HYBRID STRATEGY USING HIDDEN MARKOV MODEL FOR INVESTIGATION OF NIDS

Potdar Shruti Suresh¹, Dr.S.L.Deshpande²
¹,²Visvesvaraya Technological University, Department Of P.G. Studies Jnana Sangama, Belagavi

I. INTRODUCTION

As the technology is developing quickly, the network is facing various threats like virus, worms, Trojan horses, rootkits and spyware. These intrusions have to be compelled to be known before any sort of loss in the organization. The network is struggling with the intrusion which affects the efficiency of computer network in terms of transfer speed. The attacker uses the advance option like dynamic ports and address spoofing to avoid the detection. This sort of intrusion can be detected by identifying the patterns and matching them to the network datasets.

Why IDS needed?
1. It analyzes and monitors the activities of user and system.
2. It audits vulnerabilities and system configuration.
3. It provides the pattern matching activity to find out the attacks.
4. It analyzes the abnormal behavior of the system.

It is mainly classified into two types:

1. **Network based IDS:**
   In network based IDS, the unauthorized and malicious behavior of the network traffic is identified. It analyzes the passing traffic on the entire network. It works on promiscuous mode. The attack can be identified in such way that traffic on a network is analyzed by using the technique called Packet Sniffing. It discovers the unauthorized access to a computer network.
   It only gathers the information, identifies the malicious behavior and logged it into the system and then sends alarms to the administrator.
   
   One of the examples of Network based IDS is SNORT.

2. **Host based IDS:**
   In host based IDS the unauthorized and malicious behavior of a specific device is identified. It monitors and analyzes the internal to the computer system. It consists of agents which monitors the system which uses the combination of signature and rules to identify any malicious activity.[1]

The different techniques used in IDS are:

1. **Misuse or Signature based Intrusion Detection:**
   This technique identifies the known patterns by matching them with the observed patterns using the rules.
   
   **Advantage:**
   It finds out the known attacks so it is efficient with low false positive rate.
   
   **Disadvantage:**
   It cannot find the unknown attacks.


2. **Anomaly based Intrusion Detection:**
This technique identifies the anomalous patterns which are different from the expected behavior.

**Advantage:**
It detects the newly generated attacks which are a significant deviation from the normal behavior of the user.

**Disadvantage:**
If the non-intrusive behavior falling outside the normal range is considered as an intrusion which provides the high false positive rate.

3. **Hybrid Detection:**
It combines both Misuse and anomaly based intrusion detection techniques.

II. **LITERATURE SURVEY**

In Do-hyeon Lee et. al[1], provides the method for detection and prevention from the attacks in the system by using the Intrusion Detection System. Multistage procedure provides the efficient way to identify the attacks. Every stage used by the recent intrusion detection system which uses the Hidden Markov Model. This model determines probability of intrusion used in auditing the data. It provides the efficiency to whole system by reducing overheads of intrusion agents. This system has the low probability of error.

In MeghaBandgaret. al[2], provides the Intrusion detection system which plays an important role in detecting the malicious activities. IDS enables the administrator in securing the system. This paper provides the method called Hidden Markov Model which is the probabilistic approach. Using this method, it is easy to find out the internet attacks. In this paper IDS used to detect only the signature based attacks which encompass only specific traffic i.e. known intrusive attacks. This paper describes the single and multiple HMM for source separation both on IP and Port Information. This approach reduces the false positive rate and provides more accuracy to the system.

In Ms. Purnima Singhet. al[3], its very much essential to prevent valuable information in the network and prevent the internal resources from the unauthorized access. As the access to the internet growing so fast, the no. of attacks also goes on increasing. Intrusion detection system ensures the safety and privacy of the system and network by examining the arriving packets for known signature. Snort is signature based Intrusion Detection System which provides basic analysis and generates the alerts when an attack is detected.

Winpcap is an open source library for capturing the packets and network analysis. It has some features like filtering the packets according to user specify rules. It gathers the information of raw packets, sucha its source IP address, packet length.

In Nahla Ben Amoret. al[4], this paper provides the technique to identify the intrusion using Naïve Bayes Network. KDD’99 is the dataset used to evaluate the intrusion. In this paper, the attack can be identified at three levels, one it is considered as whole attack or grouped them into four main categories or just it is considered as normal or abnormal behavior. This structure provides efficient classifiers to find out the intrusions.

III. **PROPOSED SYSTEM**

In this project, the system is trained using the snort and Hybrid classifiers to identify the intrusion in the network. In the first phase the system is trained with the Snort to analyze the packets in the network using the KDD cup 99 data sets. This model is used to identify only the signature based intrusions in the system. While in the second phase the system trained with a Naive Bayes classifier and Hidden Markov Model. For improving the accuracy of the classifier for limited training data sets, we use a way known as Hybrid classifier. We tend to learn Hidden Markov Model and Naive Bayes classifier on Input data set and trained the rule and then use the combined classifier model
learnt to process internet traffic and find the attacks. This hybrid approach is used to identify the Anomaly based intrusions in the system. Using this combined classifier, system’s accuracy is improved.

Models:

1. **HMM:**
   This model is a probabilistic approach which underlays the probabilistic process. These processes are hidden, but these can be observed with the help of another set of probabilistic process which produces the sequence of observed symbols.
   The hidden Markov model is an anomaly based intrusion detection method, which models the normal user behavior as well as the unknown behavior. Hidden Markov Model is a machine learning method used for building the finite state machine.

2. **Naïve Bayes Classifier:**
   Hidden Markov Model had some of the issues while we try to implement it in real time. So, it could be combined with Naïve Bayes classifier to make it work in real time. Naïve Bayes classifier to make it work in real time. Naïve Bayes Classifier Profiles the network traffic, i.e. it handles the deviation in the network traffic. Naïve Bayes Classifier used for online analysis of traffic. Traffic which is considered as abnormal by the Naïve Bayes Classifier is passed to the offline Hidden Markov Model. It computes the probability of connections present in the network. So combining the Naïve Bayes Classifier and Hidden Markov Model, the Hybrid model is formed. Due to which system becomes efficient one. [4]

3. **Snort:**
   Snort is one of the network based intrusion detection system which combines the advantage of signature based and anomaly based detection methods. Snort mainly used to passively monitor the network traffic and generates the alarm when an attack is detected. Snort can detect the attacks by matching the pattern, but when a new attack comes then system can’t find so by using snort we overcome this limitation by analyzing the real time traffic when packets arrives at the network, snort checks their behavior if it degrades network performance then snort stop to process the packets. It discards those packets and stores the information in the signature database.[3]

4. **KDD 99 cup data set:**
   KDD 99 cup data set is the Knowledge Discovery Dataset. It is used with Snort which is a Network Intrusion Detection method to identify the attacks and provides the predictive approach which distinguishes between the intrusive behavior and normal behavior. This database consists of a standard set of data which is to be audited, which contains the intrusions.[4]
Methodology

Training and Testing System:

- **Feature Selection**: This module takes training data set as input; training data set is helpful for multi-stage classifier modelling.
- **Multistage classifier modelling**: This module contains Naïve Bayes model.
- **HMM Model**: Probabilistic approach based on input data set and then use the HMM model to generate further dataset and trained the classifier.

![Diagram of Training and Testing System]

The modules are showed in diagram, they are:
- **IDS**: An IDS is used to detect malicious activities.
- **Feature selection**: This module takes training data set as input; training data set is helpful for multi-stage classifier modelling.
- **Multistage classifier modelling**: This module contains Naïve Bayes model.
- **HMM Model**: Probabilistic approach based on input data set and then use the HMM model to generate further dataset and trained the classifier.

**IV. EXPERIMENTAL RESULTS**

In the first phase, the system is trained using the snort technique and KDD 99 dataset which consists of different snort rules.

By default some of the rules are available in KDD 99 datasets, like type of the attack, on which port attack has to be done, source address and destination address.

After adding the snort rules, Hidden Markov Model is trained which is specifically used for Snort. This Hidden Markov Model is constructed by using the jHMM library.

It will detect the attacks by using the snort technique.

If the attack is not detected in snort technique, then we can use the second phase where we train the system using Naïve Bayes classifier and Hidden Markov Model. This machine learning classifier used to identify the patterns.

True positive rate is the probability of alarms that are generated by an Intrusion Detection System when intrusion is detected.

True positive rate = true positive/true positive + false negative.
False positive rate is the probability of alarms that are generated by an Intrusion Detection System when intrusion is not detected. This may cause the non-availability of connection which leads to the loss of productivity due to downtime.

**Performance Graph for True Positive Rate:**

![True Positive Rate Graph](image)

**Performance Graph for False Positive Rate:**

![False Positive Rate Graph](image)
V. CONCLUSION AND FUTURE WORK

In this project work, machine learning technique is used to detect intrusion in the network. We implemented in live and captured live internet traffic on an Ethernet interface. We have implemented three classifiers and measured its accuracy. Probabilistic approach is easy to use and provide more accurate and high false positive rate due to which the system works more efficiently. As a future work we plan to do ensemble of multiple classifier to achieve the desired accuracy.

REFERENCES


