Hybrid Approach for Network Intrusion Detection System Using K-Medoid Clustering and Naïve Bayes Classification

Deepak Upadhyaya[1] and Shubha Jain[2]

**[1] Dept. of Computer Science, Kanpur Institute of Technology
Kanpur, Uttar Pradesh, India**

**[2] Dept. of Computer Science, Kanpur Institute of Technology
Kanpur, Uttar Pradesh, India**

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' 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 applies clustering on all data into the corresponding group and after that applies a classifier for classification purpose. The proposed work will explore Naïve-Bayes Classification and K-Medoid methods for intrusion detection and how it is useful for IDS. Naïve Bayes Classification can be mined to find the abstract correlation among different security features. In this, we are presenting implementation results on existing intrusion detection system and K-Medoid clustering 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 shows that the proposed approach performs better in term of accuracy, execution time, CPU utilization and memory consumption with reasonable false alarm rate.

Keyword: *Clustering; Classification; IDS, data mining; data preprocessing; association analysis, Protocol, Database, K-Medoid, Bayesian.*

1. Introduction

Nowadays, computer networks are so complex, that nearly everyone with a computer has connected it to the Internet to access information and transmit messages. As complexity increases the question of security becomes more and more familiar as well as the depth knowledge of computer network protocols namely; Transmission Control Protocol (TCP), Internet Protocol (IP) and User Datagram Protocol (UDP) etc. One of the two most publicized threat to security is intruder (other is virus) generally referred to as hacker or cracker. Our aim is to suggest a mechanism for detecting unknown intrusions by identifying packets that are normal and to flag any packets that significantly deviate from the behaviour of these normal packets. These deviations are called anomaly or outlier. This mechanism can be visualized on a 2D topological map formulated by a Data Mining. This method of detection is named anomaly detection [4].

What is intrusion detection system: An intrusion detection system (IDS) inspects all inbound and outbound network activities and identifies suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system. Basically Intrusion detection system (IDS) is a type of security management system for computers and networks. An IDS gathers and analyzes information from various areas within a computer or a network to identify possible security breaches, which include both intrusions (attacks from outside the organization) and misuse (attacks from within the organization). IDS uses vulnerability assessment (sometimes referred to as scanning), it is a technology which is developed to assess the security of a system or network [5].

Intrusion detection function includes:

- monitoring and analysis of System and User activities
- Assessing system integrity and file integrity
- Detection and prevention of network intrusions
- Reorganization of typical attack patterns
- Analysis of Abnormal activity patterns etc.

There are several ways to categorize IDS:

Misuse detection vs. anomaly detection: In misuse detection, the IDS analyzes the information it gathers and compares it to large database of attack signatures.

Essentially, the misuse detection system looks for a specific attack that has already been documented. Like a virus detection system, misuse detection software is only as good as the database of attack signatures that it uses to compare packets. In anomaly detection, the system administrator defines the baseline, or normal, state of the network, traffic load, breakdown, protocol, and typical packet size. The anomaly detector monitors network segments to compare their state to the normal baseline and look for anomalies [4].

Network-based vs. host-based systems: Network-based system or NIDS analyzes individual packet which travels over network. The NIDS can easily detect abnormal packets that are designed to be overlooked by a firewall. Host-based system or HIDS analyzes activity on each individual computer or host [4].

Passive system vs. reactive system: The passive IDS just detects abnormal behavior of the packets but it can not produce any type of alert for the user. But the reactive IDS can easily respond to the abnormal activity by logging off a user or by block the network traffic from the suspected malicious user. IDS and Firewall both are related with network security but the working is totally different from each other [4].

Research ideas and critical analysis: - Before getting into the mechanisms of IDS, it is necessary to understand the commonly used network protocols. TCP is a transport layer protocol and is situated above the IP layer in the (Open Systems Interconnection) OSI model as shown in Figure 1. The transport layer is responsible for establishing sessions, data transfer and tearing down virtual connections [3].

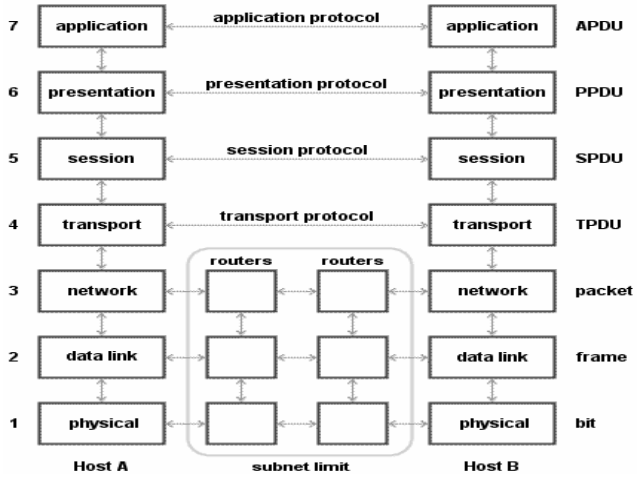

Fig.1 The OSI Model

Figure 2 shows an example of a Transport layer PDU or TPDU.

Fig. 2 Transport Layer PDU

Next is the IP layer in the OSI Model. The IP header is considered to be the next most important PDU (and is called the packet) having source and destination IP addresses with data.

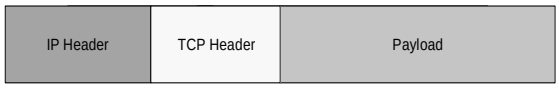

Fig. 3 TCP Packet

An IDS using unsupervised learning can effectively be used to analyse header and data (payload) components of packets and classify them as normal or malicious even though there is no knowledge of the data. The next section describes the steps that are taken for creating a self organised map and how it is used to group together similar data while a multidimensional data set (network traffic data) is given as input.

Application of data mining in computer security: Data mining is a technique which is using historical data to predict the success of a marketing campaign, discovering illegal activities during financial transaction or analyzing genome sequences. Application of data mining has presented a collection of research efforts on the use of data mining in computer security. In the context of security of the information we are seeking the knowledge of whether an information security breach has been experienced. This information can be collected in the context of discovering intrusions that aim to breach the privacy of services, data in a computer system or alternatively in the context of discovering evidence left in a computer system as part of criminal activity. Intrusion detection system is the area where data mining concentrate heavily. There are two reasons for this first an IDS is very common and very popular and extremely critical activity, Second large volume of the data on the network is dealing so this is an ideal condition for using data mining. Data mining applications are designed for the computer security to meet the needs of researcher, industrial practitioner student and professional person. Lee and Salvatore J.Stolfo, Columbia University was the two person who applied the data mining technology first time in the intrusion detection

research area [5]. The data mining process extracts effective, useful, latent, updated, and the understandable pattern from a lot of data. In the intrusion detection system, the important information comes from the host log, the network data package, the system's log data against applications, and alarm messages that other invades the detection system or the system monitors. The data mining technology has the huge advantage in the data extracting characteristic and the rule, so it is of great importance to use data mining technology in the intrusion detection. An important problem of Intrusion Detection is how to effectively separate the normal behavior and the abnormal behavior from a large number of raw data's attributes, and how to effectively generate automatic intrusion rules after collecting raw network data. To accomplish this, various data mining algorithms must be studied, such as correlation analysis, sequence analysis, classification algorithms, and so on.

2. Proposed Work

In this section we are going to present simple model of proposed Intrusion Detection System Using efficient data mining approach. The objective of this research is to provide comparative study of intrusion detection system using various techniques where we will show that our suggested technique will produce better result. The performance and strength of suggested technique is expected to be better than conventional technique of intrusion detection system and highly effective against attack.

Proposed technique: Here we are going to present general idea on a new proposed technique as showing in figure 5 for intrusion detection system which will enhance efficiency as compared to existing intrusion detection system. In the proposed technique the data mining concept is used. 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 we applied K-Medoids data mining for anomaly detection field of intrusion detection.

Whole proposed hybrid IDS is 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-Medoids Cluster Technique
3. Naïve Bayes Classification
   - Naïve Bayes Classification with K-Medoids Cluster Technique
4. Performance
   - Time Analysis
   - Memory Analysis
   - CUP Analysis
   - Probability Analysis

Database Creation (Suggested Technique):

Selecting and generating the data source: First the acquisition of data will be done. In the case of this research, Sample datasets from database will be used. The database containing high volume network traffic data and a subset of data ranging a period of 2 – 5 days will be selected.

Data scope transformation and pre-processing: For the purpose of research, the scope of data is limited to TCP/IP packets. Only six intrinsic features are extracted from each packet within the dataset. These are timestamp, Source IP, Source Port, Destination IP, Destination Port, and Service. The figure 4 below shows the scope of the input dataset.

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

Fig. 4 TCP packet frame format

For the purpose of reporting the data is extracted from the database using data base tools. We extracted the necessary features and saved data within the dataset. Once the data is loaded into the pre-processor it is prepared to be used 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.

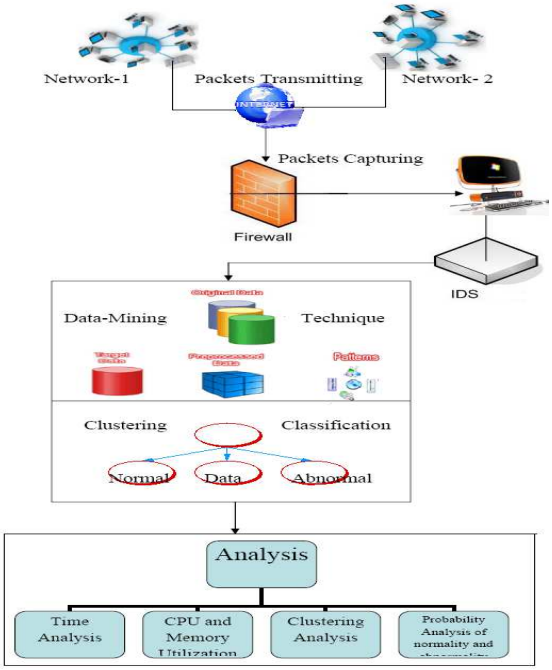

Fig. 5 Proposed Hybrid IDS Architecture

In order to maintain the high accuracy and detection rate while at the same time to lower down the false alarm rate, we propose a combination of two learning techniques. At the first stage of the proposed hybrid technique, we group similar data instances based on their behaviors by utilizing a K-Medoid clustering as a pre-classification component. Next, using Naïve Bayes classifier we classify the resulting clusters into attack classes as a final classification task. We found that data that has been misclassified during the earlier stage may be correctly classified in the subsequent classification stage.

K-Medoids Cluster Technique: Network intrusion labels are divided into four main classes, which are DoS, Probe, U2R, and R2L [1-2]. The main goal to utilize K-Medoid clustering approach is to split and to group data into normal and abnormal class. 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. Let us define j-th variable of object i as X_{ij} (i = 1….n, j= 1…..p).

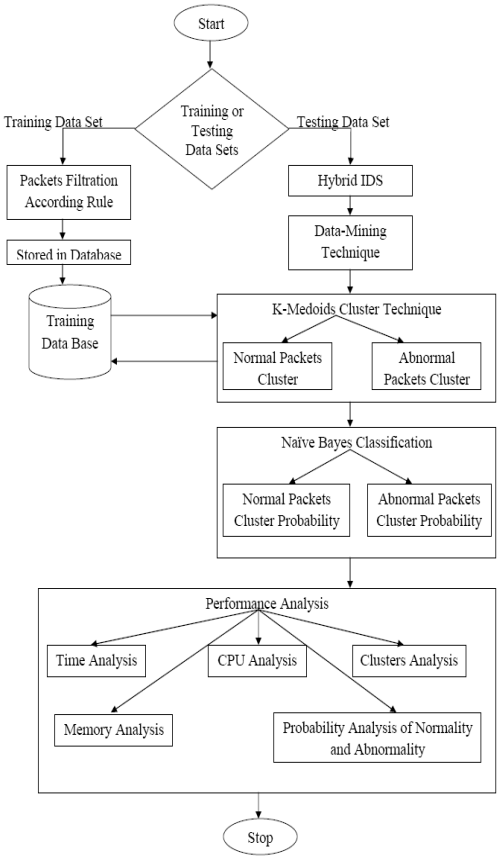

Fig. 6 Flow Chart of the Proposed Hybrid IDS

The proposed 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 all objects as follows:

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

1-2. Calculate P_{ij} to make an initial guess at the centers of the clusters.

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

1-3. Calculate $\sum_{i=1}^{n} p_{ij} \ (j = 1,\ldots,n)$ at each object and sort them in ascending order. Select 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.

In Step 1, we used a method of choosing the initial medoids. The performance of the algorithm may vary according to the method of selecting the initial medoids.

3. Performance Analysis

This section is providing analysis of the proposed system on the basis of different parameters like Normal Packet Probability, Execution Time, Memory Utilization and CPU Utilization. Microsoft Dot Net implementation has been used to test proposed system. For experiment, Intel Pentium Dual Core I3 2.53 GHz, 2 GB of RAM and Windows-7 has been used. For our experiment, we have used different file size ranges from 2 K to 14 K records data set. From the simulation we have analyzed that strength of proposed system is more than the existing one. Even with the little-bit change in the input of the proposed system produces greater change in the output.

Probability of Normal and Abnormal Packets:
Table 2 is showing the probability of normal and abnormal packets in proposed system and existing system to show the less false alarm rate in proposed technique.

Table 2: Comparison of Probability

no. of records in Record Set	*Probability of Packet*			
	Existing System		*Proposed System*	
	Normal	*Abnormal*	*Normal*	*Abnormal*
2000	6.367 E-33	3.406 E-21	6.367 E-33	3.406 E-21
3000	7.341 E-31	1.121 E-21	7.341 E-31	1.121 E-21
5000	0	2.506 E-46	9.650 E-64	2.308 E-46
10000	0	0	1.252 E-77	2.918 E-55
14000	0	0	2.252 E-104	1.594 E-60

Execution Time: - The execution time is the time that an algorithm takes to produce results. Execution time is used to calculate the throughput of an algorithm. It indicates the speed of algorithm. Table 3 is showing the execution time comparison between proposed system and existing system.

Table 3: Comparison of Execution Time

Size of Record Set (no. of records)	*Execution Time (in Second)*	
	Existing	*Proposed System*
2000	01.467	01.467
3000	01.963	01.963
5000	03.734	03.734
10000	07.576	07.576
14000	10.944	10.944

Memory Utilization: - The memory deals with the amount of memory space it takes for the whole process of Intrusion Detection System. Table 4 is showing the memory utilization comparison between proposed system and existing system.

Table 4: Comparison of Memory Utilization

Size of Record Set (no. of records)	*Memory Consumption (MB)*	
	Existing System	*Proposed System*
2000	48	40
3000	44	40
5000	47	41
10000	49	43
14000	52	44

CPU Utilization: - The CPU Utilization is the time that a CPU is committed only to the particular process of calculations. It reflects the load of the CPU. The more CPU time is used in the execution process, the higher is the load of the CPU. Table 5 is showing the CPU utilization comparison between proposed system and existing system.

Strength of the Proposed System:-

- Proposed system gives less false alarm rate.
- Proposed system is faster than existing system in terms of execution time.
- Proposed system is smaller than the existing system and easy to understand and implement.
- It does not contain complex structure. Control flow is well defined and looping structure is

minimized. Due to the above facts it takes very less time for execution.

Table 5: Comparison of CPU Utilization

Size of Record Set (no. of records)	CPU Utilization (in Second)	
	Existing System	Proposed System
2000	03.400	00.748
3000	04.492	00.639
5000	07.878	01.404
10000	15.568	02.464
14000	21.543	03.775

4. Conclusion

In This paper we have improved detecting speed and accuracy which is the prime concern of the proposed work and presented more efficient associate and cluster rules method to detect abnormal packets. Presented Approach is a hybrid approach which is the combination of K-Medoid clustering and Naïve Bayes classification. The proposed approach was compared and evaluated using our own prepared dataset. Considering the dependent relations between alerts, we 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. As compared with other method our method can find the probability from the training data as well as testing data with high efficiency.

Usually when an attack is performed, it is very possible that there exist attack cluster transitions. Based on this we use the cluster sequences to filter false alarms generated by IDS. Experimental results proved this method is effective and feasible.

Future Enhancement: 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 the 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] Skorupka, C., J. Tivel, L. Talbot, D. Debarr, W. Hill, E. Bloedorn, and A. Christiansen 2001. “Surf the Flood: Reducing High-Volume Intrusion Detection Data by Automated Record Aggregation,” Proceedings of the SANS 2001 Technical Conference, Baltimore, MD

[4]http://www.webopedia.com/TERM/I/intrusion_ detection_system.html

[5] LI Min “Application of Data Mining Techniques in Intrusion Detection” 2005

[6] KDD. (1999). Available at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[7] Changxin Song, Ke Ma “Design of Intrusion Detection System Based on Data Mining Algorithm” 2009 IEEE International Conference on Signal Processing Systems

[8] YU-XIN DING, HAI-SEN WANG, QING-WEI LIU “INTRUSION SCENARIOS DETECTION BASED ON DATA MINING” Proceedings of the Seventh IEEE International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008

[9] David Arthur and Sergei Vassilvitsk “How Slow is the kMeans Method?” ACM SCG’06, June 5–7, 2006, Sedona, Arizona, USA.

[10] Sujaa Rani Mohan, E.K. Park, Yijie Han “An Adaptive Intrusion Detection System using a Data Mining Approach” 2004

[11] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko Ruth Silverman, Angela Y. Wu “A Local Search Approximation Algorithm for k-Means Clustering” July 14, 2003 Annual ACM Symposium on Computational Geometry

[12] Eric Bloedorn, Alan D. Christiansen, William Hill “Data Mining for Network Intrusion Detection: How to Get Started” 2001