Scientific Journal Impact Factor (SJIF): 1.711

e-ISSN: 2349-9745
p-ISSN: 2393-8161

International Journal of Modern Trends in Engineering and Research
www.ijmter.com

An Intrusion Detection based on Data mining technique and its intended importance

[1] Jaymin Parekh, [2] Ronak Patel
[1,2]*Department of Computer Engineering, Ipcowala Institute of Engineering & Technology, Dharmaj, Anand, Gujarat, India- 388430*

Abstract--Intrusion detection is a pivotal and essential requirement of today's era. There are two major side of Intrusion detection namely, Host based intrusion detection as well as network based intrusion detection. In Host based intrusion detection system, it monitors the information arrive at the particular machine or node. While in network based intrusion system, it monitor and analyze whole traffic of network. Data mining introduce latest technology and methods to handle and categorize types of attacks using different classification algorithm and matching the patterns of malicious behavior. Due to the use of this data mining technology, developers extract and analyze the types of attack in the network.

In addition to this there are two major approach of intrusion detection. First, anomaly based approach, in which attacks are found with high false alarm rate. However, in signature based approach, false alarm rate is low with lack of processing of novel attacks. Most of the researchers do their research based on signature intrusion with the purpose to increase detection rate. Major advantage of this system, IDS does not require biased assessment and able to identify massive pattern of attacks. Moreover, capacity to handle large connection records of network. In this paper we try to discover the features of intrusion detection based on data mining technique.

Keywords: Data mining, Knowledge discovery data set, Intrusion detection, Intrusion detection system, Patterns.

I. INTRODUCTION

Intrusion detection is the mechanism to monitor and analyze the massive events occurs in the computer in order to detect abnormal behavior or intrusion named as security problems. Intrusions are the big problem in network and quickly growing illicit activities in the network world. The first attack and its prevention was occurred by Morris Worm in 1988 in send a mail program, then after the techniques have been developed to overcome it and provide better security at network infrastructure. ID is the emerging issue of the research area and many techniques from different area of computer science have been developed for commercial and non-commercial applications.

There are different attacks which violates the computer security policies or standard security practice. The most accurate and accepted attacks are classified and proposed by Kendall [3] in to four categories.

1. Denial of service: this type of attack is trying to disturb the network system or totally interrupt the system or service.
2. Probe: Attacker intended to gather the information about the system for sniffing the traffic and port and address scanning.
3. Remote to Local (R2L): intruder sends a packet to an specific machine in the network but they don't have access to a system and try to make some violation.
4. User to Root (U2R): intruder access to a normal user on the system by negotiating it through sniffing password and gain access to the remote system.

Drawbacks of existing ID

- Existing IDS are generally detects the known attacks but fail to detect novel malicious attacks at the level of network infrastructure.
- Data Overheard: how much data is going to analyze by analyst with proper efficiency, the amount of data growing rapidly.
- False positive: A false positive arises when legitimate attacks are misclassified and treated accordingly.
- False negative: A false negative arises when malicious attacks are classified as normal.

Data mining is one of the technology which helps to improve intrusion detection and addresses the problems arises above. The data mining is the term which designates the process of extracting useful data form huge database [1]. In this interpretation, the term knowledge discovery in databases(KDD) is used to indicate the process of extracting useful facts from huge data sets Data mining, by contrast, denotes to one specific step in this process, In addition to this, it is preceded and surveyed by additional KDD steps, which confirm that the extracted patterns actually correspond to useful knowledge. Certainly, deprived of these additional KDD steps, there is a great risk of finding worthless or uninteresting patterns [2].the KDD mechanism uses data mining technology for pre and post processing to transforms and extracts high level knowledge from low level data. Here, there are fundamental outlines of primary KDD steps.

1. Considerate the basic application domain: first is mounting an application domain and relevant background knowledge and specific goal of KDD.
2. Data Integration and Selection: second is the selecting and combining multiple data sources and choose the relevant data for the analysis task.
3. Data mining: in this step, the algorithm to extract the appropriate patterns from the huge amount of data set.
4. Pattern evaluation: this step designate the actual pattern of knowledge from respected evaluation process.
5. Knowledge illustration: this step intended to represent the discovered patterns in the form of graphical visualization.

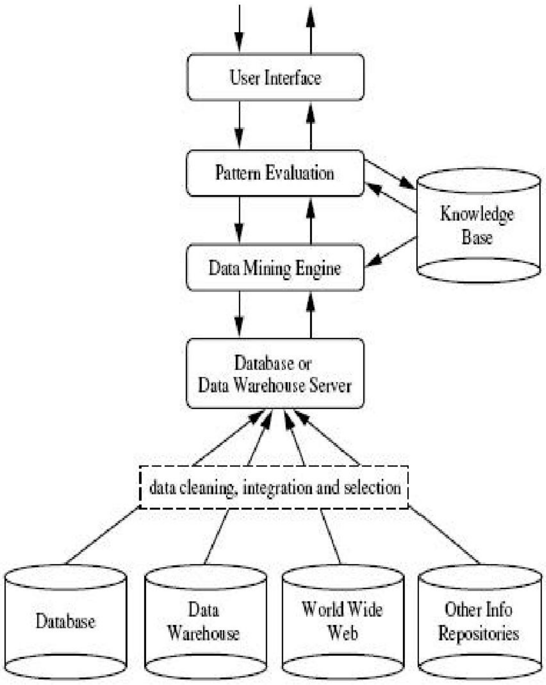


Figure 1:typical architecture of data mining[4].

The data mining provided the following merits:

-it will improve the detection of various types of attacks especially anomaly based intrusion because this approach works on signature matching and try to identify unknown intrusions.
-manage false alarm rate.Terminology of data mining manage the false positive at some acceptable level and it filter out those normal system activities to keep the alarm rate at an adequate level.
-due to learning and incremental process of data mining the activites like normal and abnormal can be detected and novel attacks could be detected precisely.
As consequence,It leads to reduce the less number of false dismissals.
-increase the effeciency.The most vital feature of data mining technology is the ability to get meaningful information from the huge amount of data.the learning feature of data mining increases the high efficiency after the feature extraction step.

II. APPROACHES OF INTRUSION DETECTION SYSTEM

There are two types of intrusion detection system.
2.1 Misuse detection Approach: The terminology behind misuse detection consists of matching network traffic through a model describing known intrusion actions.This approach is largely improved to detect the known attack but ineffective to detect the unknown threats.this taxonomy identify the known signature and represented in the form of perticular pattern.Hence,minor change in the signature may be misclassified.A signature based intrusion system can be work on matching the patterns of network traffic against the data base of signature from known malicious threats.This system is work like anti virus scanning and regular updates of signature make it defencive against the massive events occure inside the network or outside the intrusion detection system[5]. SNORT is the best IDS for signature based ID in which the researchers are able to modify the existing intrusion detection system as well as provide great benchmark function to detect the massive behavior of ID.

2.2 Anomaly based Approach: The terminology behind Anomaly detection designate to analize the profile which represents the normal network traffic behavior.The process is start with detecting the base line profile of the normal genuine traffic activity.Then after new activity that differs the normal model is considered as an anomaly.This approach is possibly recognize the unknown intrusions.On the other hand,this methodology have high false alarm rate.The incremental learning and training of this system can improve the detection accuracy as well as scalability of detecting unseen attacks.

III. CATEGORIES OF INTRUSION DETECTION SYSTEM

There are two types of intrusion detection systems.
3.1Network based Intrusion Detection System: Network based Intrusion Detection system monitor the whole traffic of the network through which the hosts are connected.This system obtain the traffic information from the different host and make decision based on that.Network based intrusion detection based system provides best real time detection of network attacks,Hence it will reduce the network intrusion and make it efficient against the malicious activities in the network.

3.2Host based Intrusion Detection System:In this terminology,the host itself monitor the traffic coming to it and analyze those network traffic and obtain decision from single user.In NIDS,intrusion detection is obtained from whole traffic network rather then single system monitoring.Host based intrusion detection System permits to collection of data on each and every network single user or host.which facilitate the single user to handle traffic and make better image what is going on at each host instead of monitoring the entire network

IV. DATA MINING TECHNIQUES

In this section,there are various data mining techniques which have been applied for detection of intrusion via different research groups.

4.1 Machine learning : Machine learning is the key area in which the problem identified via automatic computation using different algorithm.With the use of user's interest various applications are range in data mining technique that found the general rule in huge data set.on other side of the statistical method,machine learning is well suited for learning patterns with no priory knowledge and dose not intend to require what patterns may be.

4.2 Feature selection:feature selection is the mechanism in which the variables are selected for the purpose to detect a subset of features available from the data and choose for the application of learning algorithm.

4.3 Genetic algorithm: computational biology is the major research area for the genetic algorithm and have been applied for the various fields with the promising results.The REGAL System is used for learning process for the genetic algorithm to first order logic concept description[6][7].Dasgupta and Gonzalez used a genetic algorithm for exploring host baseed not network based IDS[8].

4.4Fuzzy logic: It is the process to solve the ambiguity and error.fuzzy logic is developed from fuzzy set theory dealing with the reasoning that is approximate rather then precisely deducted from classical form classical predicate logic[9].There are various researchers who have apply fuzzy logic rule to classify the normal and abnormal behavior of network traffic.

4.5Support Vector Machine: This is related to supervised learning methods based on classification and regression.support vector machine is going to use data set to separate then in to multiple class with the use of hyper plan.with the use of KDD 99 Data set many researchers uses more convention SVM to identify normal traffic and other types of massive activities.

4.6Hidden Markov Models:A Hidden Markov Model is the mechanism in which the system have been developed based on markov process with unknown parameters and most difficult to determine with known or hidden parameters from noticeable parameters.HMM is the simple dynamic bayesian network.This model is used to detect several types of intrusion which are complex with the sevral steps that may produce over an extended period of time.Authers describe that the HMMs are well in multi-step attack problem.HMMs are give better results then decision trees and neural network in detecting complex intrusion.

V. KDD 99 DATA SET

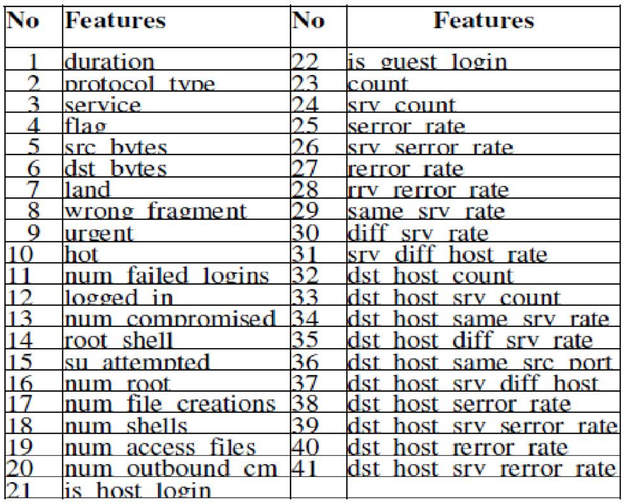
KDD(knowledge discovery Dataset) is introduce by MIT Lincon leboratery and these data set is publically available for the use of different attributes of network traffic[10].

This dataset is generally used by many researcher for detecting intrusion and cross verify the results of real time detection of intrusin detection.They have provided the five million connection records for evaluate and get results of intrusion detection systems.

There are 41 features of this dataset that describe a connection and marked as normal or an attack.

41 features:

- 1-9 stands for the basic features of packet.
- 10-22 for content features.
- 23-31 for traffic features.
- 32-41 for host based features



No	Features	No	Features
1	duration	22	is guest login
2	protocol type	23	count
3	service	24	srv count
4	flag	25	serror rate
5	src bytes	26	srv serror rate
6	dst bytes	27	rerror rate
7	land	28	rrv rerror rate
8	wrong fragment	29	same srv rate
9	urgent	30	diff srv rate
10	hot	31	srv diff host rate
11	num failed logins	32	dst host count
12	logged in	33	dst host srv count
13	num compromised	34	dst host same srv rate
14	root shell	35	dst host diff srv rate
15	su attempted	36	dst host same src port
16	num root	37	dst host srv diff host
17	num file creations	38	dst host serror rate
18	num shells	39	dst host srv serror rate
19	num access files	40	dst host rerror rate
20	num outbound cm	41	dst host srv rerror rate
21	is host login		

Table 1:Features of KDD 99 Dataset[10]

VI. FUTURE SCOPE OF INTRUSION DETECTION

In the recent years,many researchers are try to develop best intrusion detection system but still there are many problems and open issues which have scope to improve the existing system.

In order to gain high output in terms of good acuracy to detecting intrusion,high level human interaction is required.for instance SNORT require expert knowledge to get proper signature of intrusion.most of the current approaches are aim to generate automatic detection system with the use of data mining and machine learning .inappropriate adjustment in the model information is also other open issue for IDS.Selection of proper attributes of dataset may be increase the efficiency and accuracy of the existing intrusion detection system.

VII. CONCLUSION

It is concluded that,discussed technique and approaches have ability to identify intrusion with considerable level.Researchers have developed and analyze multiple data mining technique for intrusion detection system and try to increase accuracy and efficiency based on different parameters.

REFERENCES

[1] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors I 996b). Advances in Knowledge Discovery and Data Mining. AAAI Press/MIT Press.

[2] Fayyad, U. (1998). Mining Databases: Towards Algorithms for Knowledge Discovery. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 22(1):39-48.

[3] Kendall K. A database of computer attacks for the evaluation of intrusion detection systems. Master's thesis, AAI3006082; 1999.

[4] MingXue,Changjun Zhu"applied Research on Data Mining Algorithm in Network IntrusionDetetion.IEEE2009

[5] J. McHugh, A. Christie, and J. Allen, “Defending yourself: The role of intrusion detection systems,” Software, IEEE, vol. 17,no. 5, pp. 42–51, 2000.

[6] Neri, F., "Comparing local search with respect to genetic evolution to detect intrusion in computer networks", In Proc .of the 2000 Congress on Evolutionary Computation CECOO,La Jolla, CA, pp. 238243. IEEE Press, 16- 19 July, 2000.

[7] Neri,F.,"Mining TCP/IP traffic for network intrusion detection", In R. L. de M'antaras and E. Plaza (Eds.), Proc.of Machine Learning: ECML 2000,Ilth European Conference on Machine Learning, Volume 18 10 of Lecture Notes in Computer Science, Barcelona, Spain, pp. 3 13322.Springer, May 3 1- June 2, 2000.

[8] Dasgupta, D. and F. A. Gonzalez, "An intelligent decision support system for intrusion detection and response", In Proc. of International Workshop on Mathematical Methods, Models and Architectures for Computer Networks Security (MMM-ACNS), St.Petersburg.Springer- , 2 1-23 May,200 1.

[9] G. 1. Klir, "Fuzzy arithmetic with requisite constraints”, Fuzzy Sets and Systems, 9 1: 165175, 1997.

[10] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html