Robust Data Clustering Algorithms for Network Intrusion Detection

Gunja Ambica [#1], Mrs.N.Rajeswari[#2]
[1] *M.Tech (CSE),Gudlavalleru Engineering College, Gudlavalleru*
[2] *Associate Professor, Gudlavalleru Engineering College, Gudlavalleru.*

ABSTRACT:
IDS (Intrusion Detection system) is an active and driving defense technology. Intrusion detection is to detect attacks against a computer system. This project mainly focuses on intrusion detection based on data mining. Data mining is to identify valid, novel, potentially useful, and ultimately understandable patterns in massive data. One of the primary challenges to intrusion detection are the problem of misjudgment, misdetection and lack of real time response to the attack. In the recent years, as the second line of defense after firewall This project presents an approach to detect intrusion based on data mining frame work. In this framework, intrusion detection is achieved using clustering techniques. Firstly, a method to reduce the noise in the data set using improved kmeans. This system use K-means,FCM and Improved K-means data mining algorithms are used to improves the performance of intrusion detection since the traffic is large and the types of attack are various. By the more accurate method of finding k clustering center, an anomaly detection model was presented to get better detection effect. This project used KDD CUP 1999 data set to test the performance of the model. The results show the system has a higher detection rate and a lower false alarm rate, it achieves expectant aim.

I INTRODUCTION
A network intrusion attack often is any use of a network that compromises its stability and even the security of real info that would be stored on computers coupled with it. A wide range of activity is categorized as this definition, including effort to de-stabilize the network overall, gain unauthorized
access to files or privileges, simply mishandling and misuse of software. Added security measures can stop these kind of attacks. The intention of intrusion detection is to create system which could automatically scan network activity and detect such intrusion attacks. Once an attack is detected, the machine administrator could well be informed and in consequence take corrective action. Detecting such abusive simply not only provides information on damage assessment, but additionally will help to prevent future attacks. These attacks are normally detected by tools known as intrusion detection system. The most popular and well-known data to have an intrusion detection method is the audit data. An audit trail is the records among the activities on a system kept in chronological order. Since there exist note for one activity (which might even correspond to one system call) inside the system, theoretically it is more than possible manually analyze the source data and detect any abnormal activity inside the system. However, the vastness of the audit data provided by an audit collection system often makes manual analysis impractical. Therefore, an automated audit data analysis tool is considered the only solution.
An intrusion detection system (IDS) is software and/or hardware invented to detect unwanted attempts at accessing, manipulating, and/or disabling of computer system, mainly through a network, typically the internet. One of the main challenges in the security management of large-scale high-speed networks (LSHSN) happens to be the detection of anomalies in network traffic. A secure network must provide the following:
• Data confidentiality: Data that are being transferred in the network should be accessible only to those which have been properly authorized.
• Data integrity: Data should maintain their integrity from the moment they are transmitted towards the moment they are actually received. No corruption or data loss is accepted either from random events or malicious activity.
• Data availability: The network ought to be resilient to Denial of Service attacks.

Anomaly detection: It truly is based on the normal behavior associated with a subject (e.g. an individual or possibly a system). Any action that significantly deviates coming from the normal behavior is held to be as intrusive. Which means when we could generate a normal activity profile to produce a system. Our team can flag all system states varying from established profile. Misuse/Signature detection: misuse detection catches intrusions in regards to the characteristics of known attacks. Any action that conforms to the pattern of a known attack or vulnerability is considered as intrusive. The best issues in misuse detection system are tips to write a signature that encompasses all possible variations of the pertinent attack. And the best way to write signatures that don't also match non-intrusive activity.

1) Training data containing flow records of both normal and anomalous traffic are transformed into feature datasets.
2) The datasets are divided into different clusters for normal and anomalous traffic using the Kmeans clustering algorithm.

3) The resulting cluster centroids are deployed for fast detection of anomalies in new monitoring data based on simple distance calculations. While clustering monitoring data and identifying anomalies based on outlier detection has already been tried before, we are not aware of previous attempts generating additional clusters for anomalous traffic as we do. K-means [5] is definitely one of the simplest unsupervised learning algorithms that solve the famous clustering problem. The movement follows a trouble-free and way to classify a confirmed data set over a certain multitude of clusters (assume k clusters) fixed a priori. The best idea is to define the term k centroids, one for each cluster. These centroids really should be placed in a cunning way because of different location causes different result. So, the greater choice is to mark them as much as possible far-off from each other. What comes next is to accept each point belonging to a given data set and associate it into the nearest centroid. When no point is pending, your first step is finished
and an early groupage is completed. Right now it becomes necessary re-calculate k new centroids as barycenters of one's clusters as a result of the last step. When we have these k new centroids, a fresh binding you ought to do amongst the same data set points and naturally the nearest new centroid. A loop has been generated. Due to this loop we could find that the k centroids change their location step by step until eliminate changes are executed. Quite simply centroids do not move any more. Our Network data mining approach deploys the K- mean clustering algorithm in order to separate time interval with normal and anomalous traffic in the training dataset. The resulting cluster centriods are then used for fast anomaly detection in new monitoring data.

II BACKGROUND AND RELATED WORK

Supervised Methods: The main goal as to the supervised methods is to design a predictive model (classifier) to classify or label incoming patterns. The classifier has to be trained with labeled patterns in order to classify new unlabeled patterns. The given labeled training patterns are use to here are the description of classes. Some supervised methods include support vector machines, neural network and genetic algorithms to name a few. 2.3.2 Unsupervised Methods
Unsupervised methods, also termed as data clustering, use a different approach by grouping unlabeled patterns into clusters based upon similarities. Patterns contained in the same clusters are more a dead ringer for one other than they're to patterns owned by different clusters. Data clustering is extremely useful when little priori details about information is offered. Clustering methods can be classified into two categories: hierarchical clustering algorithms (Figure 1,a) and partitioned clustering algorithms(Figure 1,b).

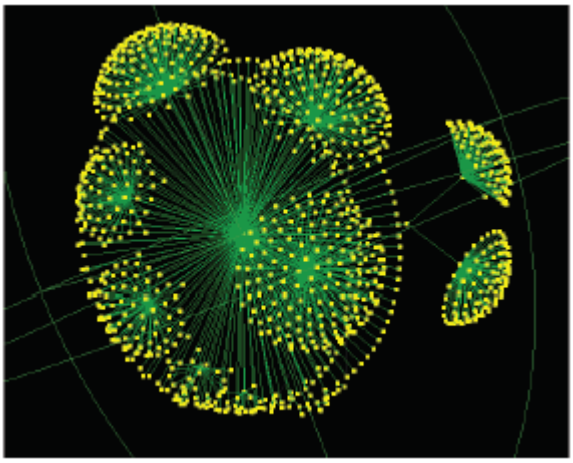
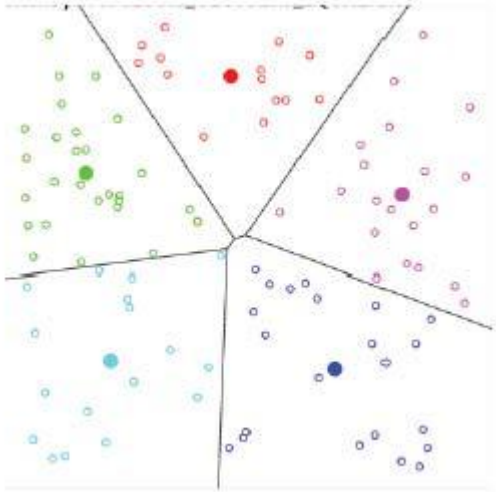


Figure 1 (a) Hierarchical clustering output



(b) Partitional clustering output

James Anderson[2] first proposed that audit trails should be utilized monitor threats. Most of the available system security procedures were geared toward denying admission to sensitive data because of an unauthorized source. Dorothy Denning [4] first proposed the concept of intrusion detection as a chemical solution the topic of providing a sense of security in computer systems. The basic idea is that intrusion behavior involves abnormal usage of sst. Dinner gown model is most definately a rule-based pattern matching system. Some models of normal usage of the internal system could well be constructed and verified against usage of the machine and the significant deviation coming from the normal usage flagged as abnormal usage. This model served as an abstract model for further developments in this particular field and it is generally known as generic intrusion detection model.

IDES [5] used expert system strategies for misuse intrusion detection and statistical techniques for anomaly detection. IDES expert system component evaluates audit records as they are produced. The audit records are viewed as facts, which map to rules in the rule-base.

Problem Definition:

- Existing system has high false alarm rate.
- Previous algorithms has low attack detection rate.
- Existing algorithms does not handle huge dataset.
- Existing techniques does not implement outlier before the clustering.
- Existing Kmeans algorithm uses static initialization as k initial centroids.
- When the numbers of data are not so many, initial grouping will determine the cluster significantly.
- The number of cluster, K, must be determined before hand.
- We never know the real cluster, using the same data, if it is inputted in a different order may produce different cluster if the number of data is a few.
- Sensitive to initial condition. Different initial condition may produce different result of cluster. The algorithm may be trapped in the local optimum.
- We never know which attribute contributes more to the grouping process since we assume that each attribute has the same weight.
- weakness of arithmetic mean is not robust to outliers. Very far data from the centroid may pull the centroid away from the real one.
- The result is circular cluster shape because based on distance.

Ko et al. at UC Davis first proposed to specify the intended behavior of some privileged programs (setuid root programs and daemons in UNIX) using a program policy specification language [42]. During the program execution, any violation of the specified behavior was considered "misuse". The major limitation of this method is the difficulty of determining the intended behavior and writing security specifications for all monitored programs. Nevertheless, this research opened the door of modeling program behavior for intrusion detection.

Leonid Portnoy [53] presented method for detecting Intrusion based on feature vector collected from network, without being given any information about classification of these vectors. He designed a system that implemented clustering technique and able to detect a large number of intrusions while keeping false positive rate reasonable low. Data clustering technique has advantage over the signature based classifier. First that no manual classification of training data is needs to done. The second is that we do not aware of new types of intrusions in order for the system to be able to detect them.

3. PROPOSED FRAMEWORK

This section, it gives an overview of the data set used for intrusion detection. This data set contains seven weeks of training data and two weeks of testing data. The raw data was about four gigabytes of compressed binary TCP dump data from the of network traffic generated. This was processed into about five million connection records, each of which is a vector of extracted feature values of that network connection. As we know, a connection is a sequence of TCP packets to and from some IP addresses, starting and ending at some well defined times. This data set of the five million connection records was used as the data set for the 1999 KDD intrusion detection contest and is called the KDD Cup 99 data. In particular, MIT Lincoln Lab's DARPA intrusion detection evaluation datasets have been employed to design and test intrusion detection systems. In 1999, recorded network traffic from the DARPA 98 Lincoln Lab dataset [4] was summarized into network connections with 41-features per connection. This formed the KDD 99 intrusion detection benchmark in the International Knowledge Discovery and Data Mining Tools Competition.

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 [3]. To do so, a simulation is made of a factitious military network consisting of three 'target' machines running various operating systems and services. Additional three machines are then used to spoof different IP addresses to generate traffic. Finally, there is a sniffer that records all network traffic using the TCP dump format. The total simulated period is seven weeks.

Each connection was labeled as normal or as exactly one specific kind of attack. All labels are assumed to be correct. There were a total of 37 attack types in the data set. The simulated attacks fell in exactly one of the four categories : User to Root; Remote to Local; Denial of Service; and Probe.

• **Denial of Service (dos):** Attacker tries to prevent legitimate users from using a service.

• **Remote to Local (r2l):** Attacker does not have an account on the victim machine, hence tries to gain access.

• **User to Root (u2r):** Attacker has local access to the victim machine and tries to gain super user privileges.

• **Probe:** Attacker tries to gain information about the target host.

Data pre-processing:

Data preprocessing comprises following components including document conversion, feature selection and feature weighting.

The functionality of each component is described as follows:
(1) Dataset prepared with DOS attack which include smurf, Neptune, back, teardrop and POD ping of death attacks / anomaly.
(2) Feature selection – reduces the dimensionality of the data space by removing irrelevant or less relevant feature selection criterion.
(3) Document conversion- converts different types of documents such as gz, tcpdump to csv file and arff (Attribute-Relation File Format) data file format.

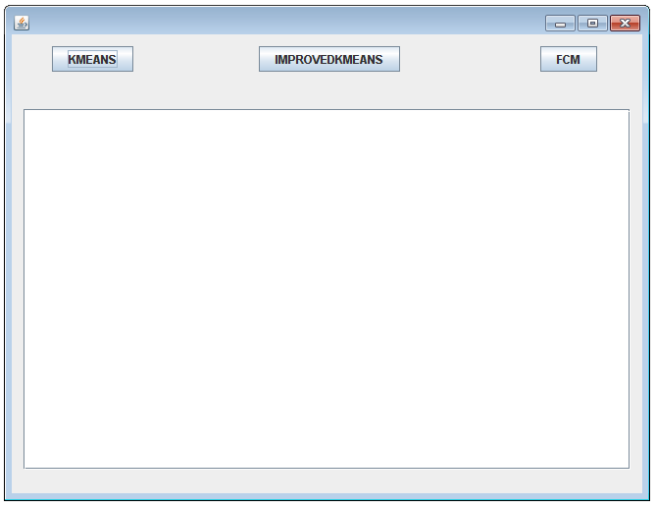
The pseudo code for the **adapted kMean algorithm** is presented as below:
/*****start of pseudo code
1. Choose random k data points as initial Clusters Mean (cluster center)
2. Repeat
3. for each data point x from D
4. Computer the distance x and each cluster mean (centroid)
5. Assign x to the nearest cluster.
6. End for
7. Re-compute the mean for current cluster collections.
8. Until reaching stable cluster
9. Use these centroid for normal and anomaly traffic.
10. Calculate distance of centroid from normal and anomaly centroid points.
11. If distance(X, Dj) > = 5
12. Then anomaly found ; exit
13. Else then
14. X is normal;
*/ end of pseudo code.

FCM ALGORITHM:

Input: n data objects, number of clusters
Output: membership value of each object in each cluster
Algorithm:
1. Select the initial location for the cluster centres
2. Generate a new partition of the data by assigning each data point to its closest centre.
3. Calculate the membership value of each object in each cluster.
4. Calculate new cluster centers as the centroids of the clusters.
5. If the cluster partition is stable then stop, otherwise go to step2 above.

Experimental Results:

All experiments were performed with the configurations Intel(R) Core(TM)2 CPU 2.13GHz, 2 GB RAM, and the operating system platform is Microsoft Windows XP Professional (SP2).



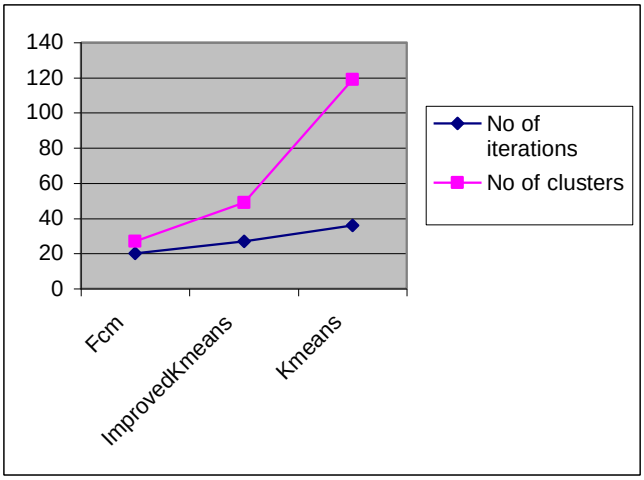
FCM results:

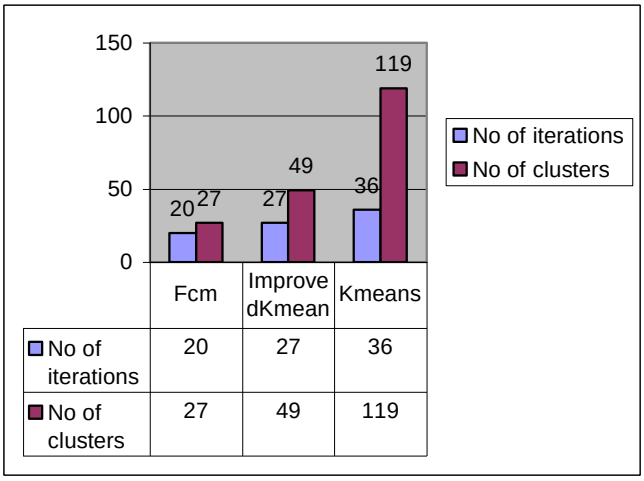
Cluster centroids:

Attribute	Full Data (5291)
duration	318.4838
protocol_type	tcp
service	http
flag	SF
src_bytes	82705.3175
dst_bytes	1934.5177
land	0
wrong_fragment	0.028
urgent	0.0002
hot	0.183
num_failed_logins	0.0015
logged_in	0.3937
num_compromised	0.1546
root_shell	0.0019
su_attempted	0.0011

KMEANS RESULTS:

```
Cluster centroids:

Attribute                               Full Data
                                           (5291)
=========================================
duration                                318.4838
protocol_type                                tcp
service                                     http
flag                                          SF
src_bytes                             82705.3175
dst_bytes                              1934.5177
land                                          0
wrong_fragment                            0.028
urgent                                   0.0002
hot                                       0.183
num_failed_logins                        0.0015
logged_in                                0.3937
num_compromised                          0.1546
root_shell                               0.0019
su_attempted                             0.0011
num_root                                  0.175


 IMPROVEDKMEANS
==============

Cluster centroids:

Cluster 0
          0.0 tcp private S0 0.0 0.0
Cluster 1
          0.0 tcp http REJ 0.0 0.0 0.
Cluster 2
          12039.0 tcp telnet SF 798.0
Cluster 3
          0.0 icmp ecr_i SF 1032.0 0.
Cluster 4
          36.0 tcp ftp SF 1463.0 4152
Cluster 5
          0.0 tcp other REJ 0.0 0.0 0
Cluster 6
          0.0 tcp http S1 236.0 29200
Cluster 7
          35682.0 tcp telnet RSTOS0 3
Cluster 8
          0.0 udp domain_u SF 29.0 0.
Cluster 9
          6.0 tcp gopher SF 0.0 0.0 0
Cluster 10
          0.0 tcp http RSTR 54540.0 8
```





Above graphical representations shows that proposed algorithm improves in decreased the error, number of clusters and iterations.

6. CONCLUSION AND FUTURE WORK

A set of normal and abnormal processes was taken for experiments. This set was fed into k-mean clustering. Using K-mean clustering algorithm centroid is calculated for anomaly and normal data. 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. A Data clustering using K-mean algorithm is a good technique to address these problems.

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.Experimental results show that proposed clustering algorithms detect better attacks with less time.In future proposed algorithms are used with classifiers in order to get better results.

REFERENCES:

[1] Data Clustering Using K-Mean Algorithm for Network Intrusion Detection, Satinder Pal Singh, Lovely Professional University, Jalandhar MAY-2010.

[2] C. Wang and J. C. Knight. Towards survivable intrusion detection. In Proceedings of the 3rd Information Survivability Workshop (ISW-2000), Boston, USA, October 2000.

[3] Tom Mitchell. Machine Learning. Mc Graw Hill, 1997.

[4] K. Aas and L. Eikvil, Text Categorisation: A Survey, http://citeseer.nj.nec.com/ aas99text.html, 1999.

[5] Carl Endorf, Gene Schultz, and Jim Mellander. Intrusion Detection and Prevention. McGraw- Hill Osborne Media, first edition, 2003.

[6] W. Lee and S. J. Stolfo. A framework for constructing features and models for intrusion detection systems. In Proceedings of ACM Transactions on Information and System Security (TISSEC), volume 3(4), pages 227–261.

[7] Wang, Q. and V. Megalooikonomou. " A clustering algorithm for intrusion detection. in SPIE Conference on Data Mining " , Intrusion Detection, Infonnation Assurance, and Data Networks Security. 2005. Orlando, Florida, USA.

[8] Jiawei Han Micheline Kamber, " Data Mining Concepts and Techniques " , Second Edition.

[9] Masakazu Seno, George Karypis, " Finding Frequent Patterns Using Length-Decreasing Support Constraints", Data Mining and Knowledge Discovery, pp.197-228, 2005.

[10] W. Lee, S. Stolfo, and K. Mok, Mining audit data to build intrusion detection models, Proc. 4th International Conf. on Knowledge Discovery and Data Mining (KDD ’98), New York City, NY, 1998, 66-72.

[11] H. Toivonen, Sampling large databases for association rules, Proc. 22nd international conf. on very large data bases (VLDB’96), Mumbai, India, 1996, 134-145.

[12] S. Lee and D. Cheung, Maintenance of discovered association rules: When to update?, Proc. 1997 ACMSIGMOD workshop on research issues on data mining and knowledge discovery (DMKD’97), Tucson, AZ, 1997.