A New Data-Mining Based Approach for Host Based Intrusion Detection System

Suyash Rithe, Omkar Pandit, Ranjeet Pasale, Sanjay Tambe,
Dept. of Computer Engg. Pimpri Chinchwad College of Engg.
Pune, India jan-2013

***Abstract* - Nowadays, as information systems are more open to the Internet, the importance of secure networks is tremendously increased. New intelligent Intrusion Detection Systems (IDSs) which are based on sophisticated algorithms. In this paper, we propose a new data-mining based algorithm such as K-means for clustering and ANN (Artificial Neural Network) algorithm for predicting the suspicious packet is whether intrusion or normal packet. We design a system which includes analysis of packets and detection of malicious packets by applying data mining algorithms. The main objective of system is to monitor the packets coming from network on host system.**

keywords:-Data mining, k-means, ANN

INTRODUCTION

Security is an important issue for all the networks of companies and institutions at the present time and all the intrusions are trying in ways that successful access to the data network of these companies and Web services and despite the development of multiple ways to ensure that the infiltration of intrusion to the infrastructure of the network via the Internet, through the use of firewalls, encryption, etc. But IDS is a relatively new technology of the techniques for intrusion detection methods that have emerged in recent years. Intrusion detection system's main role in a network is to help computer systems to prepare and deal with the network attacks.
Intrusion detection functions include:

- Monitoring and analyzing both user and system activities.
- Analyzing system configurations and vulnerability
- Assessing system and file integrity.
- Ability to recognize patterns typical of attacks.
- Analysis of abnormal activity patterns.
- Tracking user policy violations.

The purpose of IDS is to help computer systems on how to deal with attacks, and that IDS is collecting information from several different sources within the computer systems and networks and compares this information with preexisting patterns of discrimination as to whether there are attacks or weaknesses.

CLASSIFICATION OF INTRUSION DETECTION SYSTEM

Intrusion detection system are classified into three types
1. Host based IDS
2. Network based IDS
3. Hybrid based IDS

1. Host based IDS (HIDS)

This type is placed on one device such as server or workstation, where the data is analyzed locally to the machine and are collecting this data from different sources. HIDS can use both anomaly and misuse detection system.

2. Network based IDS (NIDS)

NIDS are deployed on strategic point in network infrastructure. The NIDS can capture and analyze data to detect known attacks by comparing patterns or signatures of the database or detection of illegal activities by scanning traffic for anomalous activity. NIDS are also referred as "packet-sniffers", because it captures the packets passing through the of communication mediums.

3. Hybrid based IDS

The management and alerting from both network and host based intrusion detection devices, and provide the logical complement to NID and HID - central intrusion detection management.

In this paper, we will present a new data-mining based technique for intrusion detection using an K-means algorithm for clustering of packets. We will create an initial dataset using packet analyzer tool which will contain some packets. Using K-means we will cluster the packets into three group such as attack, non-attack and suspicious. If the packet is suspicious it will predict by ANN algorithm. We are also using BPNN (Back Propagation Neural Network) and feed-forward for final prediction. The proposed paper organized as, Section 2 explains about data mining. Section 3 introduces K-means algorithm for clustering and BPNN in training phase. ANN(Artificial Neural Network) and feed forward for final prediction in Section 4 and conclusion in Section 5.

SECTION 1

Proposed System:-

Intrusion detection system is used for attempting to detect computer attacks by examining data records observed the processes on the same network. Proposed intrusion detection system is a Host based system which works online. Here we used anomaly detection analysis engine that can search for something unusual or rare, analyze system event stream to find patterns of activities appearing to be abnormal. It is rule based system based on sets of predefined rules that are provided by network administrator.

Proposed intrusion detection system is comprising of three algorithms k-means, Back propagation Neural Network and Feed Forward. Here, we used packet sniffer to capture the packets coming from the network on host system. By using packet analyzer we found signatures of every approaching packet. Signatures are nothing but properties of packets like protocol of packet, packet length etc. After analyzing packets were manually labeled, they get marked whether attacked or not-attacked. Then packet data set get transformed, each packet is vectorized by assigning a unique integer value. Transformed set was given as input to K-means for clustering. As length is the most varying signature among all the packets .we used packet length as base field for clustering using K-means. After clustering, clustered data set was stored. Every packet from Training data set is forwarded to Back propagation Neural Network as input. Basic purpose of BPNN is to classify the scattered the data based on rules defined by administrator. So here, BPNN is mainly builds a model of packets categorized as attack or non-attack. BPNN model saved once, so to use as reference for making prediction about new incoming packet. It is all about work we had performed in training phase. Now in actual implementation phase, when in real time a packet arrived it get analyzed, K-means extracts its features and at last Feed forward make a prediction about packet i.e. it is attacked or clean by referencing model created by Back Propagation Neural Network. Here, along with prediction we are also provide functionality that describe more about attack (name of attack). We also provide an user friendly feature of providing a SMS and Email alert through which user get informed about IP address of machine from which malicious data is approaching. Because of this feature user get aware about source of attack so from next time he/she can avoid acceptance of data from that resource.

We provide serialization to store the packet information which is in the form of byte code which is secure for the user, to run this application the user must have packages like jpcap and winpcap install on his system.winpcap is used to scan all ports of machine and capture the packets and jpcap is used in java application which works on these packets.

We can say that the IDS we are developing is a software application which can be install on any machine and provide efficient result.

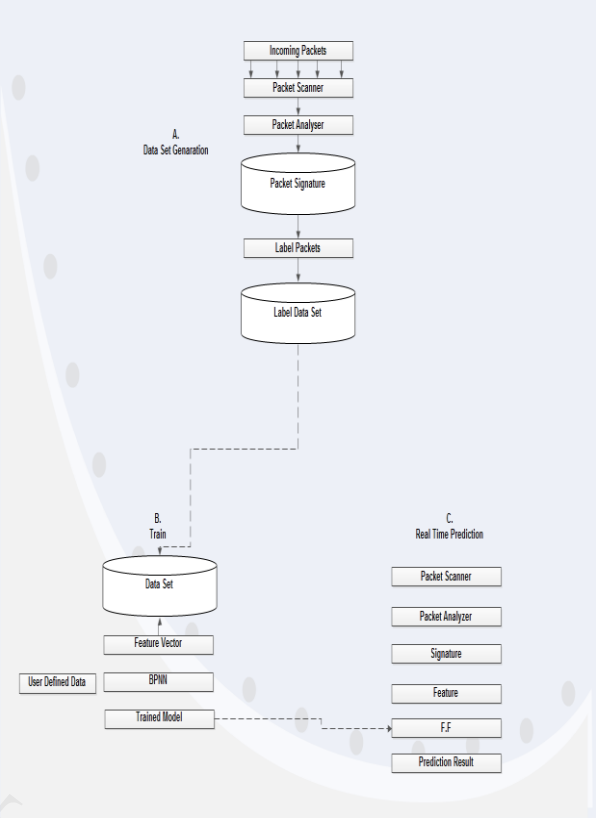


FIG: PROPOSED SYSTEM

SECTION 2

Data mining (DM), also called Knowledge-Discovery and Data Mining, is one of the hot topic in the field of knowledge extraction from database. Data mining is used to automatically learn patterns from large quantities of data. Mining can efficiently discover useful and interesting rules from large collection of data. It is a fairly recent topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition.

Data mining is disciplines works to finds the major relations between collections of data and enables to discover a new and anomalies behavior. Data mining based intrusion detection techniques generally fall into one of two categories; misuse detection and anomaly detection. In misuse detection, each instance in a data set is labeled as 'normal' or 'intrusion' and a learning algorithm is trained over the labeled data. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks, as long as they have been labeled appropriately.

Host-based intrusion detection is the first area that was explored in intrusion detection, and it involves loading a piece or pieces of software on each server that needs to be monitored. The loaded software uses log files, process accounting information, user behaviors or output data from application-based intrusion detection systems operating on the host. Network-based intrusion detection monitors the raw network packets on its network segment. It does not require software to be loaded and managed on a variety of hosts [2]. Since host-based intrusion detection involves looking at the

logs containing events that have actually occurred, it can measure whether an attack is successful or not with more detail and accuracy than a network-based intrusion detection system. It can monitor all user log activities as well as what each user does during the time that he connects to the network. Host-based intrusion detection may also check the integrity of the system files and watch for suspicious processes, such as file accesses and attempts to install executables or access-privilege service. Host-based intrusion detection systems reside on each host to monitor the activities that occur at the user level, much of which cannot be detected by a network-based system. For example, attacks from the keyboard or login within a LAN do not cross the network and, therefore, cannot be detected by network-based intrusion detection. In switched environments, it is difficult to achieve sufficient network coverage by network-based intrusion detection systems. Host-based intrusion detection provides better performance due to residing on as many critical hosts as need. Furthermore, encrypted network traffic can be hardly handled by a network based system. When the incoming traffic comes to an operating system, it has been deencrypted. A host-based system is able to monitor such events easily. Finally, a host based intrusion detection system does not need any additional hardware and is cost effective. These are the reasons why host-based intrusion detection system has been deployed first [2, 3].

SECTION 3

K-means algorithm for clustering of packets

K-means algorithm is a classical clustering algorithm. Its aim is to divide data into k clusters, and ensures that the data within same cluster has high similarity; the data in different cluster has low similarity. K-means algorithm first select K data at random as initial cluster center, for the rest data, add it to the cluster with highest similarity according to its distance to cluster center; then recalculate the cluster center of each cluster. Repeat this process until each cluster center doesn't change. Thus data is divided into K clusters. K-means algorithm is very simple, it is suitable for large scale data set. But K-means algorithm is sensitive to initial value and sequence of data object, different initial data may lead to different clustering result. The purpose of the proposed approach is to perform a clustering analysis on a set of tested connections through K means, and then compute the distribution of false alerts in these clusters. This operation is repeated several times, with a different number of clusters each time, until obtaining a final configuration of clusters where each cluster is ideally highly representative of false alerts (the percentage of false alerts is high), or it is highly representative of real attacks (the percentage of false alerts is low).[1]

Clustering: In this technique, data points are clustered together based on their similarity factors and is often nearness according to some defined distance. Clustering [4] is an effective way to find hidden patterns in data that humans might miss. It is useful for ID as it can cluster malicious and non malicious activity separately. k-means is a clustering algorithm used to cluster observations into different groups of related observations without having prior knowledge about their relationships. Here data is divided in k clusters.

where k is provided as input

Mathematical equation:-

For K-means:(clustering of packets)

$C_N=K_M (S_{DB}, N)$

Where, N- Number of clusters

S_{DB} - signature of database set

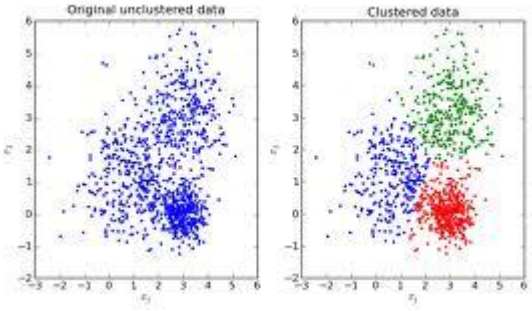
C_1, C_2, C_3 belongs to C_N -set of cluster centroid

Here we are using 3 clusters so N=3

(a) Attack

(b) non-attack

(c) suspicious



fig(a) clustering of packets

BPNN(Back propogation Neural Network)

A neural network has a natural propensity for storing experiential knowledge and making it available for use. Then the Input-Output Mapping property and capability can be provided by the ANNs [5]. One of the most commonly used supervised ANN model is BPNN.

For BPNN (training phase)

M= BPNN (S_{DB}, L_{DB})

Here

S_{DB} - signature of database set

L_{DB} -is our output - label set for signature based packet set.

SECTION 4

Artificial Neural Network and Feed-Forward (Prediction of Suspicous Packets)

An ANN is an information processing system that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a large number of highly interconnected processing elements (neurons) working with each other to solve specific problems. Each processing element (neuron) is basically a summing element followed by an activation function. The output of each neuron (after applying the weight parameter associated with the connection) is fed as the input to all of the neurons in the next layer. The learning process is essentially an optimization process in which the parameters of the best set of connection coefficients

(weighs) for solving a problem are found and includes the following basic steps [6]:
- Present the neural network with a number of inputs (vectors each representing a pattern)
- Check how closely the actual output generated for a specific input matches the desired output.
- Change the neural network parameters (weights) to better approximate the outputs. Some IDS designers exploit ANN as a pattern recognition technique. Pattern recognition can be implemented by using a feed-forward neural network that has been trained accordingly. During training, the neural network parameters are optimized to associate outputs (each output represents a class of computer network connections, like normal and attack) with corresponding input patterns (every input pattern is represented by a feature vector extracted from the characteristics of the network connection record). When the neural network is used, it identifies the input pattern and tries to output the corresponding class. When a connection record that has no output associated with it is given as an input, the neural network gives the output that corresponds to a taught input pattern that is least different from the given pattern [7].

For Feed Forward

$Pc - FF(M,S_C)$

Where

Pc- prediction about packet Pi

M - trained model by BPNN

S_C - signature of current model

SECTION 5

CONCLUSION

Intrusion detection systems (IDSs) play an important role in computer security. IDS users relying on the IDS to protect their computers and networks demand that an IDS provides reliable and continuous detection service. However, many of the today's anomaly detection methods generate high false positives and negatives.

Intrusion detection based upon computational intelligence is currently attracting considerable interest from the research community. Its characteristics, such as adaptation, fault tolerance, high computational speed and error resilience in the face of noisy information, fit the requirement of building a good intrusion detection system.

REFERENCES

[1] Clustering Based Network Intrusion Detection Using Kdd Train 20 Percent. Poonam Dabas, Rashmi Chaudhary Assistant Professor M tech Department Of Computer Sc &Engg.Uiet Kurukshetra Universty , India. 2004.

[2] Internet Security System, Network- vs. Host-based Intrusion Detection, A Guide to Intrusion Detection Technology, http:// secinf.net /info/ids/nvh_ids/, 1998

[3] Aaron Schwartzbard and Anup K. Ghosh, "A Study in the Feasibility of Performing Host-based Anomaly Detection on Windows NT," 2nd Internationa lWorkshop on Recent Advances in Intrusion Detection--RAID99.

[4] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[5] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 2003; 15: 1373-1396

[6] Sergios Theodorios and Konstantinos Koutroumbas, *Pattern Recognition*, Cambridge: Academic Press, 1999.

[7] K. Fox, R. Henning, J. Reed, and R. Simonian, "A neural network approach towards intrusion detection," Proceedings of 13th National Computer Security Conference, Baltimore, MD, pp. 125-134, 1990.