Diverse and Conglomerate Modi-operandi for Anomaly Intrusion Detection Systems

A. M. Chandrashekhar
Asst. Professor / Research scholar,
Dept. of Computer science & Engineering,
S. J. College of Engineering (SJCE),
Mysore-570006, Karnataka – INDIA

K. Raghuveer
Professor,
Dept. of Information Science & Engineering,
National Institute of Engineering (NIE),
Mysore-570008, Karnataka – INDIA.

ABSTRACT

Of late, research works on Intrusion Detection System have been receiving a lot of attention. An IDS detects hazard patterns of network traffic on the residual open parts through observing user activities [1]. There are several models available as of now, but the major loop hole in most of the existing models is the incapability of cognizing new attacks i.e. novel threats to a system. Anomaly based intrusion detection system has undoubtedly resulted in easing the pain of detecting novel threats for a system when compared to its counterpart, Signature based Intrusion Detection System. This paper gives an overview of various Anomaly Intrusion Detection System techniques like machine learning algorithms, data mining methods and its variants e.g. Entropy data mining, neural network methods etc. We also give an overview of a few hybrid techniques that have been employed and have resulted in better outcomes for e.g. a combination of Neural networks and Fuzzy logic method.

General Terms

Security, Network Security, Computer Security, Intrusion, Threat, Attack.

Keywords

Intrusion detection, Data mining, Anomaly intrusion detection, Neural Networks, fuzzy logic.

1. INTRODUCTION

Intrusions are activities that try to compromise with the security requirements of the system mainly Confidentiality, Integrity and Authentication. With increased internet usage, information flow, trade and business activists over internet, the threat levels and the aftermath effects have become indefinable. Intrusion Detection Systems have become an essential integral part of a security package for modern computer systems. Therefore it has become a research focus in computer security domain. Intrusions Detection Systems detects unauthorized or malicious attacks over a computer system which occurs primarily through internet. The intrusion detection technology uses the trace information which is left by the intruder.

IDS can be basically classified into network based and host based. This classification is done based on the information source of the detection and also as a matter of fact that Network based IDS monitors activities happening on a network, where as Host based IDS monitors activities happening on a system. They are further roughly classified into Misuse IDS and Anomaly IDS. Misuse IDS also known as Signature based IDS is the most widely deployed method. The detection is done based on the known patterns of malicious activities in a system. A threat is known in prior and from there on it is a simple pattern matching process. Apparently, a new threat signature is not present in the threat repository which makes Misuse IDS incapable of detecting a novel threat. Anomaly IDS studies the system's normal behavior and looks out for any kind deviants activities and thus detects the intrusions. Thus Anomaly IDS can detect a threat right from its very first occurrence. In order to meet up to the standards in research works, IDS are evaluated based on its performance on benchmark datasets like DARPA datasets and KDDcup99 dataset.

DARPA (Defense Advanced Research Project Agency) released two data sets in 1998 and 1999 respectively. The dataset includes various records of network log information collected at its most enticing sites. Its includes a Training sample of three weeks, which is free from all kind of attacks and a Testing sample of two weeks which includes enormous number of attacks. Since their release, they have been most widely used benchmark datasets. In recent time, researchers have resorted to using KDDcup99 dataset. It was created based on DARPA data set.

There are 39 attacker types that could be classified into 4 major categories, which we see in detail in the next section. It consists of approximately 4,900,000 data instances each of which is a vector of extracted feature values from connection record [3].

2. REVIEW OF VARIOUS THREATS AND INTRUSION DETECTION SYSSTEMS

There are four main categories of attacks [1]:

- *DOS (Denial of Service):* An attacker tries to prevent legitimate users from using a service e.g. TCP SYN Flood, Smurf. For example: Back Land Neptune, Smurf, Teardrop etc.
- *Probe:* An attacker tries to find information about the target host. For example: scanning victims in order to get knowledge about available services, using Operating System. For example: Ipsweep, Nmap, Portsweep etc.
- *U2R (User to Root):* An attacker has local account on victim's host and tries to gain the root privileges. For example: Buffer-overflow, Load module, Perl root kit etc.
- *R2L (Remote to Local):* An attacker does not have local account on the victim host and try to obtain it. For example: ftp-write, guess-password, imap, multihoop etc.

IDS are vaguely categorized into five types [2]:

- *Network IDS (NIDS):* responsible for detecting attacks related to the network. NIDSs investigate incoming and outgoing network traffic by connecting with network devices to find suspicious patterns. If a NIDS has no additional information about the protected host, the malicious attacker can easily avoid detection by taking advantage of different handling by overlapping IP/TCP fragments by IDS and a target host.

- *Host-based IDS (HIDS):* usually are located in servers to examine the internal interfaces. HIDSs can either use standard auditing tools, or specially instrumented operating system, or application platforms. It detects intrusions by analyzing system calls, application logs, file-system modifications, and other host activities related to the machine.

- *Protocol-based IDS (PIDS):* monitor the dynamic behavior and state of the protocol used the web server. PIDSs sit at the front end of a web server, monitoring and analyzing the HTTP protocol stream. It understands the HTTP protocol to protect web server by filtering IP address or port number.

- *Application protocol-based IDS (APIDS):* monitor and analysis on a specific application protocol or protocols between a process, and group of servers that is used by the computer system. APIDSs can be sitting between a web server and the database management system that monitoring the SQL protocol specific to the business logic. Generally, APIDSs look for the correct use of the protocol.

- *Hybrid IDS (HIDS):* combines two or more intrusion detection approaches. HIDS provide alert notification from both network and host-based intrusion detection devices.

An IDS has become an essential integral part of a security package for a modern computer system. It consists of several components. It needs a system that can sense activities and collect the alert log of system or a network. An effective analyzing system unit is required in the next stage that can analyze and return a synthesis or summary of the vast alert log. A database system is required for look up and storage purposes. Finally we need a decision unit that takes appropriate decisions on intrusions.

The following are the functionalities that are expected out from an Ideal IDS:

- Monitoring user's activity
- Monitoring systems activity.
- Auditing system configuration.
- Assessing the data files.
- Recognizing known attack.
- Identifying abnormal activity.
- Managing audit data.
- Highlighting normal activity.
- Correcting system configuration errors.
- Stores information about intruders.

In view of the above desired functionalities, The Common Intrusion Detection Framework (CIDF) tells that IDS must contain four major parts: Sensors, Analyzers, Database and Response Unit

3. INDIVIDUAL METHODOLIGES FOR ANOMALY IDS

There are several straight away techniques for Anomaly IDS which we discuss in this section. In the next section we discuss about some aggregate techniques for anomaly IDS.

3.1 Machine Learning Techniques

Machine learning techniques are adaptive algorithms that study audit data and generate profiles of normal system or network behaviors. Any deviating activities are termed as attacks. They make use of probability measures and event modeling for learning the system behaviors.

- *PHAD [Packet Header Anomaly Detection]:* PHAD detects anomalies in Network layer packet headers. For e.g. TCP, IP, Ethernet ICMP etc. It is a simple and efficient network intrusion detection algorithm that detects novel attacks by flagging anomalous field values in the Packer header of Network layer packets. PHAD learns the normal range of values for the fields in the packet header and in the next step any kind of deviations from the normal permitted values are termed as intrusions. PHAD was first proposed by Mathew V Mahoney. PHAD, when first demonstrated by Mathew V Mahoney and Philip K Chan resulted in finding 72 out of 201 instances of attacks. This experiment was conducted on 1999 DARPA offline Intrusion Detection evaluation dataset.

- *ALAD [Application Layer Anomaly Detection]:* ALAD detects anomalies in application layer payload. PHAD analyze the IP payload, whereas ALAD analyzes application payload. This algorithm detects novel attacks by flagging anomalous filed attributes in the application payload. For example FTP, Telnet etc. It extends the network model to the application layer. Instead of modeling single packets as in PHAD, here it model incoming TCP connections to the well known server ports (0-23). ALAD also introduces conditional rules that use conditional probability unlike PHAD. [5]

- *LERAD [Learning Rules for Anomaly Detection]:* LERAD is a drastic improvement over the Apriori versions of machine learning algorithms i.e. PHAD, ALAD. PHAD worked on network layer packet attributes and ALAD on application payload attributes, whereas LERAD detects anomalies in both Network and Application layer packet attributes. There are three main aspects of this algorithm which make it more efficient than the earlier versions. First, it model the application payload, a more difficult problem than modeling just IP addresses and port numbers, as most network anomaly detectors do. Second, it uses a non-stationary model, in which the time since an event last occurred is significant, and the frequency of occurrence is not. Third, it develops a randomized algorithm for finding the type of conditional rules that are most useful for anomaly detection.

3.2 Data mining

Data Mining, a fairly young and inter disciplinary field of computer science is the method of extracting patterns from huge data sets by combining techniques from statistics and artificial intelligence with database management. In simple terms Data mining can be defined as the process of extracting descriptive patterns from large data sets. The recent speedy development in data mining offered an extensive variety of algorithms, obtained from the fields of pattern recognition, databases, machine learning, and statistics [4].

There are several types of techniques available in data mining. Following are particularly of our interest and relevant to our intrusion detection task.

- Classification*:* This method maps a data item into one of the several predefined categories. These algorithms usually output "classifiers", for instance, in the form of decision rules or trees. An idyllic application in intrusion detection will be to collect sufficient "normal" and "abnormal" review data for a user or a program, and then apply a classification algorithm to study a classifier that will decide (future) review data as belonging to normal class or abnormal class [4].
- Link Analysis*:* This method determines relationships between fields in database. Discovering the associations in audit data will provide insight for determining the right set of system characteristic for intrusion detection [6].
- Sequence Analysis*:* This method models sequential patterns. These algorithms can assist us to understand what (time-based order of audit events are frequently encountered simultaneously). These frequent event patterns are vital elements of the behavior profile of user or program [6].

3.2.1 *Real Time Data mining [7]*

While generic Data mining techniques are applied on off-line data, real time data mining techniques aims at addressing the issues like accuracy, efficiency and usability. The architecture of such a system is a distributed one involving sensors, detectors, data warehouse and model generation components. The last mentioned component is the heart of this model.

Data mining programs used to analyze the audit information and extract behaviors that are abnormal in a system; this increases the accuracy of the system. To get better efficiency, the computational costs of features are examined and a multiple-model cost based approach is utilized to produce detection models with less cost and more accuracy. To improve usability, adaptive learning algorithms are utilized to facilitate model construction and incremental updates; unsupervised anomaly detection algorithms are employed to reduce the reliance on labeled data.

3.2.2 Entropy based Data Mining [8]

The previously discussed data mining based detection modi operandi used Apriori approach. This involves first building profiles for normal system behavior. Then anomalous deviations from the normal behavior are termed as attacks. In the first step profiles are extracted which are modeled from sequence of events in the system. The results of Apriori approach involve association rules with contradictions. In other words, the result of Apriori is noisy and the post-processing of the results is necessary in order to use those results in the intrusion detection systems.

In this approach we use Graph Based Induction (GBI) method for rule learning. This method uses the concept of mathematical theory of entropy to guide the search for association rules and can mine association rules with much more support and lesser number of contradictions. Thus, it doesn't have a defect, which is common to intrusion detection systems associated with the APRIORI based data mining method.

3.3 Neural Network

An artificial neural network comprise of a collection of processing components that are highly interconnected and renovate a set of inputs to set of desired outputs. The result of the transformation is found out by the characteristics of the components and the weights linked with the interconnections among them. Unlike expert systems, which can provide the user with a definitive answer if the characteristics which are reviewed exactly match those which have been coded in the rule base, a neural network conducts an study of the information and presents a probability estimate that the data matches the characteristics which it has been trained to recognize. While the probability of a match determined by a neural network can be 100%, the accuracy of its decisions relies totally on the experience the system gains in analyzing examples of the stated problem [9].

The neural network gains the knowledge initially by training the system to correctly identify preselected examples of the problem. The response of the neural network is reviewed and the configuration of the system is refined until the neural network's analysis of the training data reaches a satisfactory level. In addition to the initial training period, the neural network also gains experience over time as it conducts analyses on data related to the problem [9].

3.4 Unsupervised method

Proposed by Eleazar Eskin, Andrew Arnold et al. presents an algorithm that are designed to process unlabelled data. In the structure, data elements are mapped to a feature space which is normally a vector space R^d. anomalies are detected by determining which points lays in the sparse regions of feature space [3].

A large data set is used in unsupervised learning. The attacks are very rare and are buried deep within the data, so the major requirement of the proposed algorithm is the ability of processing large data. In addition, these algorithms can semi-automate the manual inspection of data [3].

These algorithms make two basic assumptions:

- The number of normal instances vastly outnumbers the anomalies.
- Anomalies themselves are qualitatively different from normal instances.

Often the assumptions don't hold good and algorithm is ineffective in these cases. Some other recent approaches also use similar technique these approaches inter- point distances between instances in the data to determine which points are outliers. The major difference between the others and the one proposed is the nature of the outliers.

3.5 Hamming Network Approach

Proposed by Muna M. Taher Jawhar *et al.* This model gives an evolving anomaly intrusion detection system built by using hamming and MAXNET Neural Network to recognize attack class in the network traffic [1].

It needs to have a stored prototype pattern set. Given a prototype, it finds the hamming distance for the input data set. The match with the minimal hamming distance is identified and reported. The hamming distance is the number of bits within an input pattern that does not match the corresponding bits in stored prototype [1].

There are three main components of the system:

- Data provider – KDDcup99 Data set was used as the data source.
- Pre-processor – processes the data to get it in the required format. Consist of three stages
    - Feature extractor
    - Encoding
    - normalization
- Hamming and MAXNET Network Classification - consist of two components*:*
    - Hamming net
    - MAXNET

3.6 Improved self adaptive Bayesian algorithm

Proposed by Dewan Md. Farid and Mohammad Zahidur Rahman last year the method presents a new approach to the alert classification to reduce false positives in intrusion detection using improved self adaptive Bayesian algorithm (ISABA) [2].

The Bayesian algorithm provides a probabilistic approach for classification, which provides an optimal way to predict the class of an unknown example [2]. Conditional probability is used to classify. The naïve Bayesian theorem has been improved by the authors to ISABA to address problems like classification rates and false positives.

The foundation of this algorithm lies on the Baye's rule. The Baye's rule provides a way to calculate the probability of a hypothesis based on its prior probability [2].

$$\text{Baye's rule : } P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Where: '*D'* is the observed data and '*h*' is hypothesis space containing possible target function. *P(h|D)* is called the posterior probability, *P(h)* is the prior probability associated with hypothesis *h, P(D)* is the probability of the occurrence of data *D and P(D|h)* is the conditional probability.

Adaptive Bayesian algorithm creates a function from KDD99 benchmark intrusion detection training data, which first estimate the class conditional probabilities for each attribute value based on their frequencies over the weights with match of same class in the training data [2]. ISABA is an improvement over Adaptive Bayesian Algorithm.

4.CONGLOMERATE METHODOLOGIES FOR ANOMALY IDS

Like we said earlier, IDS has become a focus point for many researchers in computer security domain. In the previous section we have discussed about various diverse techniques for anomaly IDS, where each method is eccentric on a single particular topic. Researchers have been trying out different method for better outcomes. A few hybrid techniques for Anomaly IDS have been proposed which are believed to show better detection rates and efficiency.

4.1 Decision Tree method in conjunction with Neural Network

The notion of anomaly IDS can be enhanced by using both Decision Trees (DT) and Neural Networks (NN). A decision tree is a decision support means that uses a tree like graph or representation of decisions and their possible consequences, as well as chance event outcomes, resource costs, and utility. DT is one means to display an algorithm. While DTs are highly successful in detecting known attacks, NNs are more interesting to detect new attacks. The hybrid of Self Organizing Map (SOM) as an unsupervised Neural Network based Intrusion Detection and supervised Neural Network based on Back propagation would be used for clustering and classifying of network unknown attacks [10].

Marjan Bahrololum, Elham Salahi et.al, First proposed this hybrid model as an effort to make improvements in the field of Anomaly IDS. They conducted various experiments on the MIT Lincoln's lab, a well known dataset. Back propagation is one of the most commonly used supervised neural networks algorithm. The main aim of using back propagation method is to train the network in order to achieve a balance between the ability to respond correctly to the input patterns that are used for training purpose and the ability to give reasonable responses to input that is similar to that used in training period. The training of a network by back propagation involves three stages: The feed forward of the input training pattern, the calculation and back propagation of the related error, and the alteration of the weights, so that the forward pass produces an output vector for a given input vector based on the current state of the network weights [10].

4.2 Neuro-fuzzy inference method

K.S. Anil Kumar and Dr. V. Nandamohan proposed a methodology with combination of three techniques such as K-Means Clustering, Fuzzy Logics and Neural Network, comprising two machine-learning paradigms [11] in 2008. The main advantage of the technique is reduced human interventions for false alarms.

The system uses different concepts for different purposes to get an efficient and user friendly system. For instance it uses K-Means to cluster normal and intrusion packets, Fuzzy logic for rule generation from the perceived traits of normal and abnormal clusters, and Neural networks for detecting abnormal packets and an artificial intelligence that detects anomalies not presented during training.

The three basic techniques are:

- K-Means Clustering Algorithm*:* As the name suggests this algorithm classifies objects into K groups based on their attributes or features. The cluster centroid is calculated first, and then the grouping is done by minimizing the sum of squares of distances between the data and the corresponding cluster centroid.

- Fuzzy logic: The term "fuzzy logic" exhibits greater degree of compatibility when dealing with the data or objects that are abstract in nature. For instance while speaking, we use terms like big or small which do not specify the exact size but gives us a fair idea about the size of the object. Here in the proposed technique the logic has been used to separate normal from abnormal behavior. Weight ages assigned to the packets decide the normality or abnormality of packets.

- Neural Network: The architecture of neural network comprises a system of programs and data structures that resemble the function and structure of human brain [9]. It can learn, differentiate and extract the relations and organize the patterns in the data fed.

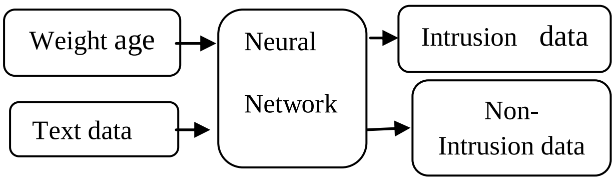

Fig-1: logical organization of the deployed neural network

5. CONCLUSION

This paper on the whole presents a survey on various Anomaly Intrusion Detection techniques. Most of the research work happening in present days is on Anomaly IDS techniques and Hybrid techniques, which as a matter of fact are the driving force for researchers. The generic method i.e. misuse detection method seems to be rather unsuitable for the kind of systems and networks used in this generation. 'Change is the essence of life' holds well in case of IDS also. As the pace of network growth goes high, security innovations with increased and enhanced features have to be made correspondingly. Thus, the trend of using hybrid techniques has started increasing because the yield factor is relatively high.

6. REFERENCES

[1] Muna M., Taher Jawhar and Monica Mehrotra" Anomaly Intrusion Detection System using Hamming Network Approach" International Journal of Computer Science & Communication, Vol. 1, No. 1, pp. 165-169, January-June, 2010.

[2] Dewan Md. Farid and Mohammad Zahidur Rahman, "Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm", Journal of Computers, Vol.5, No.1, January, 2010.

[3] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. J., "A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data", Applications of Data Mining in Computer Security, Kluwer Academic Publishers, pp. 78-99, 2002.

[4] W. Lee, S. Stolfo, and K. Mok, "A Data Mining Framework for Building Intrusion Detection Model", In Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA: IEEE Computer Society Press, pp. 120-132, 1999.

[5] Mathew V Mohoney and Philip K Chan, "Learning Non Stationary Models of Normal Network Traffic for detecting Novel Attacks" Proc. Eighth Int1. Conf. Knowledge Discovery and Data Mining, p376-385, 2002.

[6] Wenke Lee and Salvatore J. Stolfo, "Data Mining Approaches for Intrusion Detection", Proceedings of the 7th USENIX Security Symposium, San Antonio, Texas, January 26-29, 1998.

[7] W. Lee, S. Stolfo, P. Chan, E. Eskin, W. Fan, M. Miller, S. Hershkop, and J. Zhang, "Real Time Data Mining-Based Intrusion Detection", In DARPA Information Survivability Conference and Exposition II, June 2001.

[8] K.Yoshida, "Entropy Based Intrusion Detection", In Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and signal Processing, Vol. 2, pp. 840 – 843, Aug 28- 30 ,2003

[9] Cannady J, "Artificial Neural Networks for Misuse Detection", In Proceedings of the '98 National Information System Security Conference (NISSC'98), pp. 443-456, 1998.

[10] Marjan Bahrololum, Elham Salahi, and Mahmoud Khaleghi, "An Improved Intrusion Detection Technique based on two Strategies Using Decision Tree and Neural Network", Vol. 4, No. 4, pp. 96-101, 2009.

[11] K.S. Anil Kumar and Dr. V. NandaMohan, "Novel Anomaly Intrusion Detection Using Neuro-Fuzzy Inference System", IJCSNS International Journal 6 of Computer Science and Network Security, vol.8, no.8, pp.6-11 , August 2008.