Intrusion Detection System Using Data-Mining

Phadke Reshma Bhaskar
Computer Engineering Department
JSPM-NTC, Narhe
Pune, India

Bhandwalkar Shital Narayan
Computer Engineering Department
JSPM-NTC, Narhe
Pune, India

Ambike Pratiksha Pandurang
Computer Engineering Department
JSPM-NTC, Narhe
Pune, India

Chavan Vaishali Pandurang
Computer Engineering Department
JSPM-NTC, Narhe
Pune, India

Prof. Kadam G. V.
Computer Engineering Department
JSPM-NTC, Narhe
Pune, India

***Abstract*—With the very fast growth of internet, security of network is become very big issue of computer network system. Network attacks are increasing very fast. Network traffic attack is known as intrusion. Intrusion detection is used for identifying the attack on network for security purpose of information. Data mining techniques are used for detect the attack and classify these attack. Data mining techniques are for clustering, training and detection of attack. This paper will present the intrusion detection using data mining.**

***Keywords*— Data mining, Intrusion detection, Clustering,**

Cyber security, Classification, Machine Learning

I. Introduction

Now a day's people want their all things secure. In previous days security is related to homes, but now security is related to technological things like computer network. One security system is intrusion detection. Intrusion detection is nothing but the identifying unauthorized access to system which will harm the system. For detecting intrusion attack on system various data mining techniques are used. The data mining techniques are like using anomaly detection and signature database, decision tree and support vector machine, multiple level tree classifier k-nearest neighbor (KNN) etc. We are using the data mining technique in our paper is k-means clustering algorithm and artificial neural network(ANN) for detecting the intrusion. Here we are using k-means algorithm for clustering the various types of data. Also for normalization k-means is used in our paper. Artificial neural network (ANN) algorithm is used for training the dataset and detecting the attack.

II. Literature Survey

Xiang M.Y. Chang et.al. (2004) designed a multiple level tree classifier for Intrusion detection system and increase the detection rate.

Peddabachigiri S.et.al. (2007), proposed a model of intrusion detection system combining decision tree and support vector machine (DTSVM) classification techniques and produces high detection rate.

Mrutyunjaya panda et. al. (2008) compares different data mining techniques for intrusion detection system and found that accuracy & performance of Naïve bayes classifier for all classes is better than the accuracy obtained in the case of different Decision tree algorithm .M.Govindarajan et.al. (2009), proposed new K-nearest neighbor classifier applied on Intrusion detection system and evaluate performance in term of Run time and Error rate on normal and malicious dataset.

Mohammadreza Ektela et.al. (2010) used Support Vector Machine and classification tree Data mining technique for intrusion detection in network.

Song Naiping et.al. (2010), studied on IDS based on Data mining. In this paper Misuse detection and Anomaly detection are described as data mining technique.

T.Velmurugan et.al. (2010), compute the complexity between k-means and k-medoids clustering algorithm for uniform and normal distribution of data points and concluded that average time taken by k-Means algorithm is more in both the cases.

P. Amudha et.al. (2011), observed that Random forest gives better detection rate, accuracy and false alarm rate for Probe and DOS attack.

Deepthy k Denatious et.al. (2012), describe different data mining techniques applied for detecting intrusions. Also describe the classification of IDS and its working.

Roshan Chitrakar et.al.(2012), proposed a hybrid approach to intrusion detection by using k-Medoids clustering with Naïve Bayes classification and observed that it gives better performance than K-Means clustering technique followed by Naïve Bayes classification but also time complexity increases when increase the number of data points.

Roshan Chitrakar et.al. (2012) proposed a hybrid approach of combining k-Medoids clustering with Support Vector Machine classification technique and produced better performance compared to k-Medoids with Naïve Bayes classification. The approach shows improvement in both Accuracy and Detection Rate while reducing False Alarm Rate as compared to the k- Medoids clustering approach followed by Naïve bayes classification technique.

Sumaiya Thaseen et.al. (2013), analyzed different tree based classification techniques for IDS. Experimental results show that Random tree model reduces false alarm rate and has highest degree of accuracy.

III. INTRUSION DETECTION SYSTEM

A. *Basic introduction*

The concept of intrusion detection system was firstly developed by scientist named Denning in year 1987. Basically intrusion is nothing but an unauthorized access but in this paper we are basically focusing on one particular kind of problem of malicious attack by attacking(malicious) packets , that will may harm to our personal system or data .Let's see system architecture.

B. *System's flow*

The system architecture has two phase training phase and detection phase. The following process is carried out in the training phase.

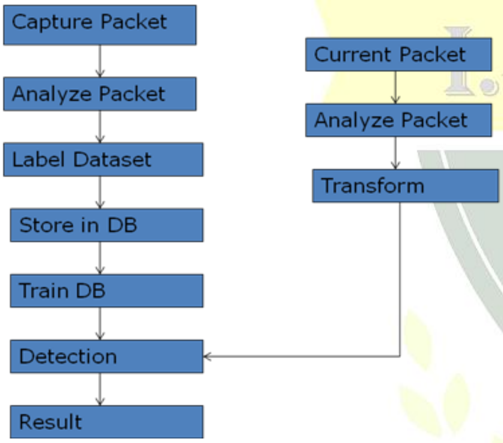

Fig. 1. System's Flow diagram.

1) Capture packet: In this system we capture the packet over the network through the adapter. But java it is not possible to capture packet through the hardware so we use third party library JPCAP and WinPCAP. JPCAP is tool for online network traffic capture and analysis. A JPCAP is also used for developing packet capture application in java. WinPCAP is used to capture packet on windows system.

2) Analyze packet: After capturing packet we need to analyze the packet. The capturing packet may be belongs to different instance for example TCP instance UDP instance etc. The main role of analyze packet to find out packet belongs to which instance.

3) Label dataset: In label dataset we normalize the features of packet. For example take two features of packet flag, length. flag may be true or false we normalize the these as considering true as 1 and false 0.And length is divided into three part low as 0,medium as 1,high as 2.After normalize we make a cluster of these packet in one database as separate cluster for TCP packet, UDP packet, SMTP packet and so on.

4) Stored in DB: Whatever packet or the feature of the packet used for the training the database is store in this stored in DB. These are used for temporary purpose to maintain the history.

5) Train DB: In train DB we train the ANN algorithm in such way when input is this for a particular output. We already store the map values for input and output. Take simple example of AND gate there have four possibility the output is one only when one input and other possibility should produce the output zero. The ANN algorithm is used to train the database.

6) Detection phase: The following process is carried out for detection phase. The current packet is capture over the network then it analyze this packet belongs to which instance then this packet is compare with the stored database. If exactly match is found with normal packet then it will display the normal packet. If match is found with attacking packet then it displays the attacking packet. If no match is found then packet get discard.

C. *Algorithm*

1) K-means clustering algorithm: Clustering is nothing but grouping according to the category .In our project we are using k-means algorithm for clustering purpose , whenever packet get received by the system , it is essential that system should make labeling on that packet so that the task of classification will become easy to the system .For that purpose we are using K-means algorithm for clustering . In system database we will make clusters for each attribute of packet according to rage. e.g. we can simplify attribute length as 1 to 50, 50 to 100,100 to 200 and like wise.

2) Artificial neural network (ANN) Algorithm: In this project we are using ANN that is artificial neural network. Basically ANN is used for the purpose of training, and in our project we are using it for training only. ANN works on the bases of neurons and axons. The structure of ANN is similar to human brain .ANN works basically on one concept, which is "weight". In our project we are using ANN to train our database.

D. *System mathematical model*

The relevant mathematics associated with the project is as follows

Let S ≡ {P, F, A, Fun}

Where, P= {p1,p2,p3,...,pn}……..........capturing packets.

F ={f1,f2,f3,...,fn}…..........…….............features

A= {a1,a2,a3,...,an}…................………attack

fun ={fun1,fun2,fun3,...,funn}. ………..functions

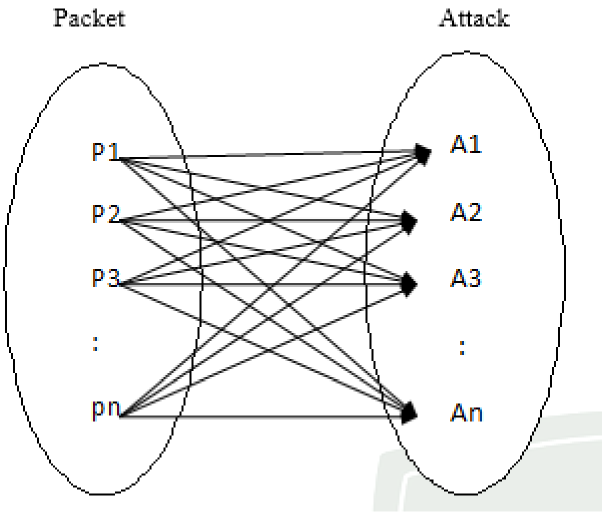

Fig. 2. Venn diagram.

E. *Functions*

Fun-be the set of function.

1. Capturing Packet

2. Training dataset (input packet) = {ANN Algorithm}

3. Run Network

In this model we are representing the relation between packets and attacks. Basically our system shows a many to many approach. The reason behind this is there can be number of packets present throughout the internal. And one important key is that it is impossible to user to predict the number of types of attacks that can be arising on our system. So there can be n number of possible packets with associated attacks.

F. *Advantages of proposed system*

- This system detects harmful attack.
- To improve the performance of system.
- Our system works on two-stage clustering

G. *Limitation of proposed system*

System is able to detect new intrusion attack but it is not able to detect specific type of new attack.

Acknowledgment

With the help of K-means algorithm for the data clustering and (Artificial Neural Network) ANN system gives better as compared to previous technologies. There were number of previous techniques that were used for intrusion detection system , but all previous systems has some disadvantages (e.g. SVM has disadvantage about identifying accurate class of those items which shows average characteristics of both clusters. We can also take another example of Signature database also has disadvantage that , it gives wrong result about those items how's data is not stored in attacking items database.ie it can give result as "Non attacking item" for attacking items also.)

As we are using ANN system will require very less amount of data to be trained properly .and previous systems (e.g. using hybrid approach like 'K-medoids with Naïve Bayes' requires very huge amount of data of training DB.

References

[1] Deepthy K Denatious & Anita John, "Survey on Data Mining Techniques to Enhance Intrusion Detection", International Conference on Computer Communication and Informatics (ICCCI -2012), Jan. 10 – 12, 2012, Coimbatore, INDIA.

[2] Deepak Upadhyaya and Shubha Jain, "Hybrid Approach for Network Intrusion Detection System Using K-Medoid Clustering and Naïve Bayes Classification", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 1, pp 231-236, May 2013

[3] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi and Lilly Suriani Affendey, "Intrusion Detection Using Data Mining Techniques", pp 200-203, IEEE, 2010

[4] Song Naiping and Zhou Genyuan, "A study on Intrusion Detection Based on Data Mining", International Conference of Information Science and Management Engineering, Pp 135- 138, IEEE, 2010

[5] P Amudha and H Abdul Rauf, "Performance Analysis of Data Mining Approaches in Intrusion Detection", IEEE, 2011

[6] Deepthy k Denatious et.al. (2012), describe different data mining techniques applied for detecting intrusions. Also describe the classification of IDS and its working.