF DARPA DATA

John Mchugh (2000) analyzed many issues that were created with the design of DARPA data execution. The author endeavors to identify the deficiency of the Lincoln Lab effort and hopes that issues would be avoided in future evolution. Many IDS researchers point out the progress, especially in its evaluation amount and some others find fault that the program has a serious unfavorable impact on the research efforts. But the Lincoln Laboratory approach of investigating missed detections and false alarms possibly is much more useful in improving IDS systems. The 1998 approach offered an approximation for a suitable unit of analysis, but the 1999 evaluation provided no unit of analysis and only reported false alarm results.

The KDD Cup99 is one of the most widely used data for the evaluation of IDSs and some drawbacks were found during the statistical analysis. Mahbod Tavallaee et al (2009) proposed the solution for the drawbacks with a new data, called NSL-KDD. The NSL-KDD consists of selected records of the complete KDD Cup99 data by avoiding the huge number of redundant records, which causes the learning algorithms to be biased towards the frequent records, and thus puts them off from learning infrequent records which are usually more harmful to networks.

2.7  SUMMARY

This chapter brought out a detailed survey on the various intrusion detection techniques based on Support Vector Machines, neural network, fuzzy, genetic algorithm, machine learning methods, boosting, data mining, and rule based methods. The study based proactive and dynamic model, k-nearest neighbor, Bayes’ theorem based classifier have also been listed. Although numerous techniques have been presented by many researchers to build efficient IDS, it is found that each one has a unique advantage regarding
classification, false alarm rate and time complexity. There is still more scope for research in IDS due to varieties of application domains and numerous users accessing them. Though there are rule based IDS filters available, this thesis proposed a rule based Multi Stage Filter (MSF) to demonstrate the effectiveness of the work. Most of the work carried out was based on KDD Cup99 data. The next chapter deals exclusively with the analysis of the dataset used in the proposed work.