Scientific Journal Impact Factor (SJIF): 1.711

e-ISSN: 2349-9745
p-ISSN: 2393-8161

International Journal of Modern Trends in Engineering and Research
www.ijmter.com

DETECTING NETWORK ANOMALIES USING CUSUM and FCM

E.M. Roopa Devi[1], R.C.Suganthe[2] , B.Bhuvaneswari3

[1,2,3] Department of Information Technology, kongu Engineering College

Abstract - The network intrusion detection techniques are important to prevent our systems and networks from malicious behaviors. However, traditional network intrusion prevention such as firewalls, user authentication and data encryption have failed to completely protect networks and systems from the increasing and sophisticated attacks and malwares. Two anomaly detection techniques – CUSUM and clustering are used to find network anomalies. CUSUM detect changes based on the cumulative effect of the changes made in the random sequence instead of using a single threshold to check every variable. It involves calculating cumulative sum and determining whether a packet is normal or not. The FCM algorithm employs fuzzy partitioning such that a data point can belong to all groups with different membership grades. Together, CUSUM and FCM become a good technique in detecting network anomalies with a very less false alarm rate.

Keywords- Network Security , Intrusion detection system, Feature selection , Data Mining,Fuzzy c-mean Clustering, KDD'99 dataset

I. INTRODUCTION

Early Internet architecture design goals did not put security as a high priority. However, today Internet security is a quickly growing concern. The prevalence of Internet attacks has increased significantly, but still the challenge of detecting such attacks generally falls on the end hosts and service providers, requiring system administrators to detect and block attacks on their own. In particular, as social networks have become central hubs of information and communication, they are increasingly the target of attention and attacks. Thus a challenge of carefully distinguishing malicious connections from normal ones is created. Previous work has shown that for a variety of Internet attacks, there is a small subset of connection measurements that are good indicators of whether a connection is part of an attack or not. These security issue creates the challenge for system administrators to distinguish normal connections of legitimate users from malicious connections. An administrator, in the face of an attack, wants to block such malicious connections without blocking legitimate users. One can never know for sure if any given connection is legitimate or malicious, but one can attempt to isolate a set of connections that stand out from normal user behaviour, so as to help system administrators detect attacks which is done by a network intrusion detection system.

1.1 INTRUSION DETECTION SYSTEM

An intrusion detection system (IDS) monitors network traffic and monitors for suspicious activity and alerts the system or network administrator. As shown in Figure 1.1, in some cases the IDS may also respond to anomalous or malicious traffic by taking action such as blocking the user or source IP address from accessing the network. IDS come in a variety of "flavours" and approach the goal of detecting suspicious traffic in different ways. There are network based (NIDS) and host based (HIDS)

intrusion detection systems. There are IDS that detect based on looking for specific signatures of known threats- similar to the way antivirus software typically detects and protects against malware- and there are IDS that detect based on comparing traffic patterns against a baseline and looking for anomalies. There are IDS that simply monitor and alert and there are IDS that perform an action or actions in response to a detected threat.

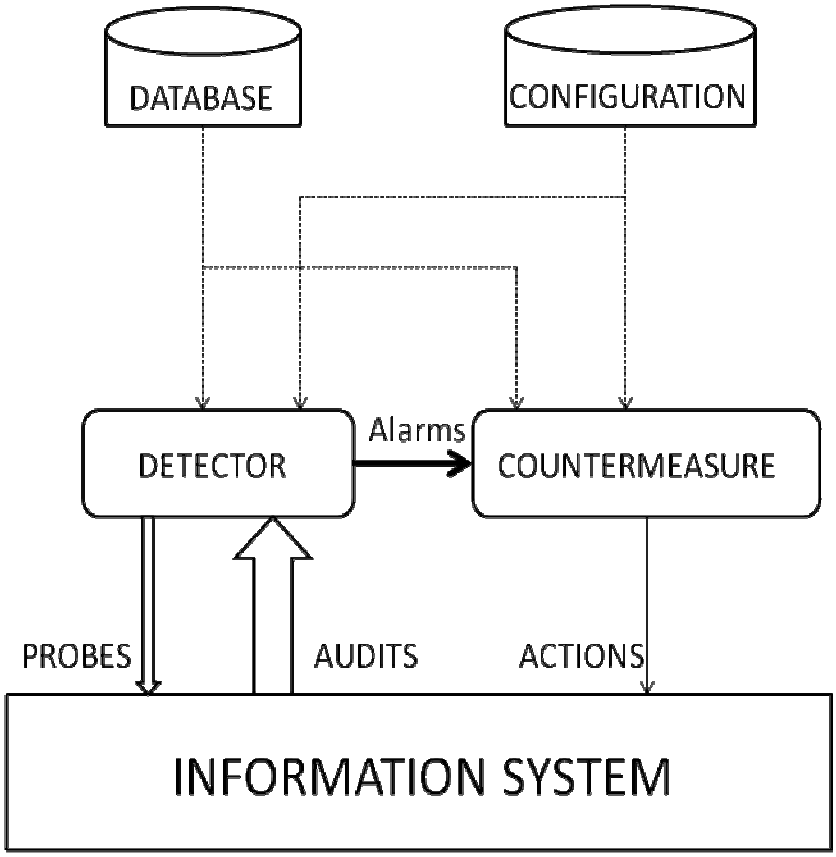

Figure 1.1 An Intrusion Detection System

Two types of intrusion detection systems are **Misuse Intrusion Detection-**Misuse or Signature intrusion detection refers to intrusion that follows well-defined patterns of attack that exploit weaknesses in system and application software. Such patterns can be precisely written in advance. **Anomaly Intrusion Detection -**An IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline will identify what is "normal" for that network- what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other- and alert the administrator or user when traffic is detected which is anomalous, or significantly different, than the baseline.

II. EXISTING SYSTEM

Different terminologies are used in various applications, such as Novelty or surprise detection and fault detection. Novelty detection is a useful technique in cases where an important class of data is represented in training set. This means that the performance of the network will be poor for those classes. If the training data only consists of examples from one class and the test data contains examples from two or more classes, the classification task is called anomaly detection or novelty detection. With enterprise networks, network analysers are often attached to the lines in order to monitor traffic and send an alarm when disruptions are detected.

Gunes Kayacik et al., (2010), KDD means Knowledge discovery and data mining. It is collection of various datasets. KDD Cup 99 dataset consists of network based datasets which was used as training set for creating a IDS Model using cascaded classification technique. The NSL-KDD is the new version KDD'99 data set. It solves some of the inherent problems of the KDD'99 data set. It can be applied to an effective data set to help researchers compare different intrusion detection methods. It runs the experiment on the complete set without the need to randomly select a small portion. In the existing system, Intrusion detection systems have been used to detect known attacks by matching misused traffic patterns or signatures. A class of IDS that leverages machine learning can also detect unknown attacks by identifying abnormal network traffic that deviates from the so-called "normal" behavior previously profiled during the IDS training phase. Most of intrusion detection systems nowadays rely on handcrafted signatures just like anti-viruses which have to be updated continuously in order to be effective against new attacks.

III. PROPOSED SYSTEM

The main of the he proposed system is to provide a hybrid framework combining the two anomaly detectors namely CUSUM and Fuzzy C-means clustering technique. KDD Cup 99 dataset consists of network based datasets. This dataset is used as an input for the process. Cumulative Sum algorithm (CuSum) is used to detect anomalies based on statistical information of packets in the networks. In brief, CUSUM detect changes based on the cumulative effect of the changes made in the random sequence instead of using a single threshold to check every variable. CUSUM is a broadly adopted algorithm for detection of abrupt traffic flow change, especially for DoS packet flooding. CUSUM is good at detecting the abrupt change of the mean of an observed sequence. By adjusting parameters of CUSUM, one can make trade off between the detection sensitivity and the degree of noise-resilience in an anomaly detection system. Unlike crisp clustering that crisply assigns each data point to a separate cluster, fuzzy clustering allows each data point to belong to various clusters with different membership degrees (or weights). Fuzzy clusters are expressed by their centers (or centroids) that are simultaneously found in the partitioning process of a fuzzy clustering algorithm. The number of clusters is often inputted as a parameter to a fuzzy clustering algorithm.

3.1 MODULES DESCRIPTION

The proposed system considers the three sub processes. They are as follows:

- Feature Selection
- CUSUM
- Clustering

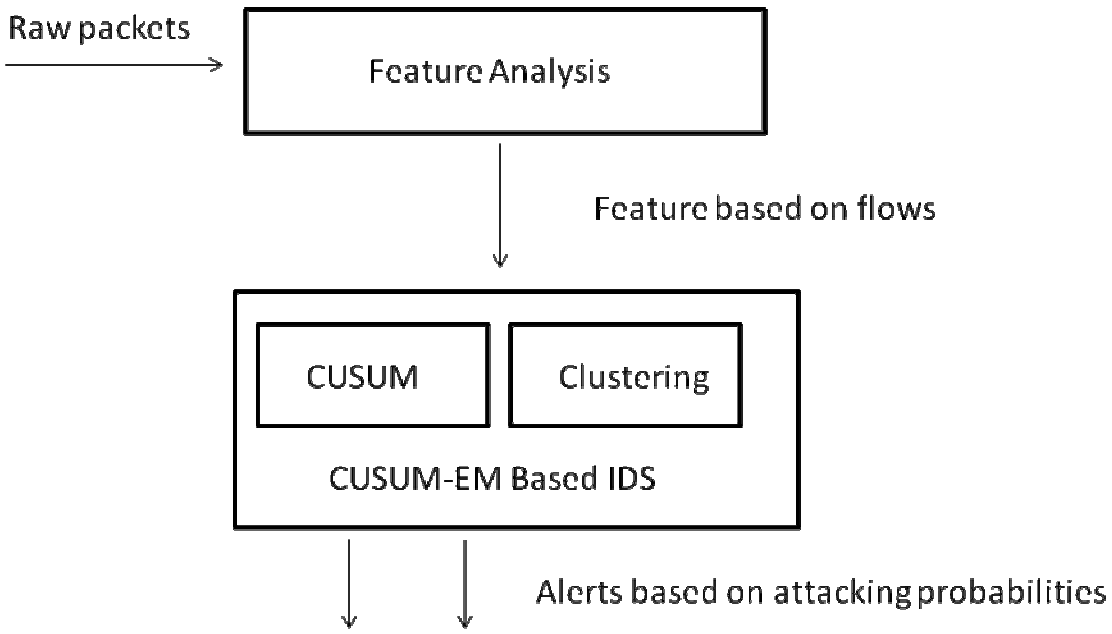

Figure 3.1 System Flow Diagram

3.1.1 Feature Selection

The KDD 99 intrusion detection datasets are based on the 1998 DARPA initiative, which provides designers of intrusion detection systems (IDS) with a benchmark on which to evaluate different methodologies. The raw data collected is usually substantially large, so it is desired that one select a subset of this data by creating feature vectors that represent most of the information we need from the data. There are 41 features in the KDD Cup 99 dataset, out of which 11 features are selected here. The selected features are: Duration, Protocol Type, Service, Flag, Count, Src_bytes, Dets_bytes, Land, Urgent, Wrong fragment, Logged in. All these features belong to the intrinsic attributes. Feature selection is done for selecting a subset of relevant features for use in a dataset. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection returns a subset of the features. Selections of these features provide two main benefits: shorter training times,Eliminate features with less or no predictive information.

3.1.2 CUSUM

A procedure, named Change-Point Detection designed to detect the change point of a network behaviour, is as follows. First, compare observed event sequence with user profiles. If any difference is significant, identify the time point the change happens so as to real time discover when the attack starts. Second, CUSUM is deployed to sequentially monitor input random variables. A simple parametric approach is often too simple to accurately model a network session due to the complexity of Internet. CUSUM with the characteristics of sequential and nonparametric light computation load can make IDS work online.

In statistical quality control, the CUSUM (or cumulative sum control chart) is a sequential analysis technique used for monitoring change detection. CUSUM involves the calculation of a cumulative sum and it can determine if a node is a zombie or not. The CUSUM algorithm is an approach to detect a change of the mean value of a stochastic process and it is based on the fact that if a change occurs, the probability distribution of the random sequence will also be changed.

Algorithm

CuSum = 0
n = 0
Repeat
n = n + 1
CuSum = CuSum + Xn
If CuSum > ThresHold
then Signal attack indication Until Finished

The cusum value is initially initialised to zero. n is the packet number. The cumulative sum is calculated. A threshold value is set and when the cusum value exceeds the cumulative sum the packet could be considered as an attack packet or else it is ignored.

3.1.3 Clustering

Cluster analysis is identifying such grouping (or clusters) in an unsupervised manners, in unsupervised approach are divides a set of objects into homogeneous groups. There have been many clustering algorithms scattered in publications in much diversified areas such as pattern recognition, artificial intelligence, information technology, image processing, biology, psychology, and marketing. Clustering algorithms can be classified into main two categories: hard clustering algorithms and fuzzy clustering algorithms. Unlike hard clustering algorithms, which require that each data point of the data set belong to one and only one cluster, fuzzy clustering algorithms allow a data point to belong to two or more clusters with different probabilities. Basically hard clustering has each document belongs to exactly one cluster. In hard clustering a hard partition of the dataset is made.

The clustering algorithm employed here is called as the Fuzzy C-means algorithm (FCM). The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1. Fuzzy C-means Clustering (FCM) is also known as Fuzzy ISODATA, is a clustering technique which is separated from hard k-means that employs hard partitioning.FCM is an iterative algorithm. The aim of FCM is to find cluster centers (centroid) that minimize a dissimilarity function. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. Fuzzy Clustering also called soft clustering. In fuzzy clustering fuzzy partition of the data set is made. Fuzzy clustering uses membership function in partition data set. To accommodate the introduction of fuzzy partitioning, the membership matrix (U) is randomly initialized.

1. $\Sigma^{c}_{i=1} u_{ij} = 1,$ for all j=1,.....,n

This function is called membership function and its value between 0 and 1.To find the value of centroid (c_i) with help of membership matrix (u_{ij}).

2. $$c_i = \frac{\Sigma^{n}_{j=1}\ u_{ij}{}^{m} X_j}{\Sigma^{n}_{j=1}\ u_{ij}{}^{m}}$$

The dissimilarity function which is used in FCM is given Equation

$$J(U,c_1,c_2,..,c_e) = \Sigma^{e}_{i=1} J_i = \Sigma^{e}_{i=1} \Sigma^{n}_{j=1}\ u_{ij}{}^{m} d^2_{ij}$$

u_{ij} is between 0 and 1

d_{ij} is the Euclidian distance between ith centroid (c_i) and jth data point ;

where $d_{ij=} \sqrt{\Sigma ni=1\ (Xi - Ci)^2}$

$$U_{ij} = \frac{1}{\Sigma^{c}_{k=1} (d_{ij}/ d_{kj})^{2/(m-1)}}$$

If $\| U(k+1) - U(k) \| < \varepsilon$

Then STOP; otherwise return to step 2.

IV. RESULTS AND DISCUSSION

The main performance measure for evaluating the intrusion detection model is detection rate. Detection rate is defined as how accurately the IDS classifies the test data set. Here CUSUM and Clustering algorithms are combined together for better efficiency. There are 11 attributes which are selected from 41 features of the KDD data set. There are around 494,020 records which is too large for our purpose. Hence a concise dataset 10% KDD Cup 99 is employed here. It could be seen that the CUSUM was able to detect around 1200 attack packets while FCM was able to find around 1450 packets which clearly shows that FCM is far better than that of CUSUM. A graph depicting the number of attack packets detected by each algorithm is shown here in Figure 4.1.

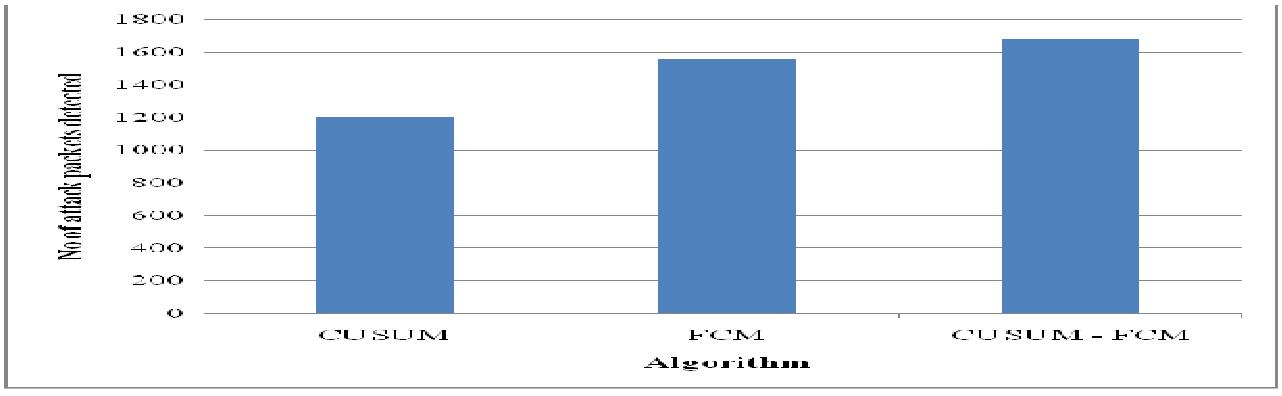

Figure 4.1 No of attack packets detected

V. CONCLUSION AND FUTURE WORKS

Intrusion Detection System plays an important role in computer security. IDS users relying on the IDS to protect their computers and networks demand that an IDS provide reliable and continuous detection service. However, many of the today's anomaly detection methods generate high false and negatives. The CUSUM and Fuzzy C means algorithm is a good technique to address the problem. All anomaly-based intrusion detection system, work on the assumption that normal activities differ from the abnormal activities substantially. In the existing system, various other novel algorithms such as K-means are employed. Here, two algorithms based on CUSUM and Fuzzy C Means is used for analyzing the behavior in the intrusion detection are evaluated by experiments. The preliminary experiments with the KDD Cup 99 dataset have shown that this approach is able to effectively detect intrusive behavior. The results also show that a low false positive rate can be achieved. More change point detection algorithms like Shiryaev-Roberts procedure could also be added in order to minimize the false alarm rate. The proposed system is an applied for dataset. Online packets could be caught and could be fed as an input to the IDS using CUSUM and clustering.

REFERENCES

[1] Alexander G Tartakovsky, Aleksey S Polunchenko and Grigory Sokolov,'Efficient Computer Network Anomaly Detection by Changepoint Detection Methods,' IEEE Journal of Selected topics in Signal Processing, Vol. 7, No. 1, pp.4-11,2013.

[2] Thatte G, Mitra U, and Heidemann J, 'Parametric Methods For Anomaly Detection In Aggregate Traffic,' IEEE/ACM Transactions on Networking, Vol. 19, No. 2, pp.512–525,2011

[3] Dash M and Liu H, 'Feature Selection for Clustering,' Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Vol. 5, No.8, pp.110–121,2000.

[4] Wu Jain, Feng Guo Rui, 'Intrusion Detection Based On Simulated Annealing And Fuzzy C-means Clustering,' MINES International Conference, Vol. 2, No.1, pp.382-385,2009.

[5] Hiren S, Jeffrey Ur, Anupam J, 'Fuzzy Clustering for Intrusion Detection,' The 12th IEEE International Conference on Fuzzy Systems, Vol.2, No.5 pp.1274-1278.2003.

