ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 4, June 2012

Model for Intrusion Detection System with Data Mining

Deepak Upadhyaya
M.Tech. Scholar
KIT, Kanpur, UP, India
upadhyay@mail.com

Shubha Jain
HOD, CSE Department
KIT, Kanpur, UP, India
shubhj@rediffmail.com

Abstract— Today internet has become very popular medium to communicate between users publicly, due to this, lots of intruder has spread across the internet that perform malicious activity and attack to destroy useful information. There are many techniques to detect cyber-attacks and malicious activities in computer systems in which networks Intrusion detection systems (IDSs) is one of them. It is well-known and widely-used security tool. This paper is presenting general study of the existing techniques of intrusion detection using data mining methods. Furthermore there is proposed model of IDS which is the combination of clustering and classification technique to. At the time of experimental analysis we will compare performance of the proposed IDS with existing IDS in terms of execution time and memory utilization.

Index Terms— Data Mining; Intrusion detection System; Security; Protocol; Data Base.

I. INTRODUCTION

Today every one is using computer networks to connect with each other and it is known as the complexity of the network. To access information and transmit information securely is an important task in the network where we know that information travels publicly. One of the two most publicized threat to security is intruder (other is virus) generally referred to as hacker or cracker. Presented paper is suggesting a mechanism for detecting known or unknown intrusions from the received packets over the network. After receiving the packets, information is extracted and intrusions are identified through matching with the rules stating behavior of normal and abnormal packets to flag any packets that significantly deviate from the behaviour of these normal packets. These deviations are called anomaly or outlier. To improve the performance of the proposed IDS, the data mining concept is introduced.

Organization of this paper is as follows: section I presents introduction of IDS, section II presents a brief survey on related work on Intrusion Detection System and Problem Identification, section III presents a proposed work. Finally, section IV presents concluding remarks and references.

II. LITERATURE SURVEY AND PROBLEM ANALYSIS

In [1] network security through Intrusion Detection Systems (IDSs) has been discussed. IDS are the most approachable and usable technique for network attacks since they allow network administrator to detect policy violations. However, traditional IDS are vulnerable to original and novel malicious attacks. Also, it is very inefficient to analyze from a large amount volume data such as possibility logs. In addition, there are high false positives and false negatives for the common IDSs. Furthermore authors have discussed also on data mining technique and how it is helpful in IDS. Thus, how to integrate the data mining techniques into the intrusion detection systems has become a hot topic recently. This presented whole technique of the IDS with data mining approache in details. In [2] Intrusion Detection System (IDS) is the most important technique to achieve higher security in detecting unknown\malicious\abnormal activities for a couple of years. Anomaly detection is one of the techniques of intrusion detection system. Current anomaly detection is often associated with high false alarm with moderate accuracy and detection rates while it's unable to detect all types of attacks correctly. To overcome this problem, they have suggested a hybrid learning approach. In this approach they combined two different techniques one is K-Means clustering and second is Naïve Bayes classification. Here they used clustering technique of all data into the corresponding group before applying a classifier for classification purpose.

Problem Identification: In previous work it has been observed that [1] used an algorithm known as apriori algorithm which has produced inefficient result for large data set. Furthermore in [2] the k-means clustering method is applied, it often requires several analysis before the number of clusters can be determined. The choice of initial cluster centers may be very sensitive. As each run does not produce same result which is also disadvantages of this algorithm. Due to this reason the resulting clusters depend on the initial random assignments. Moreover it minimizes intra-cluster variance, but does not ensure that the result has a global minimum of variance. Another disadvantage is the requirement for the concept of a mean to be definable which the case is not always. For such datasets the other variants will be appropriate. When the clusters are of different sizes and densities, problems of Empty clusters occurs.

III. PROPOSED WORK

This section is presenting simple model of proposed Intrusion Detection System Using efficient data mining approach. We have analyzed some observations in a critical manner which has leaded us to further work in this field like: research should be data mining concept oriented; data set should be high-quality training data; research should explore applications of data mining in intrusion detection system; research should deal with challenges in data mining. The proposed model will enhance efficiency of intrusion detection system. The proposed model is based on data mining concept with naïve Bayes Classification for anomaly detection field of intrusion detection.

Data Pre-processing: It's already known that lots of attacks incorporated in the dataset fall into various categories. Table 1 is presenting various categories of attacks.

Table 1: Types of Attack

S. No.	Common Attacks	TCP Attacks	UDP Attacks
1	Eavesdropping	TCP SYN Attack	ICMP Attacks
2	Snooping	TCP Sequence Number Attack	Smurf Attacks
3	Interception	TCP/IP Hijacking	ICMP Tunneling
4	Modification Attacks	TCP ACK Flood Attack	
5	Repudiation Attacks		
6	Denial-of-service (DoS) Attacks		
7	Distributed denial-of-service (DDoS) Attacks		
8	Back door Attacks		
9	Spoofing Attacks		
10	Man-in-the-Middle Attacks		
11	Replay Attacks		
12	Password Guessing Attacks		

For the purpose of research, the TCP header format is has been elected because proposed work is based on TCP packets; the scope of data is limited to tcp/ip packets. Table 2 is presenting TCP header format.

Table 2: TCP header Attribute

S. No.	Attribute	Attribute Code
1	Source Port	(SRC_PORT)
2	Destination Port	(DEST_PORT)
3	Sequence Number	(SEQ)
4	Acknowledgement Number	(ACK)
5	Checksum	(TCP_SUM)
6	Urgent Pointer	(URP)
7	Data Offset (4 bit)+Reserved (6 bit)+	Control Flags(6 bit)
8	Window	(WIN)
9	Options	(OPT)
10	Padding	
11	Data	

At the time of receiving data packet (segment) some comparisons are required which are as follows:

- RCV.NXT = NEXT is the sequence number which is expected on an incoming segments, and is the lower edge or left of the receive window
- RCV.NXT+RCV.WND-1 = LAST is the sequence number which is expected on an incoming segment, and is the upper edge or right of the receive window
- SEG.SEQ = FIRST is the sequence number which is occupied by the incoming segment
- SEG.SEQ+SEG.LEN-1 = LAST is the sequence number which is occupied by the incoming segment
- To judged through a segment to occupy valid receive sequence space portion if

RCV.NXT =< SEG.SEQ < RCV.NXT+RCV.WND

or

RCV.NXT =< SEG.SEQ+SEG.LEN-1 < RCV.NXT+RCV.WND

Through these comparisons it can be said that which packet should be received and which not. Table 3 is representing four cases for the acceptability of an incoming packet.

Table 3: Acceptability of incoming packet

Packet(Segment) Length	Receive Window	Test
0	0	SEG.SEQ = RCV.NXT
0	>0	RCV.NXT =< SEG.SEQ < RCV.NXT+RCV.WND
>0	0	not acceptable
>0	>0	RCV.NXT =< SEG.SEQ < RCV.NXT+RCV.WND Or RCV.NXT =< SEG.SEQ+SEG.LEN-1 < RCV.NXT+RCV.WND

Note that when the receive window is zero no packet (segment) should be acceptable except ACK segments. For a TCP it is possible to manage a zero receive window during transmitting data and receiving ACKs. However, even when the receive window is zero, a TCP must process the RST (rejected) and URG (urgent) fields of all incoming segments.

Proposed System Framework: Here proposed model of IDS will extract attribute (defined in table 2) from the captured packets and will apply data preprocessing to prepare training data set. There are several data extraction tools available in Public Domain. Once the data will load into the training data set, it will be prepared for use by the Data Mining approach.

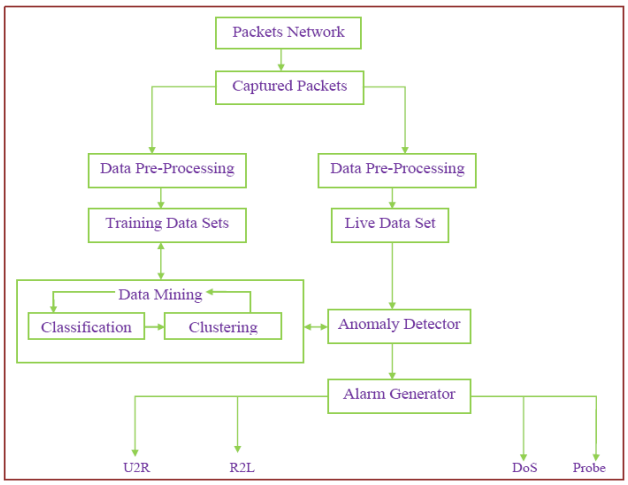

Figure 1: Proposed Model for IDS

Evaluation Measurement: For our experiment, I will use different file size ranges from 2 K to 50 K record data sets. Several performance metrics will include:

- Packet Performance
- Execution time
- CPU Utilization time
- Memory Utilization

Packet Performance: Packet performance is considered

Execution Time: - The execution time is considered the time that an algorithm takes to produce results. Execution time is used to calculate the throughput of an algorithm. It indicates the speed of algorithm.

Memory Utilization: - The memory deals with the amount of memory space it takes for the whole process of Intrusion Detection System.

CPU Utilization: - The CPU Utilization is the time that a CPU is committed only to the particular process of calculations. It reflects the load of the CPU. The more CPU time is used in the execution process, the higher is the load of the CPU.

Table 4: Behavior of Packet

Actual	**Predicted Normal**	**Predicted Attack**
Normal	TN(True Negative)	FP (False Positive)
Intrusions (attacks)	FN (False Negative)	TP (True Positive)

Reasons for Supremacy over other algorithms:-

- Proposed Model will be better than existing to find normal packet performance.
- Proposed Model will be faster than existing in terms of execution time.
- Proposed Model will be smaller than existing and easy to understand and implement.
- It will do not contain complex structure, control flow will be well defined and looping structure will be minimized. Due to the above facts it will take very less time for execution.

IV. CONCLUSION

The proposed mechanism will be compared and evaluated using benchmark dataset. The fundamental result will be the separation of packets between the potential attacks and the normal packets during a preliminary stage into different clusters. It's already known that the clusters are further classifying into more specific categories, for example Probe, R2L, U2R, DoS and Normal. Proposed mechanism will achieve very low false alarm rate while keeping the accuracy and the detection rate on average higher percentage. The proposed mechanism will be capable to classify all data correctly. In the future, we will implement the proposed mechanism for Intrusion Detection System which will be better in detecting attacks.

REFERENCES

[1] Wang Pu and Wang Jun-qing "Intrusion Detection System with the Data Mining Technologies" IEEE 2011

[2] Z. Muda, W. Yassin, M.N. Sulaiman and N.I. Udzir "Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification" 7th IEEE International Conference on IT in Asia (CITA) 2011

[3] Skorupka, C., J. Tivel, L. Talbot, D. Debarr, W. Hill, E. Bloedorn, and A. Christiansen 2001. "Surf the Flood: Reducing High-Volume Intrusion Detection Data by Automated Record Aggregation," Proceedings of the SANS 2001 Technical Conference, Baltimore, MD

[4] http://www.webopedia.com/TERM/I/intrusion_detection_system.html

[5] LI Min "Application of Data Mining Techniques in Intrusion Detection" 2005

[6] KDD. (1999). Available at http://kdd.ics.uci.edu/databases/ - kddcup99/kddcup99.html