Intrusion Detection System in Data Mining using Hybrid Approach

Sahil Sanjay Tanpure
Department of Information Technology
G. H. Raisoni College of Engineering & ManagementPune, India

JayrajJagtap
Department of Information Technology
G. H. Raisoni College of Engineering & ManagementPune, India

Gunjan D. Patel
Department of Information Technology
G. H. Raisoni College of Engineering & ManagementPune, India

ApashabiPathan
Department of Information Technology
G. H. Raisoni College of Engineering & ManagementPune, India.

Zishan Raja
Department of Information Technology
G. H. Raisoni College of Engineering & ManagementPune, India

ABSTRACT

Nowadays, computer security has become very familiar question in the society as nearly everyone has connected their computers to internet to get access to information from various informative sources and send or transmit messages in today’s much complex computer networking world. The most common security threats are intruder which is generally referred as hacker or cracker and the other is virus. To protect the computer on network from intruders, Intrusion Detection Systems are very much important defensive measure component. In this paper we propose a hybrid approach which is the combination two algorithms for clustering and classification that are K-Means and Naïve Bayes respectively. Using KDD Cup’99 dataset we’ll be evaluating the performance of our proposed approach. The evaluation will show that new type of attack can be detected effectively in the system and efficiency and accuracy of IDS will improve in terms of detection rate along with its reasonable prediction time.

Keywords

Data mining, Intrusion detection system, K-Means, Naïve Bayes, sensors.

1. INTRODUCTION

Nowadays, computer networks has become complex, as nearly everyone with a computer has connected it to the Internet to transmit messages and access information. Along with it the complexity of network increases and so the question of security becomes more and more familiar and important as people want to keep their possessions as secure as possible so with it the depth knowledge of computer network protocols increases and becomes important which includes; study of Transmission Control Protocol (TCP), Internet Protocol (IP) which are the most common ones but amongst others like User Datagram Protocol (UDP) etc., are also important...The most publicized to security is intruder generally referred to as hacker or cracker. A common example of current time is security system in a house but now that technology and attackers has advanced we have adopted security systems on our computers as well. One such security system is an Intrusion Detection System (IDS) which identify security breaches in a system which can be any action the owner of the system deems unauthorized (intrusion).

Intrusion detection system is of two types:

1. **Network Intrusion Detection System (NIDS)**

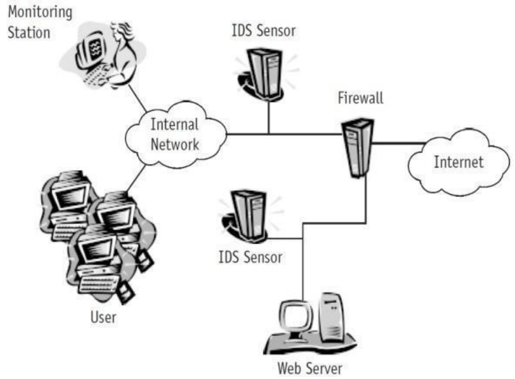


Figure (1) NIDS

Identifies intrusions by examining network traffic and monitors multiple hosts. Network Intrusion Detection Systems gain access to network traffic by connecting to a hub, network switch configured for port mirroring, or network tap. An example of a NIDS is Snort.

2. **Host Intrusion Detection System (HIDS)**

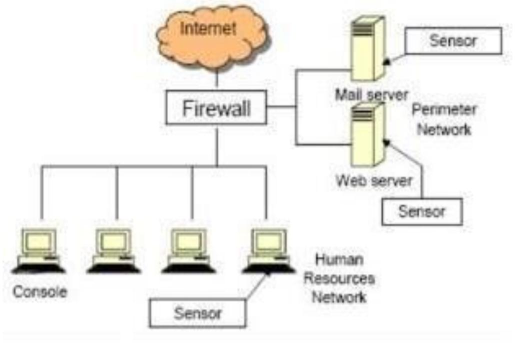


Figure (2) HIDS

It consists of an agent on a host which identifies intrusions by analysing system calls, application logs, file-system modifications (binaries, password files, capability/acl databases) and other host activities and state.

Newly used technique in intrusion detection is use of different data mining techniques. Data mining is well known as process of data retrieval which is retrieved from big collection of data which is further used to retransform into statistically significant events and structures in data. K-Means, Naïve Bayes, ID3, NB Tree, ANN, etc. are some of the data mining techniques used to keep track of clustering, classification, association, and sequence analysis.

In this paper, we are proposing a hybrid approach using K-Means clustering and Naïve Bayes Classification methods for IDS to detect and prevent computer from different attacks along with their details.

2. BACKGROUND

In this section we present a list of the most popular attacks that an intrusion detection system may need to detect. The types of attacks presented here are studied for our research. The different attacks that may occur are:

Denial of Service (DOS)-

Attacker usually occupies all system sources, disables system resources, and engages all computing or memory resources to be too busy to handle legitimate requests or deny legitimate users from accessing a machine. Examples of attacks are Smurf, Mail bomb, SYN Flooding, and Ping Flooding.

Remote to User (R2L)-

Attacker sends packets to remote machine over a network and exploits the network vulnerability to gain local

access as a user of that machine. Examples of attacks are Ftpwrite, SQL Injection, etc.

User to Root (U2R)-

Attacker takes the advantage of system leak by accessing a normal user's account on the system and exploits system vulnerabilities to get legal administrator access to the system. Examples of attacks are Load module, Perl.

Probing-

Attacker performs some preparation step before launching attacks by scanning a network of computers to gather information or to find known vulnerabilities. The attacker will use this information to determine the targets and the type of operating system. Examples of attacks are Nmap, Satan, Ipsweep, Mscan.

3. MOTIVATION

Internet has become the main factor of global information that contains commercial, social and cultural activities. As the internet plays an important role in our day to day life there is increase in number of cyber-attacks and threats. The Initial step in making or launching an attack the attackers scan the machine on the internet to discover network system and host that seems vulnerable. They exploit the system by altering the data, system configuration, planting malicious code, stealing identities, authorizing themselves. They also spread malicious code through e-mails, downloads, messaging and even file sharing. Thus attacks are increasing day by day. The attack can be launched to vulnerabilities in very short time and with little effort. Thus there are great numbers of attacks and it makes much harder for current system to protect themselves from these kinds of attacks.

There is rapid increase in usage of communication system. And some of them are working against the one protecting the cyber infrastructure. As there is a dramatic rise in legitimate activity of both network and communication system, it has made easier for the attackers to hide their activity in a much larger crowd. The increase in number of users of internet with high speed and interconnected machines, it has become easier for attackers to launch large number of attacks or spread malicious code. The flaws and vulnerabilities in protocol design and implementation, complex software code, misconfigured system and inattentiveness of the system operations, leaves a number of machines open to being co-opted by malicious hackers who infect them with malware. Thousands of machines called as Zombies can be called upon to launch attack against a resource at very short period of time. Approaches for discovering such attacks require human intervention and have very high false positive and false negative rates.

4. LITERATURE SURVEY

Intrusion Detection System (IDS) have become an important building block of any sound defense network infrastructure. Malicious attack has brought more adverse impact on the network than before increasing the need for effective approach to detect and identify such effects more effectively.

Naive Bayes is one of the classification models that predicts very fast due to the less complexity functioning of it. Fast prediction is also the reason for a lot work done in recent years using Bayesian approach. In [1] a new hybrid learning approach is proposed that combines K-Mean clustering, Naive Bayes (statistical) also known as KMNB with Decision Table Majority (rule based) approaches.in that an experiment is carried out to evaluate the performance of the proposed approach using KDD Cup '99 dataset. But Naïve Bayesian network is a restricted network that has only two layers and assumes complete independence between the information nodes and in this author has not used KDD cup

'99 benchmark data set in existing dataset. In [2] methods used for these systems include using anomaly detection or a signature database. In this research they used both anomaly detection and a signature database using data mining techniques. The solution provides a tool that would run data mining tools against a log file to detect patterns that may be considered an unauthorized activity. The tool gains additional patterns as time goes by and grows more effective. But it is not a No real time application. In [3] the author presents intrusion detection model based on Decision Tree algorithm and Apriori clustering algorithm. Both Algorithms of Data Mining in Intrusion Detection System are able to predict new type of attacks based on the training data sets. In [4] author points out differences in host and network-based intrusion detection techniques to demonstrate how the two can work together to provide additionally effective intrusion detection and protection. But a series of theoretical and practical problems to be resolved and a number of key technologies are required to make further deep study. In [5] the author uses the commercially available intrusion detection systems are signature based that are not capable of detecting unknown attacks. In this, author analyses a classification model for misuse and anomaly attack detection using decision tree algorithm. But decision trees use a pre-classified dataset to learn to categorize data based on existing trends and patterns. In [6] the author proposes efficient classifier Naïve Bayes on reduced datasets for intrusion detection. Empirical results show that selected reduced attributes give better performance to design IDS that is efficient and effective for network intrusion detection. But U2R attacks cannot be detected.

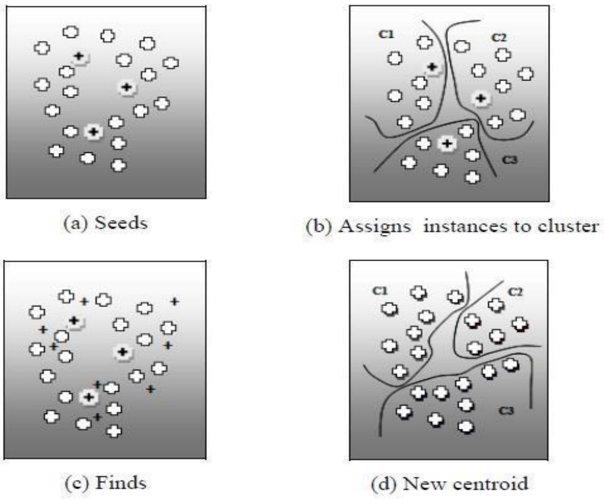
5. PROBLEM STATEMENT

To develop an application which will protect the system from several focused attacks, which are initiated by various intruders. There are various methods to develop a security application. Many of the approaches earlier accepted and performed had many problems which approximately didn't give the desired outputs and security from several intrusions. So we came up with a hybrid approach that is IDS in data mining using hybrid approach. This hybrid approach comprises of K Means clustering algorithm and Nave Bayes classification algorithm. Hybrid approach to provide security to the system from various intruders will overcome the problems which were raised using other approaches. So we proposed a Hybrid Approach to effectively detect the attack on system using hybrid approach in data mining.

5.1 DESCRIPTION

K-Means Algorithm-

The network intrusion class labels are divided into four main classes, which are DoS, Probe, U2R, and R2L. Figure (3) (a) through Figure (3) (d) shows the steps involved in the K-Means clustering process. Figure (4) will later show the final overall result with application of the classification approach. The main goal of utilizing K-Means clustering is to split and group data into normal and attack instances. K-Means clustering methods partition the input dataset into k-clusters according to an initial value known as the seed-points into each cluster's centroids (cluster centres), i.e. the mean value of numerical data contained within each cluster. In our case, we choose k = 3 in order to cluster the data into three clusters (C1, C2, C3). Since U2R and R2L attack patterns are naturally quite similar with normal instances, one extra cluster is used to group U2R and R2L attacks. Back to Figure (b), each input will be assigned to the closest centroid by squaring distances between the input data points and the centroids. New centroids will then be generated for each cluster by calculating the mean values of the input set assigned to each cluster as shown in Figure (c). The steps in Figures (b) and (c) are repeated until the result reached a convergence as shown in Figure (d).



(a) Seeds

(b) Assigns instances to cluster

(c) Finds

(d) New centroid

Figure (3) K-MeansClustering

The K-Means algorithm works as follows:

Select initial centres of the K clusters. Repeat steps 2 through 3 until the cluster membership stabilizes.

Generate a new partition by assigning each data to its closest cluster centres.

- Compute new clusters as the centroids of the clusters.

Naïve Bayes classifier

Some behaviour in intrusion instances is similar to normal and other intrusion instances as well. In addition, a lot of algorithms including K-Means are unable to correctly distinguish intrusion instances and normal instances. In order to improve this shortcoming in classification, we combined the K-Means technique with Naïve Bayes classifier. Naïve Bayes has become one of the most efficient learning algorithms. Naïve Bayes are based on a very strong independence assumption with fairly simple construction. It analyses the relationship between independent variable and the dependent variable to derive a conditional probability for each relationship.

Using Bayes Theorem we write:

$$P(H|X) = P(X|H)\, P(H) / P(X) \text{ ----- (1)}$$

Let X be the data record. Let H be some hypothesis representing the data record X, which belongs to a specified class C. For classification, we would like to determine P(H|X), which is the probability that the hypothesis H holds, given an observed data record X. P(H|X) is the posterior probability of H conditioned on X. In contrast, P (H) is the prior probability. The posterior probability P (H|X), is based on more information such as background knowledge than the prior probability P (H), which is independent of X. Similarly, P (X|H) is posterior probability of X conditioned on H. Bayes theorem is useful because it provides ways to calculate the posterior probability P(H|X) from P(H), P(X), and P(X|H).

We consider five category classes (C1 = Normal, C2 = DoS, C3 = Probe, C4 = R2L, and C5 = U2R). Given X, we can predict C1, C2, C3, C4, and C5. The Bayes rule is shown in Equation (2).

$$P(Ci|X) = \frac{P(X|Ci).P(Ci)}{P(X)} \text{ ------------(2)}$$

Where Ci represents the category of classes and X is the data record. X may be divided into pieces of instances, say x1, x2..., xn which are related to the attributes X1, X2... XN, respectively. The probability obtained is shown in the following Equation (3).

$$P(Ci|X) = \frac{P(x1|Ci).P(x2|Ci)\ldots.\ P(xn|Ci).\ P(Ci)}{P(X)} \text{ -------(3)}$$

The denominator P(X) is always constant for all classes. Thus, it can be ignored as in Equation (4).

$$P(Ci|X) = P(x1|Ci).P(x2|Ci)\ldots.P(xn|Ci).\ P(Ci) \text{ -------- (4)}$$

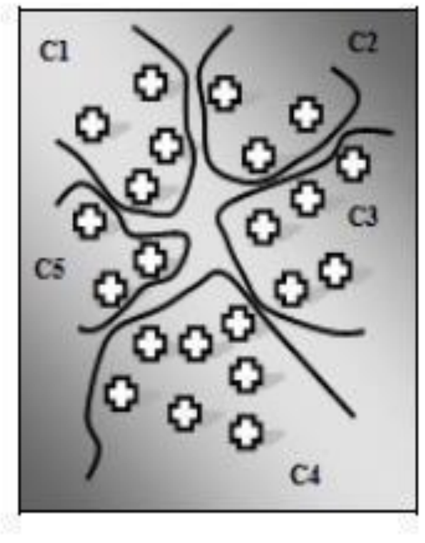


Figure (4)

Figure (4) shows Naïve Bayes classifier that are used to classify classifies all three clusters as illustrated in Figure

(3)(d) into more specific categories, which are Probe, Normal, Dos, U2R, and R2L. The combination of these classifiers with the K-Means clustering technique showed an encouraging improvement as compared to previous approaches. The results are surprisingly better in terms of accuracy, detection rate, and false alarm rate.

6. PROPOSED SYSTEM

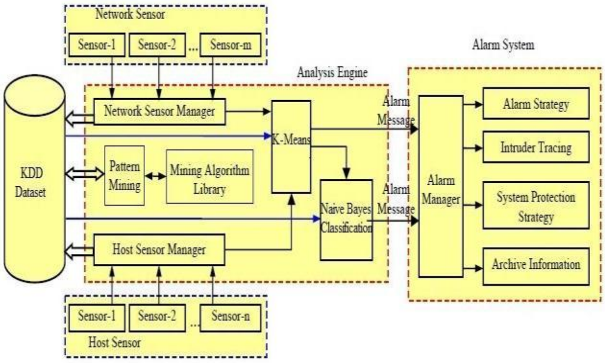


Figure (5) Proposed System Architecture

The proposed system is divided into four parts that is KDD data set, analysis machine, alarm system, host and network sensor. The KDD data set is used for storing the dataset. Network and Host sensors are used for detection of intrusion and Analysis Engine is used for analysing the data as whether the data is an intrusion or not. The alarm system is used for giving an alarm to the user about the intrusion in the system. Sometimes the alarm system can make its owndecision of tracing the intruder or put some system protection strategy or archive information. In the proposed system we use a KDD cup '99 data set for storing data. Input data taken are port numbers, host address, internet protocol addresses, etc. Data from the host sensor and network sensor is stored in KDD data set. If the data is already present then the data is sent directly to the K-Means for clustering and then the alarm message is sent. The alarm manager would then detect and apply some strategy as alarm strategy or tracing the intruder or system protection strategy or to archive information. The data from the KDD data set is put in the pattern mining and compared with the mining library algorithm. If the data set is matched with data in mining algorithm library, it is sent back to KDD dataset and passed on to Naïve Bayes classification. If the data is old or the attacker pattern is already stored then the pattern is sent to K-Means for clustering the data and then sent for alarm message. But if the pattern is new then the pattern is put in K-Means for clustering then it is sent for classification in Naive Bayes. In the Naïve Bayes the data is trained and saved for further use. Then the message is sent for alarm. Then the alarm manager would sort according to the attack pattern and perform accordingly as to give an alarm or trace an intruder or protect the system or to archive the information. The system takes decision according for the type of attack being made or the type of intruder trying to attack the system.

7. ACKNOWLEDGMENTS

It gives us great pleasure in presenting the preliminary project report on `IDS: Intrusion Detection System in Data Mining using Hybrid Approach'. We would like to take this opportunity to thank our internal guide Prof. APASHABI PATHAN for giving us all the help and guidance we needed. We are really grateful to her, for her kind support. Her valuable suggestions were very helpful.

We are also grateful to Prof. SIMARAN KHIANI, Head of Information Technology Department, G.H.Raisoni College of Engineering & Management for her indispensable support, suggestions.

8. CONCLUSION

We proposed IDS in Data Mining using hybrid approach using K-Means and Nave Bayes algorithms. Using this we will improve detecting speed and accuracy as a goal, and proposing more efficient associate and cluster rules mining method as comparing algorithm to abnormal detecting experiment based on network, and will improve the support and credit.

9. REFERENCES

[1] AditiPurohit, Hitesh Gupta, "Hybrid Intrusion Detection System Model using Clustering, Classification and Decision Table" IEEE 2013.

[2] Jonathon Ng, Deepti Joshi, Shankar M. Banik, "Applying Data Mining Techniques to Intrusion Detection" IEEE 2015.

[3] [3] TruptiPhutane, Prof. ApashabiPathan, "Intrusion Detection System using Decision Tree &Apriori Algorithm"IJCET 2015.

[4] Dr.SaurabhMukherjeea, Neelam Sharma, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction" IEEE 2012.

[5] Duanyang Zhao, QingxiangXu, ZhilinFeng, "Analysis and Design for Intrusion Detection System Based on Data Mining" IEEE 2010.

[6] Dr. M. Hanumanthappa, Manish Kumar, Dr. T. V. Suresh Kumar, "Intrusion Detection System Using Decision Tree Algorithm" IEEE 2012.