Survey of IDS Approach using Data mining tool WEKA

1st Aakshi Choudhary 2nd Sarbjit Kaur Department of Computer Sc. . MIET Mohri Kurukshetra India

Abstract— **Two common data mining techniques for finding hidden patterns in data are clustering and classification analyses. Classification is supposed to be supervised learning and clustering is an unsupervised classification with no predefined classes. Clustering tries to group a set of objects and find whether there is some relationship between those objects. In this paper we have used the numerical results generated through the Probability Density Function algorithm as the basis of recommendations in favor of the K-means clustering for weather-related predictions.**

***Index Terms*—IDS, K-mean Clustring,Attack**

I. INTRODUCTION

Information technology has become a key component to support critical infrastructure services in various sectors of our society. In effort to share information and streamline operations, organizations are creating complex networked systems and opening their networks to customers, suppliers, and other business partners. While most users of these networks are legitimate users, an open network exposes the network to illegitimate access and use. Increased network complexity, greater access, and a growing emphasis on the internet have made network security a major concern for organizations. The number of computer security breaches has risen significantly in the last three years. (Computer Security is the ability to protect a computer system and its resources with respect to confidentiality, integrity, and availability.)

Various protocols, firewalls are in existence to protect these systems from computer threats. While these traditional approaches to network security have focused on prevention, network intrusion detection has become increasingly important in recent years to enable firms to reduce undetected intrusion. Intrusion is a type of cyber attack that attempts to bypass the security mechanism of a computer system. Such an attacker can be an outsider who attempts to access the system, or an insider who attempts to gain and misuse non-authorized privileges.

IDS are now becoming important part of our security system, and its credibility also adds value to the whole system. Data mining techniques can be applied to gain insightful knowledge of intrusion prevention mechanisms. They can help detect new vulnerabilities and intrusions, discover previous unknown

patterns of attacker behaviors, and provide decision support for intrusion management. Intrusion detection is detection of intrusion behavior, it collects information of the key part of computer network and system, then analyzes them to detect whether occur the action of disobey security strategy. Intrusion Detection System (IDS) is the software or combination of software and hardware to detect intrusion behavior. IDS can examine intrusion attack before system is damaged, and make use of alerting and defense system to deport the intrusion attack. In the process of intrusion attack, It can reduce the loss resulted in. After system attacked, the related attack information is collected, and as security system knowledge, it is added to the strategy set, thus can strengthen system security defence ability, avoid system being intruded by the same intrusion again

II .COMPONENTS OF IDS

Sensors which generate security events, a **Console** to monitor events and alerts and control the sensors, and a **central Engine** that records events logged by the sensors in a database and use a system of rules to generate alerts from security events received. There are several ways to categorize IDS depending on the type and location of the sensors and the methodology used by the engine to generate alerts. In many simple IDS implementations all three components are combined in a single device or appliance. Intrusion detection can allow for the prevention of certainty, attacks severity relative to different type of attack and vulnerability of components under attack the response may be kill the connection, install filtering rules, and disable user account

III. TYPES OF ATTACK

The simulated attacks fall in one of thefollowing four categories:

- Denial of Service Attack (DoS): is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.
- User to Root Attack (U2R): is a class of exploit in which the attacker starts out with access to a normal

user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.

- Remote to Local Attack (R2L): occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.
- Probing Attack: is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

The basic steps involved in identifying intrusion using clustering technique are:

- Find the largest cluster, i.e., the one with the most number of instances, and label it normal;
- Sort the remaining clusters in an ascending order of their distance to the largest cluster;
- Select the first K1 clusters so that the number of data instances in these clusters sum up to ¼ ´N, and label them as normal, where ´ is the percentage of normal instances ;
- Label all the other clusters as attacks.Unlike traditional anomaly detection methods, they cluster data instances that contain both normal behaviors and attacks, using a modified incremental k -means algorithm. After clustering, heuristics are used to automatically label each cluster as either normal or attacks. The self-labeled clusters are then used to detect attacks in a separate test dataset.

IV. RELATED WORK

Chandola et al have tried to explain anomaly and its detection methods. Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. Hence 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. In 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. For each category, they provided a fundamental anomaly detection technique, and then demonstrate how the different existing techniques in that category are variants of the basic technique. This template provides an easier and concise understanding of the techniques belonging to each category. Further, for each category, they identified the pros and cons of the techniques in that category. They also provided a discussion on the computational complexity of the techniques since it is an important issue in real application domains.

Barford et al had written a research paper on network performance in which they considered first the localization of the anomaly which is very important. It is critical to detect as the effective operations over the network are influenced by anomaly localization. High jitter and loss events are also plays a critical role in it. In this paper they 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. The localization algorithm was designed to use existing path measurement data in such a way as to minimize the number of paths necessary for additional probing in order to identify the link responsible for an observed performance anomaly. Their results showed that the method was able to accurately detect and localize performance anomalies in a timely fashion and with lower probe and computational overheads than previously proposed methodologies. Authors have also given some important related work on detecting and localizing network performance. Which is as: Detecting and localizing network performance anomalies is an important problem for operational networks and has received substantial attention in prior work. Past efforts have largely focused on using passively collected data to detect, and possibly locate, performance anomalies [6]. For example, Huang et al. considered the problem of detecting general performance problems in a network using only passive packet measurements [7]

Ahmed et al 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 behavior 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 algorithms. The outcome of this paper indicates the area where machine learning approach can be used to detect the anomalies. To make the algorithms portable to different applications and robust to diverse operating environments, all parameters must be learned and autonomously set from arriving data. The algorithms must be capable of running in real-time despite being presented with large volumes of high-dimensional, noisy, distributed data. This means that the methods must perform chronological calculations with the complexity at each timestamp being independent of time. In this paper The OCNM algorithm proposed by Munoz and Moguerza provides an elegant means for estimating minimum volume sets [12],[13]. It assumes a sample set S comprising T, F dimensional data points.

Jhang et al has presented a 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.

Li et al has explained the issue of network security with the development of computer technology. As the growing number of network security threats and the current intrusion detection system development, this paper gave a new model of anomaly intrusion detection based on clustering algorithm. Because of the k-means algorithm's shortcomings about dependence and complexity, the paper putted forward an improved clustering algorithm through studying on the traditional means clustering algorithm. Clustering being a very powerful tool of data mining used very well in this research paper to get the graphs of normal data with context to anomalies. The new algorithm learns the strong points from the k-medoids and improved relations trilateral triangle theorem. The experiments proved that the new algorithm could improve accuracy of data classification and detection efficiency significantly. With the improved trilateral triangle theorem data interpretation become more easy and understandable. The results showed that this algorithm achieves the desired objectives with a high revealing rate.

V. K-MEAN CLUSTRING USING INTRUSION DETECTION

K-means algorithm is a classical clustering algorithm. Its aim is to divide data into k clusters, and ensures that the data within same cluster has high similarity; the data in different cluster has low similarity.

K-means algorithm first select K data at random as initial cluster center, for the rest data, add it to the cluster with highest similarity according to its distance to cluster center; then recalculate the cluster center of each cluster. Repeat this process until each cluster center doesn't change. Thus data is divided into K clusters.

K-means algorithm is very simple, it is suitable for large scale data set. But K-means algorithm is sensitive to initial value and sequence of data object, different initial data may lead to different clustering result. The purpose of the proposed approach is to perform a clustering analysis on a set of tested connections through K means, and then compute the distribution of false alerts in these clusters. This operation is repeated several times, with a different number of clusters each time, until obtaining a final configuration of clusters where each cluster is ideally highly representative of false alerts (the percentage of false alerts is high), or it is highly representative of real attacks (the percentage of false alerts is low).

The following algorithm, which is called Clustering K-mean algorithm here, is an improved algorithm to K- means In KD algorithm, suppose that a constant R stands for clustering radius threshold, C stands for cluster, dist(C,D) stands for the distance from the center of clustering C to vector D, KD clustering algorithm is following

Let $X = \{x_1, x_2, x_3, \ldots\ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center using:
$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$
Where, *'c_i'* represents the number of data points in i^{th} cluster.
5) Recalculate the distance between each data point and new

obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

V1.CONCLUSION

Intrusion Detection System (IDS) plays an effective way to achieve higher security in detecting malicious activities for a couple of years. Anomaly detection is one of intrusion detection system. Current anomaly detection is often associated with high false alarm with moderate accuracy and detection rates when it's unable to detect all types of attacks correctly.

REFERENCES

[1]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.

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

[3] 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

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

[5] 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.

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

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

[8] 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.

[9] Sophia Kaplantzis "**STUDY ON CLASSIFICATION TECHNIQUES FOR NETWORK INTRUSION DETECTION**" IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 51, NO. 8, AUGUST 2003.

