A Combined Approach for Intrusion Detection System Based on the Data Mining Techniques.

Pragya Diwan[1,] Dr. R.C Jain[2]
[1]Research Scholar, Department of Information Technology, SATI Vidisha (M.P.),
[2]Director, SATI Vidisha (M.P.)

Abstract:

All most all existing intrusion detection systems focus on low-level attacks, and only generate isolated alerts. They can't find logical relations among alerts. In addition, IDS's accuracy is low; a lot of alerts are false alerts. To reduce this problem we propose a hybrid approach which is the combination of K-Medoids clustering and Naïve-Bayes classification. The proposed approach will be clustering all data into the corresponding group before applying a classifier for classification purpose. The proposed work will explore Naïve-Bayes Classification and K-medoid methods for intrusion detection and how it will useful for IDS. The reasons for introducing Naïve Bayes Classification are the involvement of many features where there is no dividation between normal operations and anomalies. Thus Naïve Bayes Classification can be mined to find the abstract correlation among different security features. In this, we will present implementation results on existing intrusion detection system and K-medoid cluster technique with Naïve Bayes classification for intrusion detection system. An experiment is carried out to evaluate the performance of the proposed approach using our own created dataset. Result show that the proposed approach performed better in term of accuracy, detection rate with reasonable false alarm rate.

Keywords: *Association analysis, Database Protocol, Database, Data preprocessing, Data mining, Internet, Intrusion detection.*

I. INTRODUCTION

Information security technology is an essential component for protecting public and private computing infrastructures. With the widespread utilization of information technology applications, organizations are becoming more aware of the security threats to their resources. No matter how strict the security policies and mechanisms are, more organizations are becoming susceptible to a wide range of security breaches against their electronic resources. Network-intrusion detection is an essential defense mechanism against security threats, which have been increasing in rate lately. It is defined as a special form of cyber threat analysis to identify malicious actions that could affect the integrity, confidentiality, and availability of information resources. Data mining-based intrusion-detection mechanisms are extremely useful in discovering security breaches.

II. INTRUSION DETECTION SYSTEM

An intrusion detection system (IDS) is a component of the computer and information security framework. Its main goal is to differentiate between normal activities of the system and behavior that can be classified as suspicious or intrusive [1]. IDS's are needed because of the large number of incidents reported increases every year and the attack techniques are always improving. IDS approaches can be divided into two main categories: misuse or anomaly detection [1]. The misuse detection approach assumes that an intrusion can be detected by matching the current activity with a set of intrusive patterns. Examples of misuse detection include expert systems, keystroke monitoring, and state transition analysis. Anomaly detection systems assume that an intrusion should deviate the system behaviour from its normal pattern. This approach can be implemented using statistical methods, neural networks, predictive pattern generation and association rules among others techniques. In this research using naïve byes classification with clustering data mining techniques to extract patterns that represent normal behavior for intrusion detection. This research is describing a variety of modifications that will have made to the data mining algorithms in order to improve accuracy and efficiency. Using sets of naïve byes classification rules that are mined from network audit data as models of "normal behavior." To detect anomalous behavior, it will generate Naïve Byes classification probability with clustering followed from new audit data and compute the similarity with sets mined from "normal" data. If the similarity values are below a threshold value it will show abnormality or normality.

2.1 Types of Intrusion Detection System

2.1.1 Network intrusion detection systems (NIDS):

Monitors packets on the network wire and attempts to discover an intruder by matching the attack pattern to a database of known attack patterns. A typical example is looking for a large number of TCP connection requests (SYN) to many different ports on a target machine, thus discovering if someone is attempting a TCP port scan. A network intrusion detection system sniffs network traffic, by promiscuously watching all network traffic.

2.1.2 Host based intrusion detection system (HIDS):

A host based intrusion detection system does not monitor the network traffic; rather it monitors what's happening on the actual target machines. It does this by monitoring security event logs or checking for changes to the system, for example changes to critical system files or to the systems registry. Host based intrusion detection systems can be split up into:

- **System integrity checkers**: Monitors system files & system registry for changes made by intruders (thereby leaving behind a backdoor). There are a number of File/System integrity checkers, such as "Tripwire" or " LAN guard File Integrity Checker".
- **Log file monitors:** Monitor log files generated by computer systems. Windows NT/2000 & XP systems generate security events about critical security issues happening on the machine. (for example a user acquires root/administrator level privileges) By retrieving & analyzing these security events one can detect intruders .

2.2 Types of attack on IDS

- **Information gathering:**

1. Network mapping - ping sweeps.
2. DNS zone transfers.
3. E-mail recons.
4. TCP or UDP port scans **-** Enumeration of services indexing of public web servers to find web server and CGI holes.
5. OS fingerprinting.

- **Exploits:**
  Attackers make use of vulnerabilities in target servers or misconfiguration on the system/network.
- **Denial-of-service (DoS) attacks:**
  An attempt to break the system and make it inaccessible to other users. Intruders will attempt to crash a service or machine, overload network or hardware resources, such as overload the links, the CPU, or fill up the disk.

III. DATA MINING

Data mining (DM), also called Knowledge-Discovery and Data Mining, is one of the hot topic in the field of knowledge extraction from database. Data mining is used to automatically learn patterns from large quantities of data. Mining can efficiently discover useful and interesting rules from large collection of data. It is a fairly recent topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition. Data mining is disciplines works to finds the major relations between collections of data and enables to discover a new and anomalies behavior. Data mining based intrusion detection techniques generally fall into one of two categories; misuse detection and anomaly detection. In misuse detection, each instance in a data set is labeled as 'normal' or 'intrusion' and a learning algorithm is trained over the labeled data. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks, as long as they have been labeled appropriately. Data mining are used in different field such as marketing, financial affairs and business organizations in general and proof it is success. The main approaches of data mining that are used including classification which maps a data item into one of several predefined categories. This approach normally output "classifiers" has ability to classify new data in the future, for example, in the form of decision trees or rules. An ideal application in intrusion detection will be together sufficient "normal" and "abnormal" audit data for a user or a program. The second important approach is clustering which maps data items into groups according to similarity or distance between them.Anomaly detection techniques thus identify new types of intrusions as deviations from normal usage. In statistics based outlier detection techniques the data points are modeled using a stochastic distribution and points are determined to be outliers depending upon their relationship with this model. However, with increasing dimensionality, it becomes increasingly difficult and in-accurate to estimate the multidimensional distributions of the data points [1]. However, recent outlier detection algorithms that we utilize in this study are based on computing the full dimensional distances of the points from one another as well as on computing the densities of local neighborhoods [6].

IV. PROPOSED CONCEPT

Data mining techniques have been successfully applied in many different fields including marketing, manufacturing, process control, fraud detection, and network management. Over the past five years, a growing number of research techniques have applied data mining to various problems in intrusion detection. In this will apply to data mining for anomaly detection field of intrusion detection. Presently, it is unfeasible for several computer systems to affirm security to network intrusions with computers increasingly getting connected to public accessible networks (e.g., the Internet). In view of the fact that there is no ideal solution to avoid intrusions from event, it is very significant to detect them at the initial moment of happening and take necessary actions for reducing the likely damage. One approach to handle suspicious behaviors inside a network is an intrusion detection system (IDS). For intrusion detection, a wide variety of techniques have been applied specifically, data mining techniques, artificial intelligence technique and soft computing techniques. Most of the data mining techniques like association rule mining, clustering and classification have been applied on intrusion detection, where classification and pattern mining is an important technique. Similar way, AI techniques such as decision trees, neural networks and fuzzy logic are applied for detecting suspicious activities in a network, in which fuzzy based system provides significant advantages over other AI techniques. This system is anomaly-based intrusion detection makes use of effective rules identified in accordance with the designed strategy, which is obtained by mining the data effectively.

The proposed concept is using data mining techniques. Data mining techniques have been successfully applied in many different fields including marketing, manufacturing, process control, fraud detection, and network management. Over the past five years, a growing number of research techniques have applied data mining to various problems in intrusion detection. In this K-medoids data mining technique has applied for anomaly detection field of intrusion detection.

Whole proposed IDS divided into following module:

1. Database Creation (Suggested Technique)
   - Selecting and generating the data source
   - Data scope transformation and pre-processing
2. Data mining Techniques
   - K-Means Cluster Technique
   - K-Medoids Cluster Technique
3. Naïve Bayes Classification
   - Naïve Bayes Classification with K-Means Cluster Technique
   - Naïve Bayes Classification with K-Medoids Cluster Technique
4. Performance
   - Time Analysis
   - Memory Analysis
   - CUP Analysis
   - Cluster Analysis
   - Probability Analysis of Normality and Abnormality

Database Creation (Suggested Technique):

Selecting and generating the data source:

First the acquisition of data will do. In the case of this research, Sample datasets from DATA- BASE will use. The DATABASE contained high volume network traffic data, and a subset of data ranging a period of 2 – 5 days will be select.

Data scope transformation and pre-processing: For the purpose of the research, the scope of the data was limited to TCP/IP packets. Only six intrinsic features were extracted from each packet within the dataset. These were timestamp, Source IP, Source Port, Destination IP, Destination Port, and Service. The table 4.1 below shows the scope of the input dataset.

Flags	Source	Destination
TCP Port Number	Source TCP Port	Destination Port Number
IP Address	Source IP Address	Destination IP Address
Timestamp		
Services		

Table 4.1: TCP packet frame format

For the purpose of reporting for this research the data was extracted from the DATABASE set using any data base tools. Because there are several data extraction tools are available in Public Domain. This extracted the necessary features and saved the data within the dataset. Once the data was loaded into the pre-processor, it was prepared for use by the Data Mining approach.

Data Mining Technique

Anomaly learning approaches are able to detect attacks with high accuracy and to achieve high detection rates. However, the rate of false alarm using anomaly approach is equally high. In order to maintain the high accuracy and detection rate while at the same time to lower down the false alarm rate, the proposed a combination of two learning techniques. For the first stage in the proposed technique, this grouped similar data instances based on their behaviors by utilizing a K-Mediod clustering as a pre-classification component. Next, using Naïve Bayes classifier this classified the resulting clusters into attack classes as a final classification task. This found that data that has been misclassified during the earlier stage may be correctly classified in the subsequent classification stage.

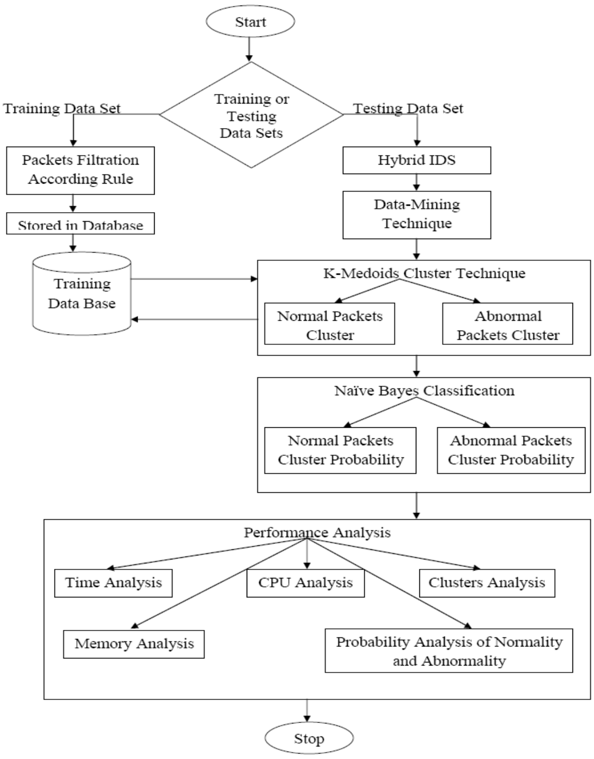
Start
Training or Testing Data Sets
Training Data Set
Testing Data Set
Packets Filtration According Rule
Stored in Database
Training Data Base
Hybrid IDS
Data-Mining Technique
K-Medoids Cluster Technique
Normal Packets Cluster
Abnormal Packets Cluster
Naïve Bayes Classification
Normal Packets Cluster Probability
Abnormal Packets Cluster Probability
Performance Analysis
Time Analysis
CPU Analysis
Clusters Analysis
Memory Analysis
Probability Analysis of Normality and Abnormality
Stop

Figure 4.1.Flow Chart of the Proposed IDS

K-Medoids Cluster Technique: Network intrusion class labels are divided into four main classes, which are DoS, Probe, U2R, and R2L [1-2]. The main goal to utilize K-Mediod clustering approach is to split and to group data into normal and abnormal. K-medoids clustering are sometimes used, where representative objects called medoids are considered instead of centroids. Because it uses the most centrally located object in a cluster, it is less sensitive to outliers compared with the K-means clustering. Suppose that we have n objects having *p* variables that will be classified into k (k<n) clusters (Assume that is given). Let us define j^{th} variable of object i as X_{ij} (i = 1….n, j= 1…..p). The proposed algorithm is composed of the following three steps:

The K-mediods algorithm is composed of the following three steps.

Step 1 : (Select initial medoids)

1.1. Using Euclidean distance as a dissimilarity measure, compute the distance between every pair of

$$d_{ij} = \sqrt{\sum_{a=1}^{p}(X_{ia} - X_{ja})^2} \quad i = 1,\dots,n;\ j = 1,\dots,n \qquad (1)$$

1.2. Calculate P_{ij} to make an initial guess at the centers of the clusters.

$$p_{ij} = \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \quad i = 1,\dots,n;\ j = 1,\dots,n \qquad (2)$$

1.3 Calculate $\sum_{i=1}^{n} p_{ij} \ (j = 1,\ldots,n)$ at each objects and sort them in ascending order. Selected objects having the minimum value as initial group medoids.
1.4. Assign each object to the nearest medoid.
1.5. Calculate the current optimal value, the sum of distance from all objects to their medoids.

Step 2: (Find new medoids)
Replace the current medoid in each cluster by the object which minimizes the total distance to other objects in its cluster.
Step 3: (New assignment)
3.1. Assign each object to the nearest new medoid.
3.2. Calculate new optimal value, the sum of distance from all objects to their new medoids. If the optimal value is equal to the previous one, then stop the algorithm. Otherwise, go back to the Step 2.
The above algorithm runs just like K-means clustering and so this will be called as 'K-means-like' algorithm. The performance of the algorithm may vary according to the method of selecting the initial medoids.

V. CONCLUSION & FUTURE WORK

The proposed research have improved detecting speed and accuracy which is the prime concern of the proposed work, and presents more efficient associate and cluster rules mining method to abnormal detecting experiment based on network. Presented Approach is a hybrid approach which is the combination of K-Mediod clustering and Naïve Bayes classifier. The proposed approach was compared and evaluated using own prepared dataset. Considering the dependent relations between alerts, it proposed an improved cluster Algorithm with naive bayes classification; this hybrid approach can find more accurate probability of normal and abnormal packets. It is applied to find the probability of an attack. Compared with other method, proposed method can find the probability from the training data as well as testing data with high efficiency. Usually when an attack performed, it is very possible that there exist attack cluster transitions. Based on this it use the cluster sequences to filter false alarms generated by IDS, experimental results proved this method is effective and feasible.
We have discussed some observations in a critical manner, which has leaded us to the following recommendations for further research:

Future research should pay closer attention to the data mining process. Either more work should address the (semi-automatic) generation of high-quality labeled training data, or the existence of such data should no longer be assumed. Future research should explore novel applications of data mining that do not fall into the categories feature selection and anomaly detection.
To deal with some of the general challenges in data mining, it might be best to develop special-purpose solutions that are tailored to intrusion detection.

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] Ma Yanchun "The Intrusion Detection System Based on Fuzzy Association Rules Mining" IEEE Conferences 2010.
[4] Lei Li, De-Zhang Yang, Fang-Cheng Shen "A Novel Rule-based Intrusion Detection System Using Data Mining" IEEE Conferences 2010.
[5] Chunyu Miao and Wei Chen "A Study of Intrusion Detection System Based on Data Mining" IEEE Conferences 2010.
[6] Ye Changguo, Wei Nianzhong, Wang Tailei, Zhang Qin and Zhu Xiaorong "The Research on the Application of Association Rules Mining Algorithm in Network Intrusion Detection" IEEE Conferences 2009.
[7] Changxin Song, Ke Ma "Design of Intrusion Detection System Based on Data Mining Algorithm" 2009 IEEE International Conference on Signal Processing Systems.