Classification of Anomaly Data Using Data Mining Approaches

Srikanth Yadav.M [1], K.Gayathri[2], S.Monisha[3], V. Baby Supriya[4]

[1]*Associate Professor, Dept. of. CSE, Tirumala Engineering College, Jonnalagadda, NRT, AP, India*
[2]*Assistant Professor, Dept. of. CSE, Tirumala Engineering College, Jonnalagadda, NRT, AP, India*
[3, 4]*U.G. Students, Dept. of. CSE, Tirumala Engineering College, Jonnalagadda, NRT, AP, India*

***Abstract*—** The rapid evolution of technology and the increased connectivity among its components imposes new cyber-security challenges. To tackle this growing trend in computer attacks and respond threats, industry professionals and academics are joining forces in order to build Intrusion Detection Systems (IDS) that combine high accuracy with low complexity and time efficiency. The present article gives an overview of existing Intrusion Detection Systems (IDS) along with their main principles. Also, this article argues whether data mining and its core feature which is knowledge discovery can help in creating Data mining based IDSs that can achieve higher accuracy to novel types of intrusion and demonstrate more robust behavior compared to traditional

***Keywords*—** *Intrusion Detection, Anomaly, Misuse, Classification, Machine Learning*

I. INTRODUCTION

In day to day life, the need for speed access to information through the internet has increased. Hence the room for maintaining security in any organization either public or private system has become fundamental. Because of increase in network connections and systems, unauthorized access and interruption of the data are triggered. As a result, it is indispensable to create a virtual access path. In general, intruders have the capacity to find out a defect in systems or networks and can spawn vulnerabilities. Even though the access control points exist in the network, they fail in providing scrupulous security to the systems. To identify intruders, developing Intrusion Detection Systems (IDSs) is the best solution to protect systems and networks. Therefore the task of IDS is not only to detect intruders but also to monitor the raid of intruders. An accurate system of protecting data and resources from illicit access, damaging and denial of use is to be built. For every system, the security perspective is to be planned based on the expected performance. Mainly security is concerned with the following aspects in a computer system.

- Confidentiality – information is to be accessed only by permissible persons.
- Integrity – information must remain unaffected by destructive or malicious attempts.
- Availability – the computer is responsible to function without downgrading of access and provide resources to legal users when they require it.

Specifically, an intrusion is defined as a set of events which are unknown and unforeseen to the user, which compromises the protection of a computer system. It can be done from the external side or internal side of the system. Earlier in 1980‟s James P Anderson has defined intrusion as the scope of illegal force to access information, defraud information, or making the computer system unsafe. Intrusion Detection System (IDS) was commercially promoted in the year 1990. From then a variety of layouts were introduced to adapt intrusion detection systems [1] [2]. It acts like a burglar alarm and detects any kind of violation and generates alarms like audible, visual and also messages like e-mail. On the whole, IDS is primarily exploited for stopping defective activities that may attack or misuse the system by identifying attacks through providing desirable support for defense management and also give constructive information regarding intrusion. But the structure of IDS should possess low fake alarms while undertaking the discovery of attacks. IDSs have become shielding mechanisms everywhere in current networks. There is no thorough and proficient methodology offered in checking the strength of these systems.

IDSs are mainly classified into following three categories based on the framework. They are namely

- Network-based IDS
- Host-based IDS
- Application-based IDS

A. Network-based Intrusion Detection System

A Network-based Intrusion Detection System (NIDS) is a passive mechanism which prevails in the computer or the network of an organization and examines the network traffic for identification of attacks. If NIDS find out any attack, it reports such malicious codes to the system administrators immediately. A NIDS is set up in the margin of the router to monitor the traffic going into and out of the network. Probably a minimum number of supervising units for the large network can be disposed of without disturbing the usual operations of networks. NIDS is not vulnerable even with a direct attack. However, NIDS is unable to detect encrypted data packets and not succeed in distinguishing some attacks.

The NIDS is not on a dangerous track for any production services or processes since a network-based IDS does not operate as a router or other critical mechanism. System failure does not have a considerable impact on this type of IDS. Network-based IDS systems are more inclined to be self-reliant than host-based systems. They usually run on a committed scheme that is modest to install; just unload the device, perform some counteractive configuration, and then plug it into your network in a position that authorizes it to manage responsive traffic. It works on the basis of signature matching, comparing attack patterns with distinguished signatures in their database. Examples of NIDS are Snort, Netprowler, and Cisco NIDS. Mostly, Network Intrusion Detection System is robust for medium to large scale organizations, because of their amount of data and resources. Accordingly, a lot of smaller companies are doubtful in installing IDS.

The main advantages of Network Intrusion Detection System are as follows:

- Large networks are monitored by installing a few devices with an excellent network design.
- Ongoing network operations are not suspended by setting up NIDS, as they are passive devices.
- These are not vulnerable to direct attacks and conceivably detected by attackers.

The disadvantages of NIDS include

- NIDS is unsuccessful in recognizing an attack when network ability becomes very large.
- Whereas many switches are restricted or no observing port capability is available, a few networks are not proficient in maintaining all the data for investigation by a Network based IDS.
- NIDS is unable to examine encrypted packets, producing some of the traffic to be not detectable by the process and decreasing the effectiveness of NIDS.
- Attacks involving fragmented or malformed packets are not simple to detect.

B. Host-based Intrusion Detection System

A Host-based Intrusion Detection System (HIDS) is the one that is located on the computer or server, called host, and inspects only the host activities. Therefore, HIDS is employed to examine the system files, stored configuration files and to detect either creation or modification and deletion of system files by the intruders. Mainly HIDS is responsible for detecting local transactions as well as attacks that are not detected by the NIDS. The configuration of HIDS pertains only on the individual host and in need of more management effort in installing and configuring in multiple hosts. Moreover, Host-based intrusion detection system is vulnerable to direct attacks and are inclined to some Denial of Service (DoS) attacks. Host-based systems possess lower false positive rates than network-based Intrusion Detection Systems. It is because of several types of commands executed on a definite host are more focused than the types of traffic flowing across a network. This is the characteristic which decreases the problems of host-based analysis engines.

Common examples of HIDS are Cisco HIDS, Tripwire, Symantec ESM. These work on the regulation of configuration and adjusting capability. An alarm is set on when a file attributes are altered, fresh files created or prevailing files deleted. Normally the majority of Host-based IDS have same architectures because most of the host systems function as host agents informing to the main console.

C. Application Based Intrusion Detection System

Application-based Intrusion Detection System (AppIDS) is the development of the HIDS which inspects every application for anomalous events by looking into the files created in an application and also abnormality occasions like void file execution, exceeding the users‟ permission, etc. So, an AppIDS studies the communication between the application and the user and able to observe the encrypted traffic also. Nevertheless, AppIDS is more prone to attack and does not have the skill to notice the software tampering.

Mostly to build expert IDS, we must guarantee in minimizing the false alarms (both positive and negative). Based on the detection method, IDSs are categorized into

- Misuse detection
- Anomaly detection

D. Misuse Detection method

Misuse detection or signature-based IDS is the one that matches signatures or patterns of the known attacks occurring in the incoming traffic of the network. So every time, the signature is used to detect attacks accurately. The major concern in misuse detection system is writing signatures that completely includes possible transformations of the relevant attack. As well as writing signatures that do not match non-intrusive events. The gratification of misuse detection method is that it holds very good accuracy in discovering notorious attacks.

E. Anomaly Detection

The anomaly detection or statistical anomaly based Intrusion Detection System makes use of statistical analysis by keeping track of the traffic which is identified to be normal and a potential baseline is evolved. The foremost functionality of anomaly detection is the capability of discovering unknown attacks [5]. The network events are

regularly observed and matched with the baseline to determine intrusions. Generally, the statistical and behavioral models which are used to detect attacks allow a low false negative rate. Moreover, behavioral patterns of users or programs are developed based on a pattern of normal and abnormal activities, which are used in detecting the existence of an attack. Accordingly, any divergence from normal activity by a user or program would be detected, thereby produces an alarm. Unfortunately, most alarms are favorable and false positives are derived as an outcome. It is also considered as behavior-based IDS. The basic principle of anomaly IDS is concerned with intrusive activity, defined to be as a section of abnormal activity. The intrusion may be identified based on anomalous actions.

II. LITERATURE REVIEW

The need for computing machines and utility of information has been growing tremendously in the last few decades. As a result, it provides a greater access to the unknowns and makes it easier for the intruder to cover their tracks. Due to the increase in a number of attacks on major sites and networks, U.S. Defence Department Intrusion Detection (ID) was developed. Intrusion Detection is a type of security management system for computers and networks. An IDS gathers, analyses information and identify unauthorized users from various areas within a computer or a network, which include both intrusions from outside the organization and from within the organization. ID uses vulnerability assessment which is a technology developed to assess the security of a computer system or network. Intrusion detection has a bit more history behind it. The intrusion detection was introduced as a formal research when Anderson [2] wrote a technical report for the U.S. Air Force.

In 1980, the concept of intrusion detection began with Anderson's [2] seminal paper where he introduced a threat classification model that develops a security monitoring surveillance system based on detecting anomalies in user behavior. Abdullah, B et al.[1] has presented one collaborate IDS module to make a real-time detection, and block intrusions before occurrences based on HIDS using sequences of system call anomaly detection. Frivold, T [3] Proposes a data mining technique to discover fuzzy classification rules based on the Apriori algorithm. In his technique, genetic algorithms are incorporated to determine minimum support and confidence with binary chromosomes.

Baghdad, et al. [4] studied the development of host-based anomaly intrusion detection, focusing on system call based HMM training. This was later enhanced with the inclusion of data pre-processing for recognizing and eliminating redundant sub-sequences of system calls, resulting in less number of HMM submodels. Experimental results on three public databases indicated that training cost can be reduced by 50% without affecting the intrusion detection performance. False alarm rate is higher yet reasonable compared to the batch training method with a 58% data reduction. Devikrishna, K. S. and Ramakrishna, B. B.. [5] has proposed a method of detecting intrusion using incremental SVM based on key feature selection. A center SVM summarizes the distributed samples and incorporates them to build the incremental SVM for locals. By eliminating the redundant features of sample dataset the space dimension of the sample data is reduced. Again in the same year, Denning, D. E. [6], use the Rough Set Theory (RST) and Support Vector Machine (SVM) to detect intrusions. First, RST is used to pre-process the data and reduce the dimensions. Next, the features are selected by RST which will be sent to SVM model to learn and test respectively. The method is effective to decrease the space density of data.

Dokas, P et al. [7] effectively introduced an Intrusion Detection System by using Principal Component Analysis (PCA) with Support Vector Machines (SVMs) as an approach to select the optimum feature subset. Combining multiple independent data sources the known intrusion and anomaly intrusion identification is made by using statistical wavelet-based detection mechanism is carried out by Duan, Z et al. [8].The properties such as attack duration, packet count, packet rate, and dominant protocol type match with the two data sets, are indicated by attack structure. At lean and heavy traffic scenarios, the demand capacity of the server was observed to give a better clarity of anomaly intrusion detection. Analysis of several traffic anomaly properties which is impossible using traditional intrusion measurements can be performed by a new model using anomaly intrusion attack measurements. Small businesses seem to be the most common targets of attacks. Traditional measures in understanding and detecting of anomaly intrusion are no more reliable to give the current trends of attacking using spoofed address sources. Endorf et al. [9] investigate the use of a one-class Support Vector Machine algorithm to detect the onset of system anomalies, and trend output classification probabilities, as a way to monitor the health of a system.

III. PROPOSED SYSTEM

Our main aim is to develop an IDS based on anomaly detection model that would be precise, not easily cheated by small variations in patterns, low in false alarms, adaptive and be of real time. The Figure 1 describes the proposed system architecture were the intrusion packets are received from the internet then SNORT is used to collect the datasets. Initially, the features extracted from data packets then it forwarded to our proposed IDS. Then, proposed IDS compute the distance between the extracted features and trained model. Here, the trained model consists of big datasets with distributed storage environment to improve the performance of Intrusion Detection system. Thus, the

outlier value is greater than the specified threshold then it generates the false alarm.

The benefits of the proposed approach are Clusters are formally defined as maximal sets of density-connected objects. Here a simple two-dimensional dataset is taken with a much larger number of examples in cluster C1 than C2. So the cluster density of C2 is extensively higher than that of C1 cluster density. For each example consider an object q inside the cluster C1, the distance between the example q and its nearest neighbor is greater than the distance between the example p2 and the nearest neighbor from the cluster C2, and the example p2 will not be considered as an outlier.

A. MODULE DESCRIPTION:

Classification is one of the best – known solution approaches. National Institute of Standards and Technology (NIST) organization provides guidance document on Intrusion Detection Systems. Intrusion Detection System briefly classified into three different categories:

a) Host-based IDS
b) Network-based IDS and
c) Vulnerability-assessment IDS

There are two basic models used to analyze the events and discover attacks: Misuse detection model – Intrusion Detection System detect intrusions by looking for similar activities such as vulnerabilities or known intrusion signatures. Anomaly detection model - IDS detect intrusions by searching « abnormal » network traffic.

The misuse detection model is commonly referred as IDS commercial tool; always Vendors must update intrusion signatures. Anomaly detection based IDS model has the capability to detect attack symptoms without specifying attack models, but these models are very sensitive to false alarms. In the present study, we have utilized the proposed IDS approach's based on the anomaly detection model. The Proposed architecture is shown in Figure 1.

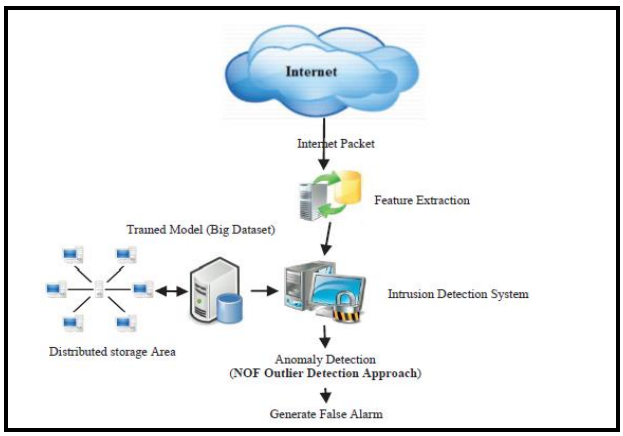

Fig 1: Proposed Architecture

IV. IMPLEMENTATION

The Network packets are collected and features extraction is done to identify the anomaly packets from the network packets. We have designed Intrusion Detection System using Neural Networks in this code. Java was previously used in Phase1 to prepare datasets and weka were used in Phase2 to carry out Data Cleaning, and R was used for building a neural network. We have used the 'neural net' package available in R for implementing the neural network. Our implementation consists of a multi-layered network that consists of numerous neurons, which are arranged into levels that are interconnected. The neural net package used an Input layer, the output layer which will provide the classification and between them, there is a number of hidden layers. The results obtained show great accuracies and the implementation took a minimal time for completion. Resource consumption was noticeably low too. Data files prepared in Phase1 of the project were used. Dataset was divided into training and testing sets so that we could train the Neural Network and carry out Intrusion Detection. The specifications of the proposed system are as follows.

- Implementing and learning how a neural network would perform when detecting an intrusion in the system.
- Testing and learning how a neural network would perform in differentiating the different kinds of attacks from normal behavior.
- Testing and learning the ability of the neural network to distinguish between different types of attacks.
- We design and implement a misuse detection system capable to detect a particular attack where we train our Neural Network on our training set with two output states: Attack of a specified type/ Other Case. After training, we test our Neural Network on the testing set.
- We design and implement a misuse detection system that can identify between few different attack types where we train our Neural Network on our training set with outcomes identifying each of the five attacks and then test our Neural Network on the testing set
- We design an anomaly detection system where we train our Neural Network to detect a normal case against any other cases and after training our Neural Network, we test our Neural Network on a testing set.
- We also record time is taken and memory consumed in both training and testing for all above approaches

V. EXPERIMENT RESULTS

For carrying out intrusion detection for Anomaly-based attacks and Misuse based attacks we had two data sets Dataset_Anomaly.csv and Dataset_Misuse.csv in the preparation phase. In the anomaly detection data set, the class or prediction variable is either Normal which represents a normal case or an Attack. Contrary to the anomaly detection data set, the misuse detection data set has a class variable Normal or Name of the attack which

represents a specific type of attack such as Smurf, NMap, Rootkit, etc. We carry out data cleaning on Dataset_Anomaly.csv using Weka's to obtain Dataset_Anomaly_AttributeSelection.csv and which has fewer attributes that help speed our NN.

The 'neural net' package is available in R and is open source. It was used for our neural network based IDS and Analysis. The package provides functions to both generate the neural network and carry out classification. Our attribute values were used to create a formula that is supplied to 'neural net' function. The neural net function returned an object that has all relevant information about the neural network and can be further used to derive our results. For classification, we took into consideration 4500+ instances of normal cases and attack cases. We chose 10 types of attacks including Neptune, NMap, PortSweep, Satan, Smurf, BufferOverflow, FTPWrite, GuessPassword, Back and Rootkit attacks. The Accuracy obtained during the execution is shown in Figure 2.

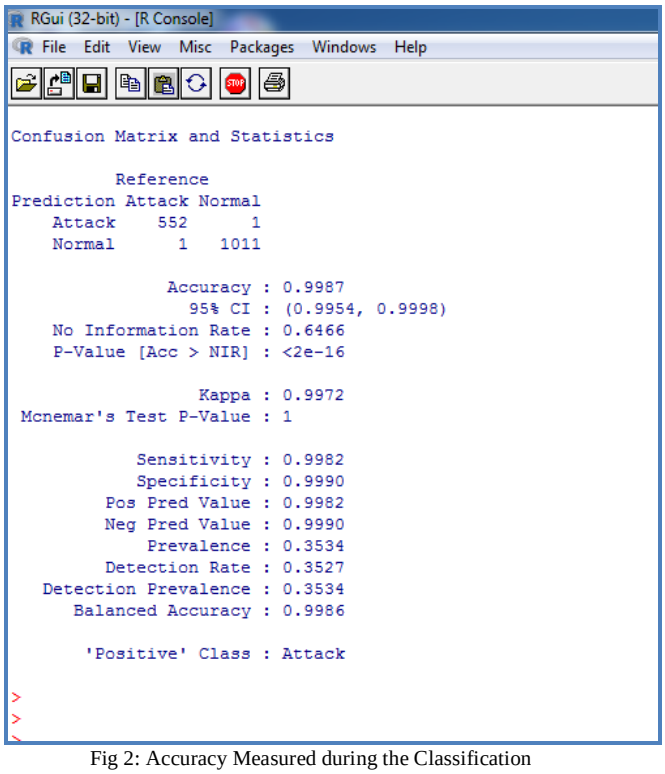

Fig 2: Accuracy Measured during the Classification

VI. CONCLUSION

The proposed IDS perform well by detecting attacks. This approach helps to overcome the human interaction toward preprocessing. Our experimental results proved that the proposed algorithms solve the above-mentioned issues and detects the attacks in an effective manner compared with other existing works. Thus, it will pave the way for an effective means of intrusion detection with better accuracy and reduced false alarm rates.

VII. REFERENCES

[1]. Abdullah, B., Abd-algafar I., Salama G. I. and Abd-alhafez A. Performance Evaluation of a Genetic Algorithm Based Approach to Network Intrusion Detection System, Proceedings of 13th International Conference on Aerospace Sciences and Aviation Technology (ASAT-13), Military Technical College, Cairo, Egypt, 2009;1-5.

[2]. Anderson, J. P. Computer security threat monitoring, and surveillance. Technical Report, Fort Washington, PA, USA.,1980;911.

[3]. Anderson, D., Frivold, T. and Valdes, A. Next-generation intrusion detection expert system (NIDES): A summary Technical Report SRI–CSL–95–07, Computer Science Laboratory,SRI International, May 1995.

[4]. Beghdad, R. Critical study of neural networks in detecting intrusions. Computers and Security, 27(5-6): 2008;168–175.

[5]. Devikrishna, K. S., and Ramakrishna , B. B. .An Artificial Neural Network based Intrusion Detection System and Classification of Attacks", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Jul-Aug 2013, 3(4): 1959-1964.

[6]. Denning, D. E.. An intrusion detection model, IEEE Transactions on Software Engineering, CA,. IEEE Computer Society Press;1987.

[7]. Dokas, P., Ertoz, L., Lazarevic, A., Srivastava, J.and Tan, P. N. Data mining for network intrusion detection. Proceeding of NGDM, 2002;21–30.

[8]. Duan, Z., Chen, P., Sanchez, F., Dong, Y., Stephenson, M. and J. M. Barker, M. (2012). Detecting spam zombies by monitoring outgoing messages, IEEE Trans. Dependable and Secure Computing, Apr 2012; 9(2):198–210

[9]. Endorf, C., Schultz, E. and Mellander, J. (2004). Intrusion detection and prevention. California: Mc Graw-Hill.

[10]. Forrest, S., Hofmeyr, S. A. , Somayaji, A. and Longstaff, T. A. A Sense of Self for Unix Processes, IEEE Symposium on Research in Security and Privacy, Oakland, CA, USA, 1996;120--128.

[11]. Gaikwad, Sonali Jagtap, D.P. Kunal Thakare and Vaishali Budhawant. Anomaly Based Intrusion Detection System Using Artificial Neural Network and fuzzy clustering., International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, November- 2012; 1(9).

[12]. Goyal, A. and Kumar, C. .GA-NIDS: A Genetic Algorithm based Network Intrusion Detection System, Electrical Engineering and Computer Science, North West University, Technical Report;2008.

[13]. Gu, G., Porras, P., Yegneswaran V., Fong, M. and Lee, W. BotHunter: detecting malware infection through IDS-driven ialog correlation, Proc. of 16th USENIX Security Symp. (SS '07), Aug. 2007; 12:1–12:16.

[14]. Gu, G., Zhang, J. and Lee, W. (2008). BotSniffer: detecting botnet command and control channels in network traffic, Proc. of 15th Ann. Network and Distributed Sytem Security Symp. (NDSS '08), Feb. 2008.

[15]. Jaiganesh, V., Sumathi, P. and Mangayarkarasi, S. ,An Analysis of Intrusion Detection System using back propagation neural network, IEEE Computer Society Publication;2013.