A Review of Intrusion Detection System Using Fuzzy K-Means and Naive Bayes Classification

Aman Mudgal[1] Rajiv Munjal[2]
[1]CBS Group of Institutions, Fatehpuri, Jhajjar, Haryana

Abstract— Intrusion Detection Systems (IDSs) are proposed to improve computer security because it is not feasible to build completely secure systems. In particular, IDSs are used to identify, assess, and report unauthorized or unapproved network activities so that appropriate actions may be taken to prevent any future damage. Intrusion Detection System is classified on the basis of the source of Data and Model of Intrusion. There are some challenges faced by the Intrusion Detection System. Fuzzy K-Mean and Naïve Bayes classification are the approaches through which the challenges can be overwhelmed. Anomaly in the Anomaly based Intrusion Detection System can be detected using various Anomaly detection techniques. Dimension Reduction can be done using Principle Component Analysis. Support Vector Machine can be used to specify the classifier construction problem. The paper describes the various approaches of Intrusion detection system in briefly.

Key words: Intrusion Detection, Fuzzy K-Mean, Naïve Bays

I. INTRODUCTION

A. *Instruction Detection System:*

Malicious users and crackers seek weak targets such as unpatched systems, systems infected with Trojans, and networks running insecure services The assurance of integrity and safety should be applied to computer systems and data. The Internet has made the information flow to the large extent. Also at the same time it has to face many threats and attacks. Thus the security alert is required to control the attacks and threats. Intrusion Detection Systems (IDSs) are proposed to improve computer security because it is not feasible to build completely secure systems [1]. In particular, IDSs are used to identify, assess, and report unauthorized or unapproved network activities so that appropriate actions may be taken to prevent any future damage [2]. Based on the information sources that they use, IDSs can be categorized into two classes: network-based and host-based. Network intrusion detection systems (NIDSs) analyse network packets captured from a network segment, while host- based intrusion detection systems (HIDSs) such as IDES (Intrusion Detection Expert System) [3] examine audit trails or system calls generated by individual hosts.

B. *Ideal Intrusion Detection System:*

An ideal intrusion detection system [4] should address the following issues, regardless of mechanism it is based on:

(1) The system must run continually without human supervision. It must be reliable enough to allow it to run in the background of the system being observed.
(2) It should not be a "black box". That is, its internal workings should be examinable from outside.
(3) It must be fault tolerant in the sense that it must survive a system crash and not have its knowledge-base rebuilt at restart.
(4) It must resist subversion. The system can monitor itself to ensure that it has not been subverted.
(5) It must impose minimal overhead on the system. A system that slows a computer to a crawl will simply not be used.
(6) It must observe deviations from normal behavior.
(7) It must be easily tailored to the system. Every system has a different usage pattern, and the defense mechanism should adapt easily to these patterns.
(8) It must deal with changing system behavior over time as new applications are being added. The system profile will change over time.
(9) It must be difficult to fool.

All the above listed are the features that an ideal Intrusion Detection System must have. So that the system becomes perfect to defend the attacks and the intrusions.

C. *Working Of Intrusion Detection System:*

The working of the intrusion detection system is quite similar as that of the other programs used to prevent the computer system from dangerous threats like malware, spyware, spam and many more. The job of the intrusion detection system starts from the recording the information about the problem and check the occurrence and the nature of the threat. When the system monitors the problem and collects the data about it, then it sends this information to the administration department of the intrusion detection system which makes several preventive measures to protect the system and keep the system in the safe hands. Intrusion detection system can work in the specific manner by monitoring some important things. These important things are as follows.

(1) Monitoring the activity of the network and activity of the threat in the network.
(2) This system has ability to detect the viruses, malware, spyware and different form of viruses and the important thing about this it can also locate their restore point.
(3) Intrusion detection system can work by observing the unauthenticated and unauthorized use of different programs of networking.

So, the whole working of the intrusion detection system based on the examination of such events of networking.

D. Types of Intrusion Detection System:

IDSs can also be categorized according to the detection approaches they use. Basically, there are two detection methods: misuse detection and anomaly detection. The major deference between the two methods is that misuse detection identifies intrusions based on features of known attacks while anomaly detection analyzes the properties of normal behavior. IDSs that employ both detection methods are called hybrid detection-based IDSs. Examples of hybrid detection-based IDSs are Hybrid NIDS using Random Forests [5] and NIDES [6].

E. Stack Based Intrusion Detection System (SIDS):

Stack based Intrusion Detection System (SIDS) is latest technology, which works by integrating meticulously with the TCP/IP stack, allowing packets to be watched as they traverse their way up the OSI layers. Watching the packet in this way allows the IDS to pull the packet from the stack before the OS or application has a chance to process the packets.

F. Network Based Intrusion Detection System (NIDS):

Network based Intrusion Detection System (NIDS) monitors the traffic as it flows to other host. Monitoring criteria for a specific host in the network can be increased or decreased with relative ease. NIDS should be capable of standing against large amount of network traffic to remain effective. As network traffic increases exponentially NIDS must grab all the traffic and analyze in a timely manner.

G. Host Based Intrusion Detection System (HIDS):

Host based Intrusion Detection System (HIDS) keeps record of the traffic that is originated or is projected to originate on a particular host.. HIDS controls the privileged access of the host to monitor specific components of a host that are not readily accessible to other systems.. HIDS has limited view of entire network topology and they cannot detect attack that is targeted for a host in a network which does not have HIDS installed

H. Anomaly Based Intrusion Detection System:

Anomaly based Intrusion Detection System examines ongoing traffic, activity, transactions and behavior in order to identify intrusions by detecting anomalies. It works on the notion that ―attack behavior‖ differs enough from ―normal user behavior‖ such that it can be detected by cataloging and identifying the differences involved. The system administrator defines the baseline of normal behavior. Anomaly-based IDS systems are very prone to a lot of false positives .Anomaly-based IDS systems can cause heavy processing overheads on the computer system.

I. Signature Based Intrusion Detection System:

Signature based Intrusion Detection System use a set of rule to identify intrusions by watching for patterns of events specific to known and documented attacks. It is typically connected to a large database which stocks attack signatures. These types of systems are able to detect only attacks ―known‖ to its database. Thus, if the database is not updated with regularly, new attacks could slide through. Signature based IDS's affect performance when intrusion patterns match several attack signatures. In such cases, there is a noticeable performance lag. Signature definitions stored in the database need to be specific so that variations on known attacks are not missed. This can lead in building huge databases which eat up a chunk of space.

II. RELATED WORK

A. SK Sharma, P Pandey, SK Tiwar et.al.[6]:

As network attacks have increased in number and severity over the past few years, intrusion detection system (IDS) is increasingly becoming a critical component to secure the network. Due to large volumes of security audit data as well as complex and dynamic properties of intrusion behaviors, optimizing performance of IDS becomes an important open problem that is receiving more and more attention from the research community.

B. M,Varaprsad Rao et.al.[7]:

The k-Means clustering algorithm partition a dataset into meaningful patterns. Intrusion Detection System detects malicious attacks which generally include theft information. modified k-Means by applying preprocessing and normalization steps. As a result the effectiveness is improved and it overcomes the shortcomings of k-Means. This approach is proposed to work on network intrusion data and the algorithm is experimented with KDD99 dataset and found satisfactory results.

C. Zhenglie Li et. al [5]

K-means clustering algorithm is an effective method that has been proved for apply to the intrusion detection system. Particle swarm optimization (PSO) algorithm which is evolutionary computation technology based on swarm intelligence has good global search ability.. proposed algorithm has overcome falling into local minima and has relatively good overall convergence. Experiments on data sets KDD CUP 99 has shown the effectiveness of the proposed method and also shows the method has higher detection rate and lower false detection rate.

D. Thaksen J.Parvat et. al.[8]:

Network attacks are a serious issue in today's network environment. The different network security alert system analyse network log files to detect these attacks. Clustering is useful for wide variety of real time applications dealing with large amount of data. Clustering divides the raw data into clusters. These clusters contain data points which have similarity between themselves and dissimilarity with other cluster data points. If these clusters are given to these security alert systems, they will take less time in analysis as the data will be grouped according to the criteria the security system needs. This can be done by using k means clustering algorithm

E. Ashoor et.al.[9]:

the idea of IDS and its importance to researchers and research centers, security, military and to examine the importance of intrusion detection systems and categories , classifications, and where can put IDS to reduce the risk to the network. Another definition of IDS given by the authors is "The goal of intrusion detection is to monitor network assets to detect anomalous behavior and misuse in network. This concept has been around for nearly twenty years but only recently has it seen a dramatic rise in popularity and incorporation into the overall information security infrastructure".

F. Hamdan et.al.[10]:

explained the process of intrusion detection which is the major part of network activity and security policies adopted over the network to secure it. In this research paper four intrusion detection approaches, which include ANN or Artificial Neural Network, SOM, Fuzzy Logic and SVM have considered for research. ANN which an oldest systems that have been used for Intrusion Detection System (IDS), which presents supervised learning methods is considered in this paper along with SOM or Self Organizing Map, which is an ANN-based system, but applies unsupervised methods. Another approach is Fuzzy Logic (IDS-based), which also applies unsupervised learning methods. The ultimate aim of this research paper is to draw an image hybrid approaches using these supervised and unsupervised methods.

G. Chandola et. al.[11]:

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains. Anomaly detection techniques are very important they also have been significantly developed over the time for certain domains. In this research paper different existing technique are grouped into different categories based on the underlying approach adopted by each technique. For each category they have identified key assumptions, which are used by the techniques to differentiate between normal and anomalous behavior. On applying a given specific technique on a particular domain, these assumptions can be used as procedure to assess the efficiency of the technique in that domain.

H. Barford et al[12]:

presented a framework for detecting and localizing performance anomalies based on using an active probe-enabled measurement infrastructure deployed on the periphery of a network. Their framework has three components: an algorithm for detecting performance anomalies on a path, an algorithm for selecting which paths to probe at a given time in order to detect performance anomalies (where a path is defined as the set of links between two measurement nodes), and an algorithm for identifying the links that are causing an identified anomaly on a path (i.e., localizing). The problem of detecting an anomaly on a path was addressed by comparing probe-based measures of performance characteristics with performance guarantees for the network (e.g., SLAs). The path selection algorithm was designed to enable a tradeoff between ensuring that all links in a network are frequently monitored to detect performance anomalies, while minimizing probing overhead.

I. Ahmed et.al.[13]:

proposed machine learning approach in detecting the anomalies in the network. In this research paper it is explained that Machine learning techniques enables the development of anomaly detection algorithms that are non-parametric, adaptive to changes in the characteristics of normal behaviour in the relevant network, and portable across applications. For this purpose they have used two different datasets, pictures of a highway in Quebec taken by a network of webcams and IP traffic statistics from the Abilene network, as examples in demonstrating the applicability of two machine learning algorithms to network anomaly detection. They investigated the use of the block-based One-Class Neighbour Machine and the recursive Kernel-based Online Anomaly Detection Algorithm

J. Jhang orithms. et. al.[14]:

survey on anomaly detection over the computer network. The reason to include this paper in literature review is that this paper presented anomaly detection in a very systematic way. The authors has included test and train both type of data for the survey. In order to distinguish between the different approaches used for anomaly detection in networks in a structured way, they have classified those methods into four categories: statistical anomaly detection, classifier based anomaly detection, anomaly detection using machine learning and finite state machine anomaly detection. They described each method in details and gave examples for its applications in networks.

III. FUZZY K-MEANS ALGORITHM

The clusters produced by the k-means procedure are sometimes called "hard" or "crisp" clusters, since any feature vector **x** either is or is not a member of a particular cluster. This is in contrast to "soft" or "fuzzy" clusters, in which a feature vector **x** can have a degree of membership in each cluster

- Make initial guesses for the means $\mathbf{m}_1$, $\mathbf{m}_2$,..., $\mathbf{m}_k$
- Until there are no changes in any mean:
- Use the estimated means to find the degree of membership u(j,i) of $\mathbf{x}_j$ in Cluster i;
  for example, if a(j,i) = exp(- $\| \mathbf{x}_j - \mathbf{m}_i \|^2$), one might use u(j,i) = a(j,i) / sum_j a(j,i)
- For i from 1 to k
- Replace $\mathbf{m}_i$ with the fuzzy mean of all of the examples for Cluster i --

$$\mathbf{m}_i = \frac{\sum_j u(j,i)^2 \mathbf{x}_j}{\sum_j u(j,i)^2}$$

end_for

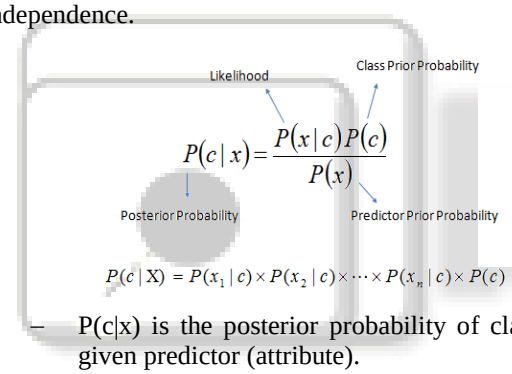
– end_until It has the advantage that it more naturally handles situations in which subclasses are formed by mixing or interpolating between extreme examples, so that it makes more sense to say that **x** is 40% in Cluster 1 and 60% in Cluster 2, rather than having to assign **x** completely to one cluster or the other.

IV. NAÏVE BAYES CLASSIFICATION ALGORITHM

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely use because it often outperforms more sophisticated classification methods.

A. Algorithm:

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.



– P(c|x) is the posterior probability of class (target) given predictor (attribute).
– P(c) is the prior probability of class.
– P(x|c) is the likelihood which is the probability of predictor given class.
– P(x) is the prior probability of predictor.

V. CONCLUSION

In this thesis we proposed a method for classification of intruder in system Intrusion detection is the major task in networking. There are so many solution provided by the researchers for detection of intruder in the network. Like Pattern Matching, Measure Based method, Data Mining method and Machine Learning Method.

Here we detected intrusion through data mining method by combining two data mining technique fuzzy K means and Naive Bayes classification and formed a hybrid technique.

We combined these different methods for measured different aspects of intrusions. Combined these rules find the intruder attack more quickly from the exiting one.

VI. FUTURE ASPECT

In future, an association rule based approach or IF-THEN rules could be effective in classified the traffic in different classes. However accuracy of the algorithms plays an important role to correctly cluster the datasets. Standalone algorithms may not be able to provide efficient results. An another hybrid approach to data clustering can also be applied for analysis and to obtain low inter-cluster similarity.

REFERENCES

[1] LI Yongzhong,YANG Ge,XU Jing Zhao Bo "A new intrusion detection method based on Fuzzy HMM "IEEE Volume 2, Issue 8, November 2008.
[2] Tarem Ahmed, Boris Oreshkin and Mark Coates, Department of Electrical and Computer Engineering McGill University Montreal, QC, Canada "Machine Learning Approaches to Network Anomaly Detection" in Workshop on Tackling Computer Systems Problems with Machine Learning Techniques, 2007
[3] Li Tian, "Research on Network Intrusion Detection System Based on Improved K-means Clustering Algorithm", Computer Science-Technology and Applications, 2009. IFCSTA '09. International Forum,
[4]http://www.cerias.purdue.edu/about/history/coast_reso u rces/idcontent/detection.html.
[5] Zhenglie Li"Anomaly Intrusion Detection Method Based on K-Means Clustering Algorithm with Particle Swarm Optimization "Springer Volume 4, Issue 2, April 2011.
[6] SK Sharma, P Pandey, SK Tiwar "An improved network intrusion detection technique based on k-means clustering via Naïve bayes classification" IEEE Volume 2, Issue 2, February 2012, Issn 2151-961.
[7] M,Varaprsad Rao "Algorithm for Clustering with Intrusion Detection Using Modified and Hashed K – Means Algorithms "Published by IEEE Computer Society,2012.
[8] Thaksen J.Parvat" Network Log Clustering Using K-Means Algorithm'In IEEE Pasfic asia workshop of networking 2011.
[9] Asmaa Shaker Ashoor (Department computer science, Pune University) Prof. Sharad Gore (Head department statistic, Pune University), "Importance of Intrusion Detection System (IDS)", International Journal of Scientific & Engineering Research, Volume 2, Issue 1, January-2011 ISSN 2229-5518.
[10] Hamdan.O.Alanazi, Rafidah Md Noor, B.B Zaidan, A.A Zaidan, "Intrusion Detection System: Overview "Journal Of Computing, Volume 2, Issue 2, February 2010, Issn 2151-961
[11] Varun Chandola University Of Minnesota Arindam Banerjee University Of Minnesota And Vipin Kumar University Of Minnesota "Anomaly Detection : A Survey", ACM Computing Surveys, September 2009.
[12] Paul Barford University of Wisconsin, Nick Duffield AT&T, Amos Ron University and Joel Sommers Colgate, "Network Performance Anomaly Detection and Localization" Infocom 2009.
[13] Tarem Ahmed, Boris Oreshkin and Mark Coates, Department of Electrical and Computer Engineering

McGill University Montreal, QC, Canada "Machine Learning Approaches to Network Anomaly Detection" in Workshop on Tackling Computer Systems Problems with Machine Learning Techniques, 2007

[14] Weiyu Zhang; Qingbo Yang; Yushui Geng, "A Survey of Anomaly Detection Methods in Networks", Computer Network and Multimedia Technology, 2009. CNMT 2009. International Symposium

