A Probabilistic Cell Generation Based Improved Decision Tree Approach for Intrusion Detection

Shelly Sachdeva
Department of Computer Science and Engineering, Sat Priya Group of Institutions, Rohtak, Haryana, India.

Rupali Malhotra
H.O.D in Computer Science Department, Sat Priya Group of Institutions, Rohtak, Haryana, India.

Abstract – Communication is performed in some distributed communication network; it suffers from various kinds of internal attacks. To perform reliable communication over the network, there is the requirement to identify these attacks over the network. These attack detection systems are either host based or network based. In this present work, a network based system is defined to identify the attack based on communication statistics analysis. The work will be here defined in three main layers. In first layer, the Bayesian network will be applied to perform the relational probabilistic analysis. In second stage of work, a cell formation approach will be defined under the danger theory. According to this approach, the safe communication will be represented by safe cell and critical communication will be identified by danger cell based on attribute groups. In the final layer, the decision tree will be applied to perform the cell classification so that the attack class will be identified over the network.

Index Terms – DDOS, Intrusion Prevention, Safe Cell and Danger Cell.

1. INTRODUCTION

Intrusion prevention measures, such as encryption and authentication, can be used in ad-hoc networks to reduce intrusions, but cannot eliminate them. For example, encryption and authentication cannot defend against compromised mobile nodes, which often carry the private keys. Integrity validation using redundant information (from different nodes), such as those being used in secure routing, also relies on the trustworthiness of other nodes, which could likewise be a weak link for sophisticated attacks. To secure mobile computing applications, we need to deploy intrusion detection and response techniques, and further research is necessary to adapt these techniques to the new environment, from their original applications in fixed wired network.

1.1. Wireless Network Attack

1.1.1. Network discovery attacks

Wireless LAN discovery tools such as Net-Stumbler are designed to identify various network characteristics. Although the use of these tools in not characterized as a real attack, it aims at discovering as much useful information about the network as possible.

Eavesdropping/Traffic analysis:

Eavesdropping and traffic analysis attacks allow the aggressor to monitor, capture data and create statistical results from a wireless network. Since all 802.11 packet headers are not encrypted and travel through the network in clear text format they can be easily read by potential eavesdroppers.

Masquerading/Impersonation attacks

This category of attacks considers aggressors trying to steal and after imitate the characteristics of a valid user or most importantly those of a legitimate AP. The attacker would most likely trigger an eavesdropping or a network discovery attack to intercept the required characteristics from a user or an AP accordingly. Then, he can either change his MAC address to that of the valid user or utilize software tools like the well-known Host AP that will enable him to act as a fully legitimate AP.

1.1.2. Man-in-the-Middle attacks

The most advanced type of attack on a wireless or wired network is the "Man-In-The-Middle" attack. The attacker attempts to insert himself as middleman between the user and an access point. The aggressor then proceeds to forward information between the user and access point, during which he collects log on information. As a result, the attacker can maliciously intercept, modify, add or even delete data.

Denial-of-Service attacks

The main goal of Denial-of-Service (DoS) attacks is to inhibit or even worse prevent legitimate users from accessing network resources, services and information. More specifically, this sort of attack targets the availability of the network i.e. by blocking network access, causing excessive delays, consuming valuable network resources, etc. A denial of service occurs when an attacker has engaged most of the resources a host or network has available, rendering it unavailable to legitimate users. More specifically, this sort of attack targets the availability of the network i.e. by blocking network access, causing excessive delays, consuming valuable network resources, etc. [1][2].

A Defence mechanism in case of DOS attack is described by H.W.Chanl[3]. Analysis of DoS / DDoS attacks and actual test common hacker using a wireless network to hacking, cracking WEP in actual operation [4]. Bayesian Networks based DOS attack detection is performed by Wei Wang, Sylvani Gombault[5]. In case of Adhoc network a work is performed by Yinan Jing [6] using Trace back scheme.

1.2. Detection Approaches

Anomaly detection techniques establish a "normal activity profile" for a system; we could, in theory, flag all system states varying from the established profile by statistically significant amounts as intrusion attempts.

(1) Anomalous activities that are not intrusive are flagged as intrusive.

(2) Intrusive activities that are not anomalous result in false negatives.

The main issues in anomaly detection systems thus become the selection of threshold levels so that neither of the above 2 problems is unreasonably magnified, and the selection of features to monitor. Anomaly detection systems are also computationally expensive because of the overhead of keeping track of, and possibly updating several system profile metrics.

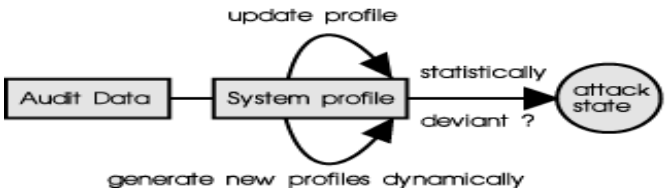

Figure 1 Anomaly Detection System

There have been a few approaches to anomaly intrusion detection systems, some of which are described below.

1.2.1. STATISTICAL APPROACHES

In this method, initially, behavior profiles for subjects are generated. As the system continues running, the anomaly detector constantly generates the variance of the present profile from the original one. There may be several measures that affect the behavior profile, like activity measures, CPU time used, number of network connections in a time period, etc. In some systems, the current profile and the previous profile are merged at intervals, but in some other systems profile generation is a onetime activity.

2. RELATED WORK

Varun Chandola(2010) has provided a reference adaptive architecture for controlling the communication under mining operation. The reference communication based model under sequence adaptive analysis is here performed to generate the category adaptive data. The sequence communication is here analyzed using RNA network so that the infected communication will be analyzed [7]. The multivariate analysis is able to provide more effective and accurate results song Song Hwang (2007) has provided mining approach using intrusion detection system. Author presented a three tier architecture is here defined using multi class specification under SVM approach. Author defined the black list node filters with attack specification so that the early detection of bad nodes will be done. Author provided the recognition under the attack specification. Author defined an architecture specific model for attack detection and to provide the safe communication in the network [8].In Year 2006, Carrie Gates performed a work, "Challenging the Anomaly Detection Paradigm a provocative discussion". In this paper Author question the application of Denning's work to network based anomaly detection, along with other assumptions commonly made in network-based detection research [9]. Authors examine the assumptions for selected studies of network anomaly detection and discuss these assumptions in the context of the results from studies of network traffic patterns. In Year 2011, ungsuk Song performed a work, "Statistical Analysis of Honeypot Data and Building of Kyoto 2006+ Dataset for NIDS Evaluation". In this paper, Author present a new evaluation dataset, called Kyoto 2006+, built on the 3 years of real traffic which are obtained from diverse types of honeypots. Kyoto 2006+ dataset will greatly contribute to IDS researchers in obtaining more practical, useful and accurate evaluation results. Author provide detailed analysis results of honey pot data and share Presented experiences so that security researchers are able to get insights into the trends of latest cyber-attacks and the Internet situations [10]. In Year 2012, A.S. Aneetha performed a work, "Hybrid Network Intrusion Detection System Using Expert Rule Based Approach". In this paper Author has proposed a new frame work based on a hybrid intrusion detection system for known and unknown attacks in an efficient way. This frame work has the ability to detect intrusion in real time environment from the link layer. The detection rate of the hybrid system has been found to increase as the unknown attack percentage increases whereas in misuse, detection rate is found to decrease and in anomaly detection rate remains constant [11]. In Year 2011, Huu Hoa Nguyen performed a work, "An Efficient Local Region and Clustering-Based Ensemble System for Intrusion Detection". In this thread, Author propose a novel design method to generate a robust ensemble-based IDS[12]. In Presented approach, individual classifiers are built. In Year 2009, Mrutyunjaya Panda performed a work, "Ensemble of Classifiers for Detecting Network Intrusion". In this paper, Author present an intrusion detection model based on Ensemble of classifiers such as AdaBoost, MultiBoosting and Bagging to gain more opportunity of training misclassified samples and reduce the error rate by the majority voting of involved classifiers. Presented main goal is to build an efficient intrusion detection model based on the error rate of classifiers if unfair distribution exists either within or between data sets

[13]. In Year 2012, Kola Sujatha performed a work," Network Intrusion Detection System using Genetic Network Programming with Support Vector Machine". Nowadays Internet Services spread all over the world. In this work dataset is classified into two datasets; namely positive kernel and negative kernel. Positive Kernel is used for creating the rules. After classifying the dataset, fuzzification is applied to that datset and then the rules have been created by Genetic Network Programming which based on direct graph structure. In the testing phase the system has been used to detect the misuse activities. By combining SVM with Genetic Network Programming increases the performance of the detection rate [14].In Year 2002, Jerzy Bala performed a work, "Application of a Distributed Data Mining Approach to Network Intrusion Detection". In this approach, classification rules are learned via tree induction from distributed data to be used as intrusion profiles. Agents, in a collaborative fashion, generate partial trees and communicate the temporary results among them in the form of indices to the data records. The process is terminated when a final tree is induced. This communication mechanism does not involve any data transfers, and in addition, a compression approach is used to reduce the communication bandwidth of data index transfers [15].

3. PROPOSED MODEL

In this present work, a communication criticality cell is generated based on communication feature analysis. This communication feature set is been used as the main vector for probabilistic analysis for Bayesian network approach. The presented work is about to perform the intrusion detection by using the proposed improved hybrid algorithm. The work will be implemented on a dataset to identify the intrusion so that the overall security of the system will be improved. The work is about to achieve the high detection ratio.

In this presented work we are going to present the work as an intrusion detection system which we can be implemented in the real environment of peer to peer network or any other centralized system. We need the statistics of such network with n number of nodes. We need to collect the information regarding the node definition and the expected parameters that can help to analyze the network traffic. This all information can be collected by studying the properties of P2P network. The information can be collected either by some external source in the form of secondary data or it can be build our self as the primary data source. In this present work, we have used KDD dataset to present the work effectively. The dataset is defined with following features: The separate the datasets are already defined in the form of training and the testing dataset. There are no duplicate records in the dataset. Dataset is defined under different classes of the attacks. Number of records in training and testing datasets is reasonable. Proposed system start working by selecting a training dataset, we have provided a browse button for selecting the training data set and testing data set than intrusion detection system perform its task on a dataset. Dataset represent the data as rows of TCP\IP dumps where each row consist of computer connection. Packet information in TCP dump file is summarized into connections. A connection is a sequence of TCP packet starting and ending at some defined time and data flows between source IP address and target IP address under well define protocol [Kayacik, G et al (2005)]. or normal connection).Each computer connect ion has 41 features and these features are grouped in to four categories Before feeding the data to the Bayesian network, for either learning or testing, raw network traffic has to be preprocessed and summarized into connections or high-level events. Each connection is described with a set of features. We use 9 of the 41 features. We used the labeled training dataset to train our Bayesian model, and the testing dataset to test for the correct discovery of intrusions. The presented work model is given below.

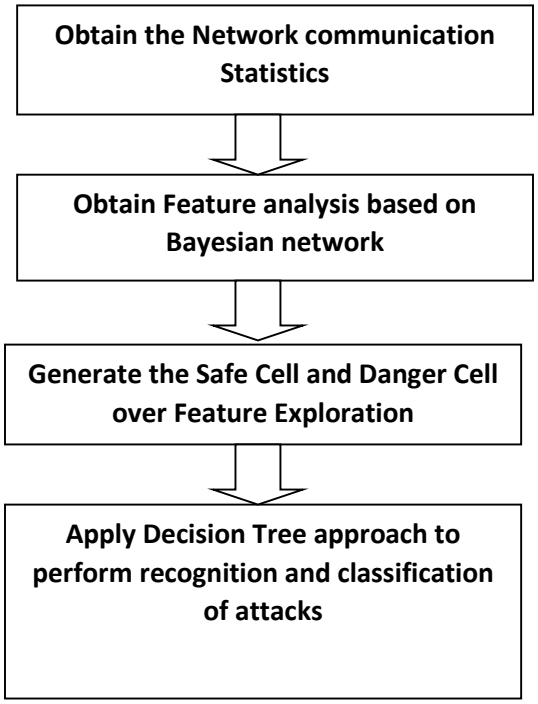

Figure 2 Flow of Work

These nine features are given as: Protocol type: type of the protocol, e.g. tcp, udp, etc. Service: network service on the destination, e.g., http, telnet. Land: 1 if connection is from/to the same host/port, 0 otherwise. Num, And failed logins: number of failed login attempts. Logged in, 1 if successfully logged in, 0 otherwise. Root shell: 1 if root shell is obtained, 0 otherwise. Is guest login: 1 if the login is a "guest" login, 0 otherwise.

Each computer connection has 41 features and these features are grouped in to four categories.

3.1. Basic Features

Basic features can be derived from packet header without inspecting payload. This category encapsulates all the attributes that can be extracted from a TCP/IP connection. Most of these features leading to an implicit delay in detection.

3.2. Content Features

Domain knowledge is used to assess the payload of original TCP packet .This include features such as number of failed login attempts. Unlike most of the DOS and Probing attacks, the R2L and U2R attacks don't have any Intrusion frequent sequential patterns. This is because the DOS and Probing attacks involve many connections to some host(s) in a very short period of time; however the R2L and U2R attacks are embedded in the data portions of the packets, and normally involves only a single connection. To detect these kinds of attacks, we need some features to be able to look for suspicious behavior in the data portion, e.g., number of failed login attempts. These features are called content features.

3.3. Time-Based Traffic Features

These features are designed to capture the properties that mature over a 2 second temporal window. One example of such feature would be number of connection to same host over 2 second interval.

However, there are several slow probing attacks that scan the hosts (or ports) using a much larger time interval than 2 seconds, for example, one in every minute. As a result, these attacks do not produce intrusion patterns with a time window of 2 Seconds. To solve this problem, the "same host" and "same service" features are re-calculated but based on the connection window of 100 connections rather than a time window of 2 seconds. These features are called connection-based traffic features.

4. RESULTS

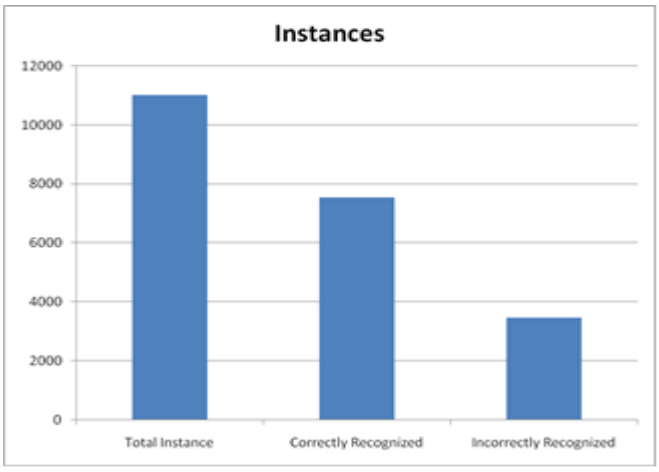

Figure 3 Instance Based Recognition (Normal Instance)

The figure here shows the instance based recognition obtained from the work for normal instances. The figure shows the total number of instances, correctly. The instance adaptive recognition is shown in the figure.

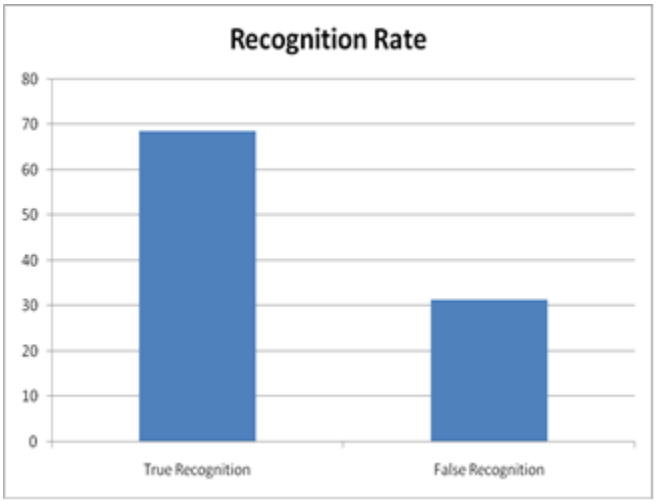

Figure 4 Recognition Rate Analyses (Normal Instances)

The figures show the instance based recognition obtained from the work. The figure shows the recognition rate obtained for correctly and false recognition aspects. The figure shows that the work has provided the recognition rate about 80%.

The figure here shows the instance based recognition obtained from the work for normal instances. The figure shows the total number of instances, correctly. The instance adaptive recognition is shown in the figure.

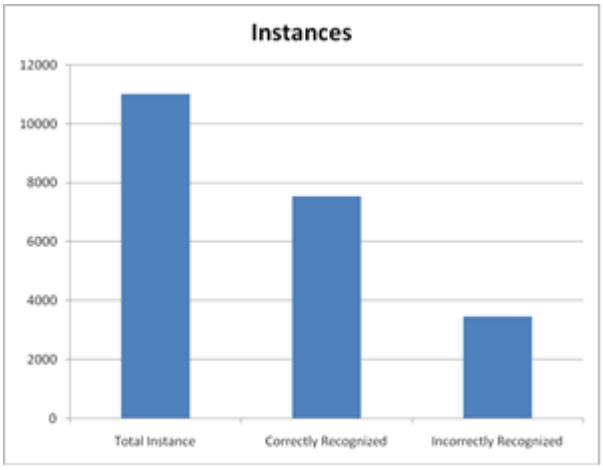

Figure 5 Instance Based Recognition (Attack Instance)

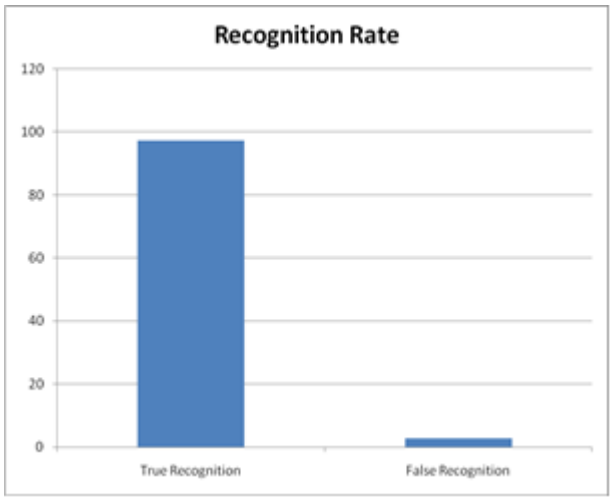

Figure 6 Recognition Rate Analyses (Attack Instances)

The figures show the recognition rate obtained for correctly and false recognition aspects. The figures show that the work has provided the recognition rate about 97%.

5. CONCLUSION

The work is here defined as a layered model. According to this presented model, at first the attribute level feature analysis is performed under weighted probabilistic model. Once the weighted value is obtained, the cell formation is done. This cell is defined under the threshold specification. In the final stage, the cell values are processed under decision tree to perform the attack prediction. The work has provided the effective recognition rate for normal and attack instances.

REFERENCES

[1] T.Subbulakshmi," Detection of DDoS Attacks using Enhanced Support Vector Machines with Real Time Generated Dataset", IEEE-ICoAC 2011 978-1-4673-0671-3/11©2011 IEEE

[2] Vera Marinova-Boncheva," Applying a Data Mining Method for Intrusion Detection", International Conference on Computer Systems and Technologies - CompSysTech'07

[3] Neelam Sharma," Layered Approach for Intrusion Detection Using Naive Bayes Classifier", ICACCI'12, August 3-5, 2012, Chennai, T Nadu, India. ACM 978-1-4503-1196-0/12/08

[4] C.I. Ezeife," NeuDetect: A Neural Network Data Mining Wireless Network Intrusion Detection System", IDEAS10 2010, August 16-18, Montreal, QC [Canada]; Editor: Bipin C. DESAI; ACM 978-1-60558-900-8/10/08

[5] Wenke Lee," Mining in a Data-flow Environment: Experience in Network Intrusion Detection", KDD-99 San Diego CA USA 1999 l-581 13-143 7/99/08

[6] LTC Bruce D. Caulkins," A Dynamic Data Mining Technique for Intrusion Detection Systems".

[7] Varun Chandola," A Reference Based Analysis Framework for Analyzing System Call Traces", CSIIRW '10, April 21-23, Oak Ridge, Tennessee, USA ACM 978-1-4503-0017-9

[8] Tsong Song Hwang, "A Three-tier IDS via Data Mining Approach", MineNet'07, June 12, 2007, San Diego, California, USA. ACM 918-1-59593-792-6/07/0006

[9] Carrie Gates, "Challenging the Anomaly Detection Paradigm A provocative discussion", NSPW 2006, September 19-22, 2006, Schloss Dagstuhl, Germany. ACM 978-1-59593-857-2/07/0007

[10] Jungsuk Song," Statistical Analysis of Honeypot Data and Building of Kyoto 2006+ Dataset for NIDS Evaluation", BADGERS '11 April 10-13, 2011, Salzburg.

[11] A.S. Aneetha, "Hybrid Network Intrusion Detection System Using Expert Rule Based Approach", CCSEIT-12, October 26-28, 2012, Coimbatore [Tamil nadu, India] ACM 978-1-4503-1310-0/12/10

[12] Huu Hoa Nguyen, "An Efficient Local Region and Clustering-Based Ensemble System for Intrusion Detection", IDEAS11 2011, September 21-23, Lisbon [Portugal] Editors: Bernardino, Cruz, Desai ACM 978-1-4503-0627-0/11/09

[13] Stefano Zanero," Unsupervised learning techniques for an intrusion detection system", SAC'04 March 1417 2004, Nicosia, Cyprus ACM 1-58113-812-1/03/04

[14] Mrutyunjaya Panda, "Ensemble Of Classifiers For Detecting Network Intrusion", ICAC3'09, January 23–24, 2009, Mumbai, Maharashtra, India., ISBN-978-1-60558-351-8.ACM 978-1-60558-351-8

[15] Jerzy Bala, "Application of a Distributed Data Mining Approach to Network Intrusion Detection", AAMAS'02, July 15-19, 2002, Bologna, Italy. ACM 1-58113-480-0/02/0007