International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 20-25

Intrusion Detection Using Data Mining Techniques

G. Mageswary[1], Dr. M. Karthikeyan[2]
[12]*Assistant Professor/Programmer*
Department of Computer and Information Science, Annamalaiuniversity, India.

Abstract: *Intrusion detection is used to detect attacks against a computer system. It is an important technology in business sector as well as an active area of research. In Information Security, intrusion detection is the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resource. Various Data mining techniques such as clustering and classification are used to detecting intrusions in the network. Data mining clustering is the process of grouping the content into one or more categories. A K-means clusteringalgorithm is a data mining technique based on this, the network data can be categorize into either normal or abnormal. In this paper K-means algorithm is used for detecting the normal or Denial of Service(DOS) attack category. The KDD Cup 1999data set is used for evaluate the proposed model.*

Keywords: *Data mining, Denial of Service, Intrusion Detection, KDD Cup 1999, K-means.*

I. Introduction

Intrusion Detection Systems (IDSs) are proposed to improve computer security because it is not feasible to build completely secure systems[1]. In particular, the 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. 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) analyze network packets captured from a network segment[2]. It examine audit trails or system calls generated by individual hosts. The TCP dumpdata into connections that contain context informationof network sessions. As the volume of network traffic increases, many NIDSs employ multiplesensors and distributed computing to improve their processing capability. NIDSs can also detectIP-based attacks such as denial- of-service attacks which involve multiple computers. A host- based IDS has difficulty detecting these attacks since it monitors only information gathered fromthe computer system. NIDS is gaining popularity since more and more systems are connectingover networks.IDSs can also be categorized according to the detection approaches they use. Basically, thereare two detection methods: misuse detection and anomaly detection. The major deference between the two methods is that misuse detection identifies intrusions based on features ofknown attacks while anomaly detection analyzes the properties of normal behavior.

1.1. Misuse detection

Misuse detection catches intrusion in terms of the characteristics of known attacks. This technique looks patterns and signatures of already known attack in the network traffic. A constantly updated database is usually used to store the signatures of known attacks. Any action that conforms to the pattern of a known attack or vulnerability is considered as intrusive. The Fig.1 shows the block diagram of misuse detection system as follows

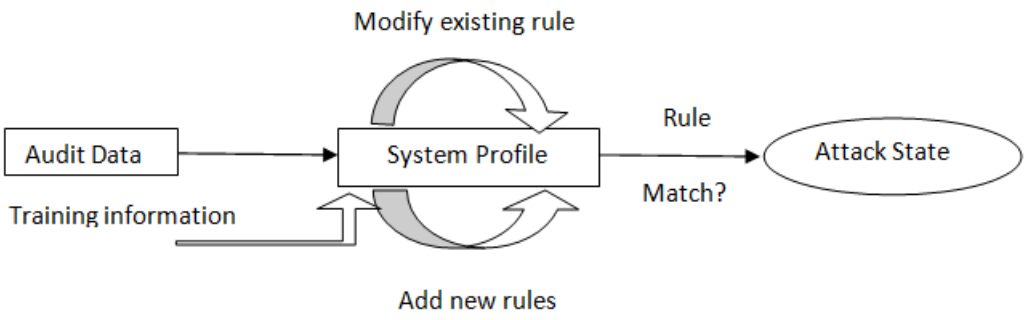

Fig.1 Misuse based detection system

1.2. Anomaly detection

This technique is used based on the detection of traffic anomalies. The deviation of the monitored traffic from the normal profile is measured. Various different implementations of this this technique has been proposed, based on the metrics used for measuring traffic profile deviation. The Fig.2 shows the block diagram of anomaly detection system as follows

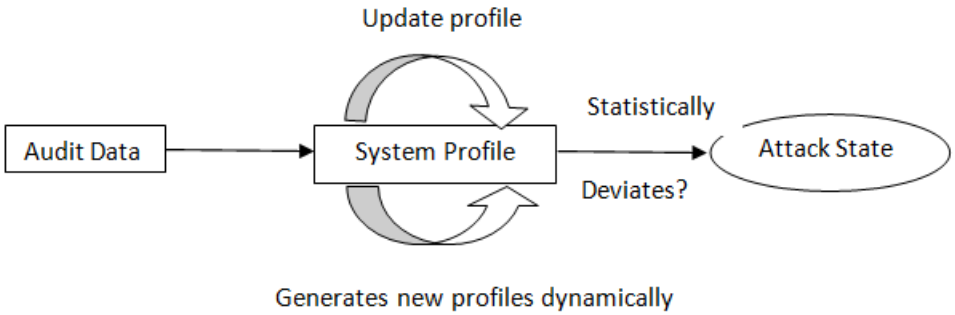

Fig.2 Anomaly based detection system

This paper is organized as follows. In Section 2, we present the review of related literature. In Section 3, the proposed k-means algorithm and the network intrusion detectionare expounded. In Section 4, the experimental results are discussed. In Section 5, concludes this paper.

II. Literature Review

In this section, we review instruction about the intrusion detection system by AurobindoSundaram et al.in Crossroads, 2(4):3–7, 1996[3]. Dr. D.Aruna kumara et al. described the data mining technique for intrusion detection,like clustering and classification. In 2015 WathiqLaftah Al-Yaseen et al. aims to design a model that deals with real intrusion detection problems in data analysis and classify network data into normal and abnormal behaviors.[4]. In 2017 ShadiAljawarneh et al. developed a new hybrid model that can be used to estimate the intrusion scope threshold degree based on the network transaction data's optimal features that were made available for training.[5]. Theodoros and KonstantinosPelechrinis mostly focused on the data mining techniques that are being used for prevension of network intrusion. Gerhard Munz use data mining techniques make it possible to search large amounts of data for characteristic rules and patterns and gives an introduction to Network Data Mining, i.e. the application of data mining methods to packet and how data captured in a network, and present novel flow-based anomaly detection scheme based on the K-mean clustering algorithm [6].

III. Proposed Method

Here the proposed algorithm used for the intrusion detection system and details about the networkintrusion activities.

3.1. K-Means algorithm

K-meansalgorithm is one of the simplest unsupervised clustering algorithms. 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 group age is done. At this point we need to re-calculate k new centroids 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 is as follows.

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

Where $\|x_i^{(j)} - c_j\|^2$ 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.
4. 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.

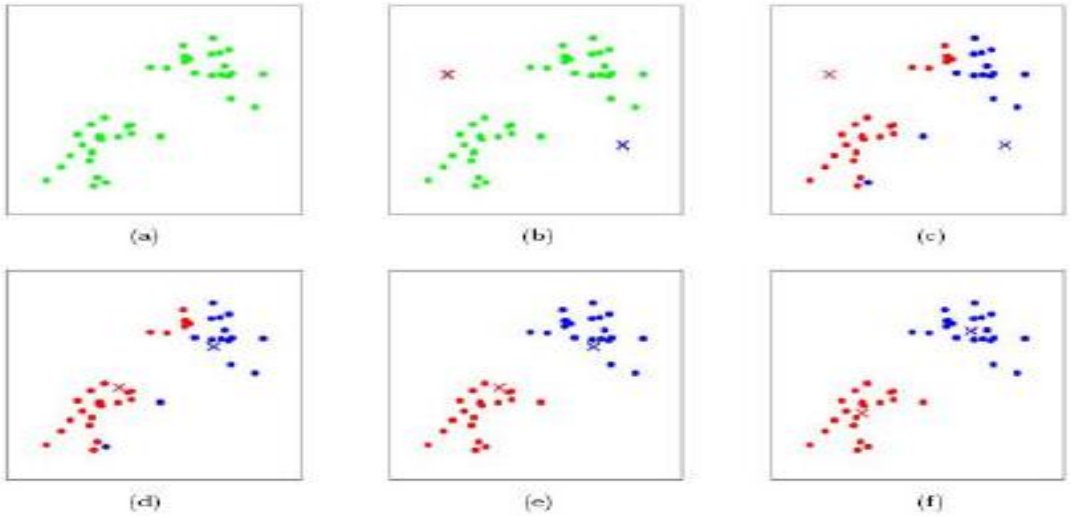

Fig.3 Running k-means clustering.

In Fig.3 the K-means algorithm. Training examples are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) Random initial cluster centroids (in this instance, not chosen to be equal to two training examples). (c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid (shown by "painting" the training examples the same color as the cluster centroid to which is assigned); then we move each cluster centroid to the mean of the points assigned to it.

3.2. Data clustering K-mean Intrusion Detection Algorithm

The Fig.4 shows the four different relations that an attack can have with other attacks and normal behaviors.

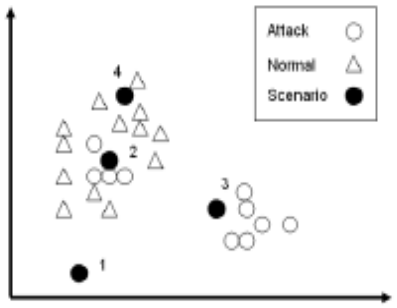

Fig.4 Attack relation scenarios

In Fig. 4 the triangles represent normal behaviors mapped in feature space, while the circles represent attacks. The solid circles show the four relation scenarios of an attack. It observes that the algorithm can only detect attacks in Scenario 1 and 3, because they deviate significantly from the normal set. And it can identify attacks in Scenario 2 and 3, since they are much closer to the attack samples than to the normal samples.

Unfortunately, Scenario 4 is difficult to detect for either measure since it neither deviate significantly from the normal data nor stays close to any reference attack samples.
The pseudo code for the adapted K-mean algorithm is presented as below:
/**********start of pseudo code
1. Choose random k data points as initial Clusters Mean (cluster center).
2. Repeat
3. for each data point x from D
a. Computer the distance x and each cluster mean (centroid)
b. Assign x to the nearest cluster.
4. End for
5. Re-compute the mean for current cluster collections.
6. Until reaching stable cluster.
7. Use these centroid for normal and anomaly traffic.
8. Calculate distance of centroid from normal and anomaly centroid points.
9. If distance(X, Dj) > = 5
a. Then anomaly found ; exit
10. If distance(X, Dj) > = 5
a. X is normal;
*/ end of pseudo code.

3.3 Attacks in the data set

Here we are using the KDD CUP 1999 data set. Each connection was labeled as normal or as specific kind of attack. The attacks fell in exactly one of the four categories : User to Root; Remote to Local; Denial of Service; and Probe. In this experiment we are going to concentrate on DOS attack.

Denial of Service (DOS) - Attacker makes the resource unavailable to its intended user.
Remote to Local (r2l)- Unauthorized access from a remote machine. (guessing pass word)
User to Root (u2r) - Unauthorized access to local super user privileges.
Probe- Attacker scans the network to know the information about the target host.

3.3.1 Different types of DOS Attacks

Smurf Attack The attacker sends a ping request to a broadcast address at a third-party on the network. This ping request is spoofed to appear to come from the victims network address. Every system within the broadcast domain of the third-party will then send ping responses to the victim.
Neptune attack There are two typical threats to web servers.
Teardrop attack It involves sending mangled IP fragments with overlapping, over-sized payloads to the target machine. This can crash various operating systems due to a bug in their TCP/IP fragmentation re-assembly code.
A ping of death (POD) POD is malformed or otherwise malicious ping to a computer A ping is normally 56 bytes in size (or 84 bytes when IP header is considered); historically, many computer systems could not handle a ping packet larger than the maximum IP packet size, which is 65,535 bytes. Sending a ping of this size could crash the target computer.

Back attack: It involves sending forged requests of some type to a very large number of computers that will reply to the requests. Using Internet protocol spoofing, the source address is set to that of the targeted victim, which means all the replies will go to (and flood) the target.

IV. Experimental Results

The experiments and results for the k-mean clustering based intrusion detection technique First the dataset used for the evaluation is described and finally the experimental results are presented. Using k-mean clustering technique centroid of normal and anomaly network flow are calculated. Similarly centroid for normal and DOS attack traffic (Smurf, Neptune, Back, Teardrop, POD) are calculated. The test data without label again centroid calculated then distance is compared with the training and tested data using Euclidean Distance in equation (2)

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \qquad (2)$$

In order to ensure that all possible normal program behaviors are included, a large training data set is preferred for anomaly detection. K-mean algorithm is applied on test data using Weka 3.6.2 and centroid are calculated

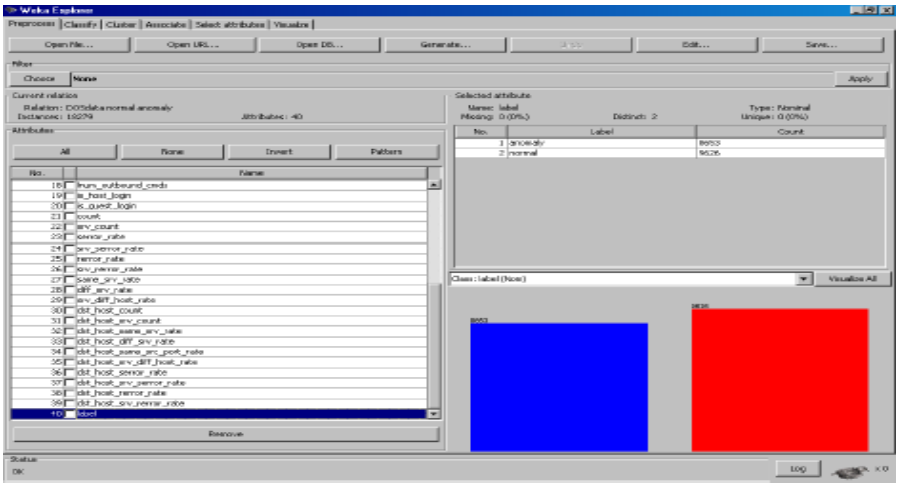

Fig.5 Weka Explorer with dataset showing normal and anomaly dataset

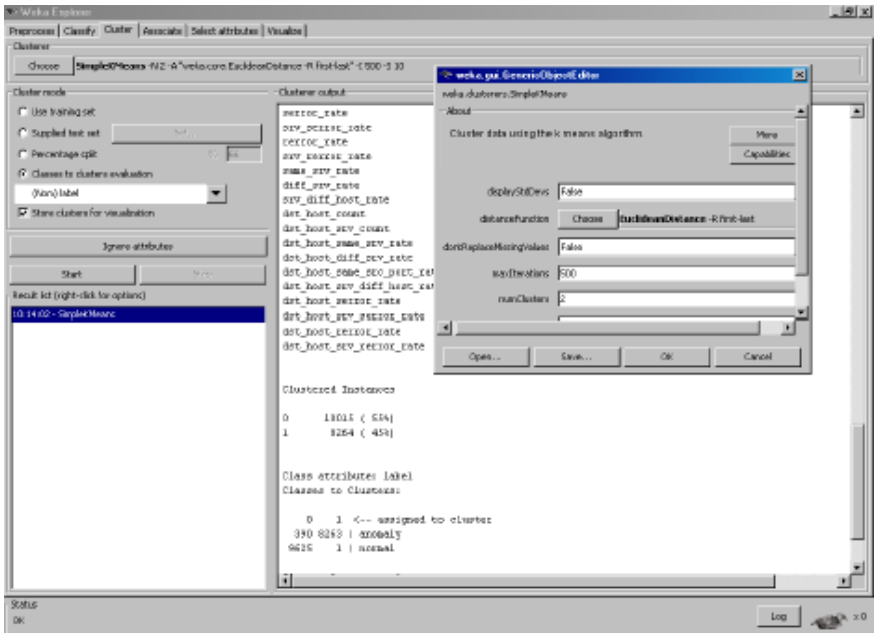

Fig.6 Experiments were performed with k-means algorithm.

The Fig. 7 shows the graph for the various attacks that are available in the Denial of service attack and also the normal data.

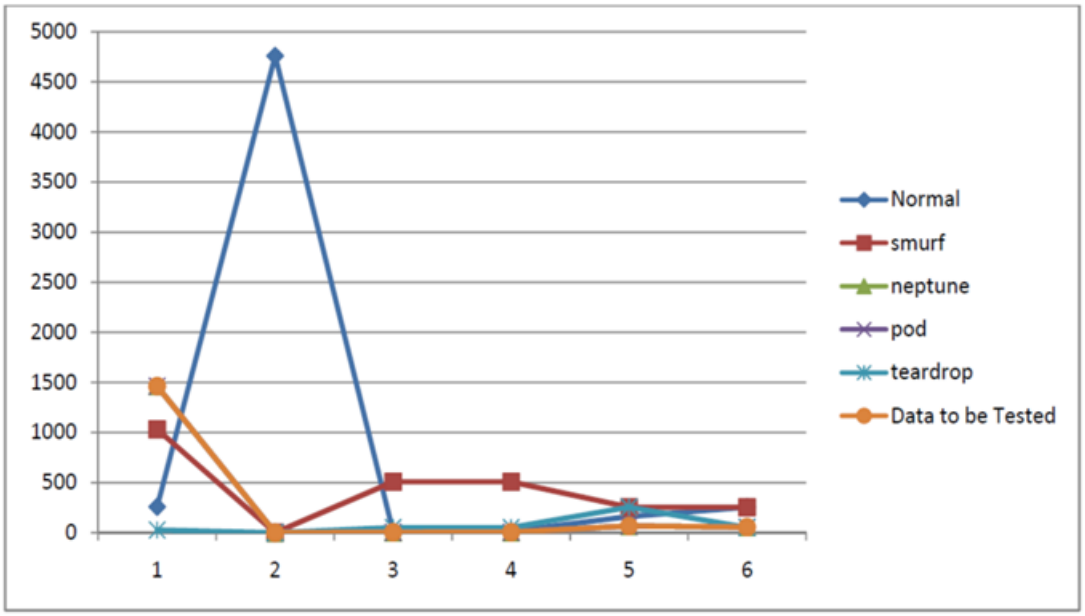

Fig.7 Anomaly found in tested data

V. Conclusion

Intrusion detection play a vital role in computer network. Intrusion Detection System used to protect the 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. The k-means clustering algorithm is used for analyzing the network data and intrusion detection is evaluated. The KDD Cup 1999 data set were used for experiments. Compared to other methods using normal and anomaly based on TCP attributes by using k-means clustering algorithm the intrusion category was determined. Experimental results shows the low false positive rate can be achieved.

References

[1]. AurobindoSundaram. "An introduction to intrusion detection". Crossroads, 2(4):3–7, 1996.
[2]. MithcellRowton, Introduction to network security intrusion detection, December 2005.
[3]. Dr. D. Aruna kumara, Ntejeswani, G, Saravani, R. Phani Krishna., "Intrusion detection using data mining technique, International journal of Computer Science and Information Technology, Vol. 6 (2) 2015.
[4]. WathiqLaftah Al-Yaseen,Zulaiah Ali Othman, MohdZakree Ahmad Nazri , "Multi-Level Hybrid Support Vector Machine and Extreme Learing Machine Based on Modified K-Means for Intrusion Detection System", International journal of Experts systems with applications, Sepetember 2016.
[5]. ShadiAljawarneh, MontherAldwairi, MuneerBaniYasin "Anomaly –based intrusion detection system though feature selection analysis and building hybrid efficient model", journal of Computational science, March 2017.
[6]. Theodoros and KonstantinosPelechrinis, "Data mining Techniques for (network) intrusion detection system".
[7]. Gerhard Munz, Sa Li: Traffic Anomaly Detection Using K-Means Clustering, Computer Networks and Internet, Wilhelm Schickard Institute for Computer Science, University of Tuebingen, September 2007.
[8]. ChaoukiKhammassi, SaoussenKrichen, "A GA_LR Wrapper Approach for feature selection in Network Intrusion Detection, journal of Computer & Security, June 2017.
[9]. Weka data mining tool, http://dataminingtools.net/wiki/weka.php