Network Intrusion Detection System

ROHIT ROUT
Maharaja Agrasen Institute Of Technology, Delhi, India

Abstract- Intrusion Detection System (IDS) is a system namely a security order that symbolize a care layer for the foundation and watches for some hateful or malicious ventures int the network. Exponential usage of web results in lifting the concerns about by what method to defend the digital information in a securing conduct. Throughout the age, the IDS science has grown extremely to equal the progress of calculating crime. Nowadays, hackers use various types of attacks for listing our calculating's private protected information. Many interruption discovery methods, designs and algorithms will act as a shield towards these attacks.

This main aim concerning this paper search out supply a complete study about the description of intrusion discovery, record, biological clock, and interruption detection patterns, types of attacks, various forms and methods, challenges with allure uses

Indexed Terms- Intrusion detection, Network, IDS attacks, Functionality, Tools, Techniques

I. INTRODUCTION

An Intrusion Detection System is used to discover all types of hateful network traffic and calculating custom that can't be discovered by a common firewall. This contains network attacks against ready aids, dossier compelled attacks on requests, host located attacks in the way that rights escalation, pirated logins and approach to impressionable files, and malware. Security is wanted for the consumers to secure their methods from outside undesired force. Firewall method is individual of the well-known care methods that is used to cover the network. IDS are secondhand in network connected actions, healing uses, charge card frauds, Insurance instrumentality. An IDS is collected of the following three elements: Sensors: - that sense the network traffic or arrangement action and produce events. Console: - to monitor occurrences and alerts and control the sensors, Detection Engine: - that records occurrences record for one sensors in a table and uses a scheme of rules to create alerts from the taken freedom occurrences. There are various habits to classification an IDS contingent upon the type and part of the sensors and the methods secondhand apiece instrument to produce alerts. In many plain IDS implementations all three parts are linked in a sole maneuver or machine.

In this paper, we use different machine learning methods to analyze the types of intrusion . Here we use machine learning methods like Naïve Bayes Classifier,Random Forest, SupportVector Machine method and Maximum Entropy method. Here we will compare these methods based on their accuracy and precision and see which method gives the best result. All the above methods are supervised learning methods. So, in all these cases we need to first train the data.

II. IDS TECHNIQUES

1. 1. Network-based IDS (NIDS): This type of IDS monitors network traffic for suspicious activity and is typically placed at a strategic point within the network to monitor traffic from multiple sources.
2. Host-based IDS (HIDS): This type of IDS monitors the system and files of a single host for suspicious activity.
3. Signature-based IDS: This type of IDS uses a database of known attack patterns (signatures) to identify malicious activity.
4. Anomaly-based IDS: This type of IDS uses machine learning algorithms to identify abnormal behavior that may indicate an intrusion.
5. Behavior-based IDS: This type of IDS monitors the behavior of users and systems to identify suspicious activity.
6. Hybrid IDS: This type of IDS combines the capabilities of multiple IDS types, such as signature-based and anomaly-based detection, to provide more comprehensive protection

- Data Gathering

There is various data set available on internet around intrusion detection system, so for this project I have taken a data set from Kaggle. One caneasily download it from the Kaggle site. This data is appropriate for our project as It has more than 1.2 Lakh rows and with 43 columns. Because of the large sizes of the dataset we can easily split it into multiple test and train dataset to observe our accuracy .

B. Pre-Processing

Different types of attack class that we have in our dataset :-

In attack class normal means 0, DOS means 1, PROBE means 2, R2L means 3 and U2R means 4.

There are various further types of attacks in DOS , PROBE , R2L and U2L

ATTACK CLASS:

DOS: Denial of service is an attack category, which depletes the victim's resources thereby making it unable to handle legitimate requests – e.g. syn flooding. Relevant features: "source bytes" and "percentage of packets with errors"

Probing: Surveillance and other probing attack's objective is to gain information about the remote victim e.g. port scanning. Relevant features: "duration of connection" and "source bytes"

U2R: unauthorized access to local super user (root) privileges is an attack type, by which an attacker uses a normal account to login into a victim system and tries to gain root/administrator privileges by exploiting some vulnerability in the victim e.g. buffer overflow attacks. Relevant features: "number of file creations" and "number of shell prompts invoked,"

R2L: unauthorized access from a remote machine, the attacker intrudes into a remote machine and gains local access of the victim machine. E.g. password guessing Relevant features: Network level features – "duration of connection" and "service requested" and host level features - "number of failed login attempts"

The task is to build network intrusion detection system to detect anomalies and attacks in the network. There are two problems.

Binomial Classification: Activity is normal or attack

Multinomial classification: Activity is normal or DOS or PROBE or R2L or U2R

III. LIST OF COLUMNS FOR THE DATA SET

["duration","protocol_type","service","flag","src_bytes","dst_bytes","land",
"wrong_fragment","urgent","hot","num_failed_logins","logged_in",
"num_compromised","root_shell","su_attempted","num_root","num_file_creations",
["num_shells","num_access_files","num_outbound_cmds","is_host_login",
"is_guest_login","count","srv_count","serror_rate",
"srv_serror_rate",
"rerror_rate","srv_rerror_rate","same_srv_rate",
"diff_srv_rate",
"srv_diff_host_rate","dst_host_count","dst_host_srv_count","dst_host_same_srv_rate",
"dst_host_diff_srv_rate","dst_host_same_src_port_rate",
"dst_host_srv_diff_host_rate","dst_host_serror_rate","dst_host_srv_serror_rate",
"dst_host_rerror_rate","dst_host_srv_rerror_rate","attack", "last_flag"]

1) Handling Outliers

Handling outliers is the process of identifying and dealing with data points that are significantly different from the majority of the data. Outliers can have a negative impact on the accuracy and performance of machine learning models, so it is important to identify and handle them properly.

2) defining relationships (between Y and numerical independent variables by comparing means)
3) giving text output a numerical meaning for anylasis.
4) Removing missing data from dataset

Dropping columns based on data audit report

- *Based on low variance (near zero variance)*
- *High missings (>25% missings)*
- High correlations between two numerical variables

'land','wrong_fragment','urgent','num_failed_logins',"root_shell","su_attempted","num_root",
"num_file_creations","num_shells","num_access_files","num_outbound_cmds","is_host_login","is_guest_login",

'dst_host_rerror_rate','dst_host_serror_rate','dst_host_srv_rerror_rate','dst_host_srv_serror_rate', 'num_root','num_outbound_cmds','srv_rerror_rate','srv_serror_rate'

These are the columns selected to be dropped from the dataset as they were not contributing to the final result of the system .
Variable reduction using Select K-Best technique

The Select K-Best methodology for variable reduction is a strategy for selecting the most important characteristics or variables from a bigger set of data. This method assesses the link between each variable and the target variable using statistical methods, and then chooses a subset of the most important variables based on their score.

The Select K-Best technique is frequently used in data analysis and machine learning to enhance the performance of prediction models by deleting duplicate or unnecessary features and lowering the number of input variables. This can lessen overfitting and enhance the model's interpretability.

Final list of variable selected for the model building using Select KBest
attack_neptune, attack_normal, attack_satan, count, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_same_srv_rate, dst_host_srv_count, flag_S0, flag_SF, last_flag, logged_in, same_srv_rate, serror_rate, service_http

- *Classifying Methods*

There are several Machine Learning Classifying method we will apply here are:
1) Naïve Bayes Classifier
2) Support Vector Machine
3) Random Forest
4) K – nearest Neighbor
5) Decision Tree

COMPARISON OF ACCURACY AND PRECISION OF DIFFERENT MACHINE LEARNING METHOD

METHODS	ACCURACY
Bernouli Naïve Bayes	77.837 %
Guassian Naïve Bayes	79.16 %
Linear SVC	79.590%
SVC	71.019%
Decision Tree	81.244%
K- nearest neighbor	75.398%
Logistic Regression	83.768%
Ridge Classifier	76.050%

IV. FUTURE SCOPE

The deployment's success determines how IDS will be implemented. For the design and implementation phases, planning is crucial. Most of the time, it is preferable to implement a hybrid IDS system that combines network- and host-based IDS. Organizations may make different choices. Because it can monitor numerous systems and because, unlike host-based IDS, it does not call for software to be put on a production system, a network-based IDS is an obvious choice for many enterprises. Several businesses offer hybrid solutions. Therefore, before installing a host-based sensor, a system needs to have the resources it needs. The IDS technology still relies on attack signatures and is reactive rather than proactive.Every time a new type of attack is identified and recorded in the database, the signature database must be updated. The frequency of signature updates varies from vendor to vendor. In the future, we can further reduce the attacks.

CONCLUSION

This paper's main goal is to give an overview of the value and need for intrusion detection systems. The whole research of IDS kinds, life cycles, different domains, forms of attacks, and tools is provided in this work. IDS are growing crucial for current day network user and business security. The term "IPS" refers to security prevention methods. The stages and the phases of the lifecycle are depicted. There are still more obstacles to conquer. Additional strategies can be employed in addition to those expressly illustrated

for anomaly and misuse detection. Further Comparative examination of a few well-known data mining algorithms used in IDS will be done, as well as improving a classification-based IDS through the use of selective feedback techniques.

REFERENCES

[1] Research paper by International Journal of Computer Science and Information Technologies 10. Research paper by Bharathiar University, Coimbatore.

[2] A survey on Intrusion Detection System SVITS , indore

[3] Karthikeyan .K.R and A. Indra- "Intrusion Detection Tools and Techniques a Survey"

[4] Intrusion detection systems using classical machine learning techniques vs integrated unsupervised feature learning and deep neural network Shisrut Rawat, Aishwarya Srinivasan, Vinayakumar Ravi, Uttam Ghosh

[5] Proctor, Paul E. The Practical Intrusion Detection Handbook.

[6] Bace, Rebecca. "An Introduction to Intrusion Detection and Assessment: for System and Network Security Management."

[7] Research paper by Engineering Research Council of Canada and Dalhousie University Electronic Commerce Executive Committee.

[8] Research paper by School of Future Studies & Planning, Devi Ahilya University, Indore.

[9] Corinne Lawrence- "IPS – The Future of Intrusion Detection"- University of Auckland - 26th October 2004.