Volume 3, Issue 3, March 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey of Network Intrusion Detection Using K-Mean Algorithm

Poonam Dabas*
Assistant Professor
Department Of Computer Sc &Engg.
Kurukshetra Universty, India.

Rashmi Chaudhary
M. Tech UIET
Department Of Computer Sc &Engg.
Kurukshetra Universty, India.

Abstract— Intrusion Detection System (IDS) due to novel attack method or upgraded. Because many current IDSs are constructing by point instruction of professional knowledge, changes to IDSs are costly and slow. Intrusion detection techniques can be categorize into irregularity detection and mistreat detection. Anomaly detection systems, for example, IDES Intrusion detection systems (IDS) process large amount of monitoring data. As an illustration, host-based IDS examine log les on a computer (or host) in classify to identify suspicious activities. Network-based IDS, on the other hand, searches network monitor data for harmful packets or packet flows. In the behind 1990s, development in data mining research and the essential to find recovered methods for network and host based intrusion detection resulted in research activities attempting to organize data mining techniques for anomaly and attack detection. IDS may be a straightforward assessment trail process, or a filter process using a traffic control system, like screening routers, packet filters, firewalls, etc. It is an major technology in business sector as well as an dynamic area of research. In Information Security, intrusion detection is the proceed of detecting proceedings that attempt to cooperate the confidentiality, integrity or availability of a resource. It acting a particularly significant role in attack detection, security check and network inspect.

Keywords— Instruction Detection System, K-Mean Algorithm, Data Clustering, Machine Learning.

I. INTRODUCTION

Instruction Detection System

Intrusion Detection Systems (IDSs) are proposed to improve computer security because it is not feasible to build completely secure systems [12]. 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 [9]. 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) [11] examine audit trails or system calls generated by individual hosts.

1.1Types of IDS

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 [3] and NIDES [1]. The following subsections explain the two detection approaches.

1.1.1 Misuse Detection

Misuse detection catches intrusion in terms of the characteristics of known attacks. Any action that conforms to the pattern of a known attack or vulnerability is considered as intrusive. The main issues in misuse detection system are how to write a signature that encompasses all possible variations of the pertinent attack. And how to write signatures that do not also match non-intrusive activity. Block diagram of misuse based detection system is as following.

Misuse detection identifies intrusions by matching monitored events to patterns or signatures of attacks. The attack signatures are the characteristics associated with successful known attacks The major advantage of misuse detection is that the method possesses high accuracy in detecting known attacks. However, its detection ability is limited by the signature database. Unless new attacks are transformed into signatures and added to the database, misuse-based IDS cannot detect any attack of this type. Deferent techniques such as expert systems, signature analysis, and state transition analysis are utilized in misuse detection.

1.1.2 Anomaly Detection System

It is based on the normal behavior of a subject (e.g. a user or a system). Any action that significantly deviates from the normal behavior is considered as intrusive. That means if we could establish a normal activity profile for a system, then we can flag all system states varying from established profile. There is a important difference between

anomaly based and misuse based technique that the anomaly based try to detect the compliment of bad behavior and misuse based detection system try to recognize the known bad behavior. In this case we have two possibilities:

(1)False positive: Anomalous activities that are not intrusive but are flagged as intrusive.

(2) False Negative: Anomalous activities that are intrusive but are flagged as non intrusive.

The block diagram of anomaly detection system is as following:

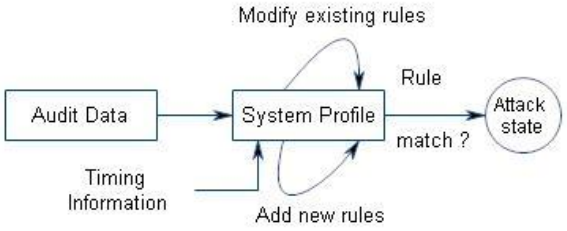

Fig 1.1 Misuse Detection Systems

Fig 1.2 Anomaly Detection Systems

Anomaly detection assumes that intrusions are anomalies that necessarily differ from normal behavior. Basically, anomaly detection establishes a profile for normal operation and marks the activities that deviate significantly from the profile as attacks. The main advantage of anomaly detection is that it can detect unknown attacks However, this advantage is paid for in terms of a high false positive rate because, in practice, anomalies are not necessarily intrusive. Moreover, anomaly detection cannot detect the attacks that do not obviously deviate from normal activities. As the number of new attacks increases rapidly, it is hard for a misuse detection approach to maintain a high detection rate. In addition, modeling attacks is a highly qualified and time- consuming job that leads to a heavy workload of maintaining the signature database [25]. On the other hand, anomaly detection methods that discover the intrusions through heuristic learning are relatively easy to maintain.

II. Related Work

SK Sharma, P Pandey, SK Tiwar et.al.[1] 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.

M,Varaprsad Rao et.al.[2] 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.

Zhenglie Li et. al [3] 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.

Thaksen J.Parvat et. al.[4] 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

Ashoor et.al.[5] 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".

Hamdan et.al.[6] 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.

Chandola et. al.[7] 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.

Barford et al[8]. 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.

Ahmed et.al.[9] 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

Jhang orithms. et. al.[10] 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. K-mean Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.

Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. The k-means algorithm can be run multiple times to reduce this effect.

For Example

Suppose that we have n sample feature vectors $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ all from the same class, and we know that they fall into k compact clusters, k < n. Let $\mathbf{m}_i$ be the mean of the vectors in cluster i. If the clusters are well separated, we can use a minimum-distance classifier to separate them. That is, we can say that $\mathbf{x}$ is in cluster i if $\| \mathbf{x} - \mathbf{m}_i \|$ is the minimum of all the k distances. This suggests the following procedure for finding the k means:

- 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 classify the samples into clusters
    - For i from 1 to k

Replace $\mathbf{m}_i$ with the mean of all of the samples for cluster i

- end_for
- end_until

Here is an example showing how the means $\mathbf{m}_1$ and $\mathbf{m}_2$ move into the centers of two clusters.

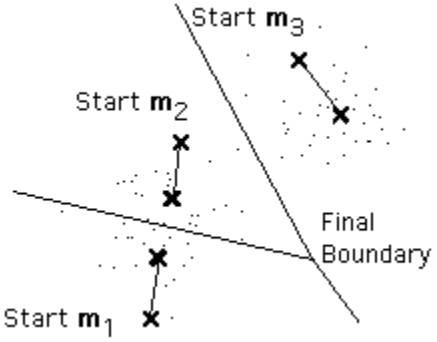

Fig-3.1Running K-mean Clustering

IV. CONCLUSIONS

Intrusion detection systems (IDSs) play an important role in computer security. IDS users relying on the IDS to protect their computers and networks demand that an IDS provides reliable and continuous detection service. However, many of the today's anomaly detection methods generate high false positives and negatives.

All anomaly-based intrusion detection systems work on the assumption that normal activities differ from the abnormal activities (intrusions) substantially. In the case of IDS models that learn a program's behavior, these deference may manifest in the form of (a) the frequency of system calls (Src_bytes, Dst_bytes), and (b) the duration of system calls used by the processes under normal and abnormal execution.

V. Future Work

To improve the usability of the IDS, the future work can be done as follows : more checking rules can be developed and implemented for the AA to improve its ability to detect compromised components; host-based IDSs can be introduced to the IDS and used to monitor the hosts and the IDS components ; an intelligent system can be employed to analyze the intrusion alerts generated by the k-mean based IDS and aid the intrusion-tolerant mechanism in treating. the compromised components; and finally, a response mechanism can be introduced in order to stop intrusions before a failure occurs.

REFERENCES

[1] 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.

[2] M,Varaprsad Rao **"Algorithm for Clustering with Intrusion Detection Using Modified and Hashed K – Means Algorithms** "Published by IEEE Computer Society,2012

[3] Zhenglie Li"Anomaly **Intrusion Detection Method Based on K-Means Clustering Algorithm with Particle Swarm Optimization "**Springer Volume 4, Issue 2, April 2011.

[4] Thaksen J.Parvat" Network Log Clustering Using K-Means Algorithm'In IEEE Pasfic asia workshop of networking 2011.

[5] 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.

[6] 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

[7] 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.

[8] Paul Barford University of Wisconsin, Nick Duffield AT&T, Amos Ron University and Joel Sommers Colgate, "**Network Performance Anomaly Detection and Localization**" Infocom 2009.

[9] 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

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

[11] 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

[12] LI Yongzhong,YANG Ge,XU Jing Zhao Bo "A new intrusion detection method based on Fuzzy HMM "IEEE Volume 2, Issue 8, November 2008.

[13] Kurutach.W "**Combination Artificial Ant Clustering and K-PSO Clustering Approach to Network Security Model** "Published by IEEE Computer Society,2006.

[14] Yau.l "Evaluation **of Fuzzy K-Means And K-Means Clustering Algorithms In Intrusion Detection Systems**" IEEE TRANS. INF. & SYST., Vol. E84-D, No. 5, 570-577 2006.

