ISSN (Online): 2455-7838
SJIF Impact Factor (2017): 5.705

EPRA International Journal of

Research & Development

(IJRD)

Monthly Peer Reviewed & Indexed International Online Journal

Volume: 3, Issue:8, August 2018

Published By :
EPRA Journals

SJIF Impact Factor: 5.705 Volume: 3 | Issue: 8 | August | 2018 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)

DDoS ATTACK DETECTION USING DATA MINING CLUSTER ANALYSIS

Ms. Ambika G N
Assistant Professor, Dept of CSE, BMS Institute of Technology and Management, Bangalore-560064

ABSTRACT

DoS/DDoS attacks are detected by invoking a statistical approaches that compares source IP addresses normal and current packet statistic to discriminate whether there is a DoS/DDoS attack. It first collects all resource IP's packet statistics so as to create their normal packet distributions. To detect the DoS/DDoS attack, feature data points are extracted from IP packet statistics dataset and are given as input to the clustering algorithms (K-Means algorithm, Fuzzy-c means algorithm and K-Medoids algorithm) which determines the presence of attacked packets. Analysis is made on the performance of each algorithm.

DDoS attacks are detected with the help of clustering algorithms each giving different false alarm rate, accuracy and execution time depending upon the different input data size.

KEYWORDS: *Denial of Service (DoS), Distributed Denial of Service (DDoS), K-means, Fuzzy c-means and K-medoids Algorithms.*

INTRODUCTION

Network Security is one of the most important issues that can be considered by commercial organizations to protect its information from malicious attacks. The problems of detecting malicious traffics have been widely studied and still as a intrested research topic in the recent decades. Many researches have been designed and implemented an Intrusion Detection System (IDS) to analyze, detect and prevent the malicious activities such as Distributed/Denial of Service Attack (DDoS/DoS).

IDS's can be classified into two types :

- Misuse Intrusion Detection (MIS) and
- Anomaly-Intrusion Detection (AID).

Misuse detection constructs from known attack behavior based on the pattern matching, which can Be used later as signature-based for attack possibilities. However, Anomaly-Intrusion Detection creates from the long term of normal usage behavior profile of network traffic. In general, IDS's can be approached by data mining techniques to detect unusual access or attacks to secure internal networks.

Denial of Service attack consists of dangerous threats that are able to disturb a CIA (Confidentiality, Integrity and Availability) services on the network. It consists of a series of attacks able to degrade the network quality service in highly predictable manner. A very common example of this type of attack is Distributed Denial of Service (DDoS) attack. In this instance, multiple computers are being used to send attacks to a victim in the same time during the attack. Zombies are common names given for the computers under the control of the attacker through Handlers. Handlers are the software packages that attacker uses for communication with the zombies. Zombies may or may not be known that they are attacking a victim of network. In general, the attacker acquires the control with zombies by communicating with

any number of handlers to determine which agents are running to schedule attacks. Usually, the attacker tries to install the handler software on a compromised routers or server that handles a large volume of traffic. Figure 1 illustrates the general model of DDoS attack.

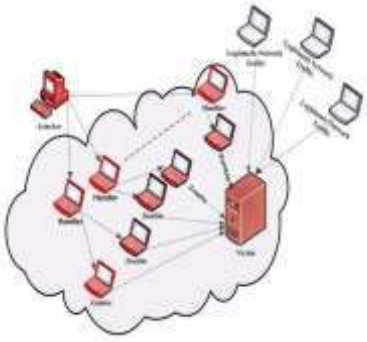

Figure 1: DDoS zombies-handlers attack model

Data mining techniques may play major roles in detecting the malicious on the network traffic such as DDoS attacks. IDS's can also be approached by data mining machine learning techniques. Data mining techniques can be classified into main approaches:

- Unsupervised and
- Supervised

In data mining supervised methods, there is a predefined class (target variable) and the algorithms from many examples where the value of class is provided. In this instance, we have a training phase to construct a predictor model and testing phase to assign each unknown value to the class variable that obtained from the training phase. The goal of the unsupervised techniques is to extract a new useful knowledge from a large data set by clustering the similar objects and separating the dissimilar objects based on some defined dissimilarity measure.

Motivation

With the amount of information and resources available on the Internet, it may even be possible for a user with some technical knowledge to download and run a simple script that performs a DDoS attack. Regular Internet users may attempt to attack a big company's website simply because they can. Being able to take down a large companies or organization's website are made enticing to average, insignificant computer user.Personal conflict may even motivate some users to perform DDoS attacks on different user's home computer, for the sake of revenge. Also, within online communities, like "hackers world", attackers are able to make the way in taking down a target may receive some form of fame or recognition in their community, mainly due to the childishness of their motivation.

Some theories reported that company websites under attack are being attacked by competing companies. This can disrupt the targeted company's services, which may very well boost the sales or amount of website views for the competing company.

There are some cases where DDoS has been used against companies for ransom purposes. The attackers would attack first and take control of the company's website until it is almost unusable or until it is down and demand ransom money in change for stopping the attacks. There have also been cases where attackers do not initially attack the company's website, but will claim that they will initiate an attack on the company's website if they do not pay the ransom amount. Some attackers even "rent" out or sell their botnets. All of these cases are motivated by financial reasons, although most reports show they have not been very successful.

Scope

Keeping user's online experience always available requires quick Distributed Denial of Service (DDoS) detection to avoid denial of service outages. Why is early detection so important? Time is literally money when mitigating a DDoS attack, data breach or other types of cyber-attack. Faster detection allows for DDoS mitigation to begin more quickly, which saves money and reduces the impact of damaged brand reputation, lost customers and declining stock prices.

Problem Statement

- Distributed Denial of Service (DDoS) attack generates enormous packets by a large number of agents and can easily exhaust the computing and communication resources of a victim within a short period of time.
- Denial-of-service attack can also lead to problems in the network 'branches' around the actual computer being attacked. For example, the bandwidth of a router between the Internet and a LAN may be consumed by an attack, compromising not only the intended computer, but also the entire network or other computers on the LAN.

LITERATURE SURVEY

Title: Distributed Denial of Service (DDoS) Attacks Detection Mechanism

Author: Saravanam Kumarswamy and Dr.R.Asokan

Method: Pushback

The process is based on solving puzzles from client side

Advantages: Authentication check between client and server.

Disadvantages: Increase complexity in puzzles.

Title: Unsupervised Intrusion Detection Based on FCM and Vote Mechanism

Author: Longlong Li, Qin Chen, Shuiming Chi and Xiaohang Liu

Method: FCM & Vote mechanism

The process is based on misuse detection and anomaly detection. The clustering is based on vote mechanism

Advantages: Robustness and Accuracy.

Disadvantages: Reduces Bandwidth and Difficult to maintain.

Title: DDoS Detection System Based on Data Mining

Author: RuiZhong and GuangxueYue

Method: Data Mining

Advantages: Real Time DDoS attack detection.

Disadvantages: Handlers can be used to give attack instructions.

Title: Statistical Approaches to DDoS Attack Detection and Response

Author: Laura Feinstein, Dan Schnackenberg, RavindraBalupari, Darrell Kindred

Methods: Entropy, Chi-Square statics & network trace

Advantages: High performance and accuracy.

Disadvantages: Time limitation.

ALGORITHMS

K-Means Clustering Algorithm

K-means clustering is one of the simplest unsupervised learning algorithms. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters).The main idea is to define 'k' centers, one for each cluster.

Advantages of K-means clustering:

1. It is fast, robust and easier to understand.
2. Efficient.
3. Gives best result when data set is distinct or well separated from each other.

Disadvantages of K-means clustering:

1. The learning algorithm is not invariant to non-linear transformations
2. Randomly choosing of the cluster center cannot lead us to the fruitful result.
3. Applicable only when mean is defined.
4. Algorithm fails for non-linear data set.

Fuzzy c-means Clustering Algorithm

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. More the data is near to the cluster center, more is its membership towards the particular cluster center.

Advantages of fuzzy c-means clustering:

1. Gives best result for overlapped data set and comparatively better than k-means algorithm.
2. Data point is assigned membership to each cluster center as a result of which data point may belong to more then one cluster center.

Disadvantages of fuzzy c-means clustering:

1. Apriori specification of the number of clusters.
2. With lower value of β we get the better result but at the expense of more number of iteration.
3. Euclidean distance measures can unequally weight underlying factors.

K-Medoids Clustering Algorithm

K-medoid is a classical partitioning technique of clustering the data set of n objects into k clusters known a priori. It is more robust to noise and outliers as compared to k-means because it minimizes a sum of pair wise dissimilarities instead of a sum of squared Euclidean distanced.

Advantages of k-medoids clustering:

1. More robust than k-means in the presence of noise and outliers; because a medoid is less influenced by outliers or other extreme values than a mean.

Disadvantages of k-medoids clustering:

1. Relatively more costly & not so much efficient
2. Need to specify *k*, the total number of clusters in advance.

3. Result and total run time depends upon initial partition.

Difference Between Fuzzy c-means and K-means Algorithms

- The objective functions are virtually identical. K-Means classifies a given set of n data objects in k clusters
- With regards to performance, the FCM needs to perform k (i.e. number of clusters) multiplications for each point, for each dimension.
- A centroid is defined for each cluster. All the data objects are placed in a cluster.
- FCM is quite slower than K-Means.
- FCM/Soft-K-Means is less "stupid" than K-Means when it comes for example to elongated clusters.

Difference Between K-means and K-medoids Algorithms

- K-Means classifies a given set of n data objects in k clusters, where k is the number of desired clusters and it is required in advance.
- In K-medoids, k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster having medoid nearest
- K-medoids method overcomes this problem by using medoids to represent the cluster rather than centroid.
- K-medoids is more robust than k-means in the presence of noise and outliers.

Flow Chart of Clustering Algorithms

Flow Diagram of K-means Clustering Algorithm

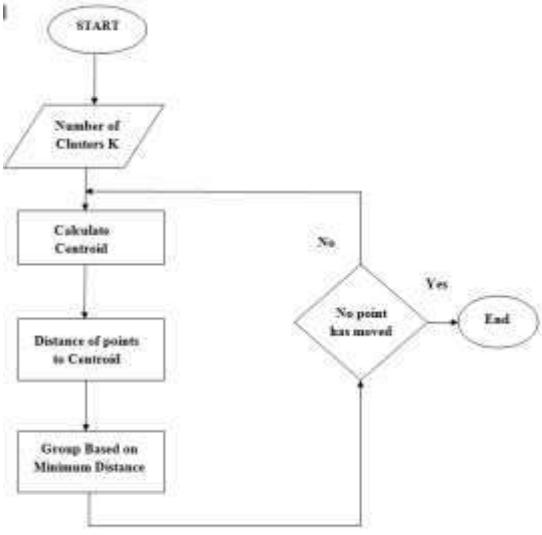

Figure 2: Flow Diagram of K-means Clustering Algorithm

Figure 2 shows the flow chart of k-means clustering algorithm. The algorithm begins by taking data set input. Then we specify the 'k' value to form 'k' clusters for the given data set input. In the next step, we randomly select 'k' centroids. Then we calculate the distance of each data points with the centroid.

In the next step, the data point which is at the minimum distance from the centroid is grouped with nearest centroid to form a cluster. New mean is calculated and reassigning is done. This process keeps on repeating until none of the data points is moved.

Flow Diagram of Fuzzy c-means Clustering Algorithm

Figure 3 shows the flow chart of fuzzy c-means clustering algorithm. The algorithm begins by taking data set input. In the next step, we randomly select cluster centers that are far away from each other. Then we calculate the distance of each data points with the centroid.

In the next step, we assign the membership to each data point corresponding to each cluster center. Then we find the summation of each data point. This process keeps on repeating until the algorithm satisfies the termination criteria.

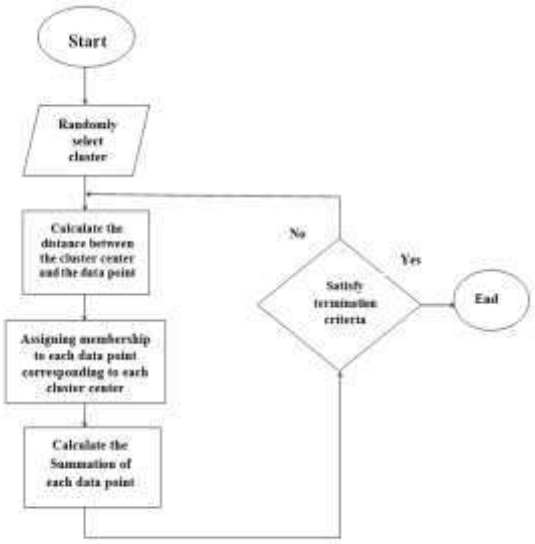

Figure 3: Flow Diagram of Fuzzy c-means Clustering Algorithm

Flow Diagram of K-medoids Clustering Algorithm

Figure 4 shows the flow chart of k-medoids clustering algorithm. The algorithm begins by taking data set input. Then we specify the 'k' value

to form 'k' clusters for the given data set input. In the next step, we randomly select 'k' medoids (which are the data point's center for the clusters) that are far away from each other.

Then we associate each data points to the closest medoid. In the next step, we compute the total cost of configuration of the medoid and non-medoid data points. Then we select the configuration with the lowest cost. New medoid is calculated and reassigning is done. This process keeps on repeating until there is no change in the medoids.

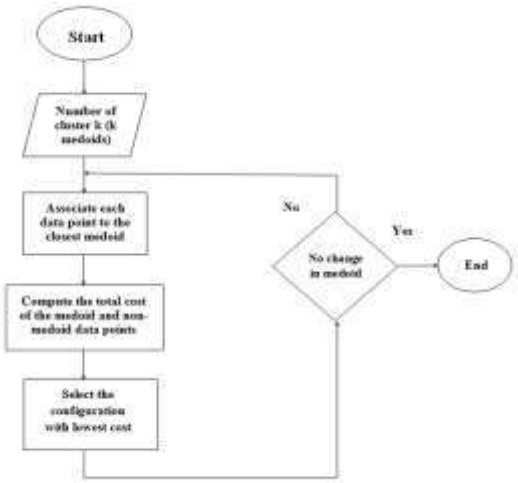

Figure 4: Flow Diagram of K-medoids Clustering Algorithm

Design Constraints

There are two constraints under consideration:

1. Input dataset from CAIDA (Cooperative Association of Internet Data Analysis) is not stationary i.e., varies continuously.
2. Regularity of the data.
   Regularity might be daily, weekly, monthly, yearly or even combination of these.

System Architecture

The system architecture is as shown in Figure 5 The block diagram of the DDoS Detection using Clustering Analysis has 3 main components:

1. Input data sets of CAIDA.
2. Clustering Algorithm which performs the cluster formation of different values.
3. Displays the result.

Figure 5: Model Diagram for DDoS

Data Flow Diagram (DFD) for DDoS Attack Detection

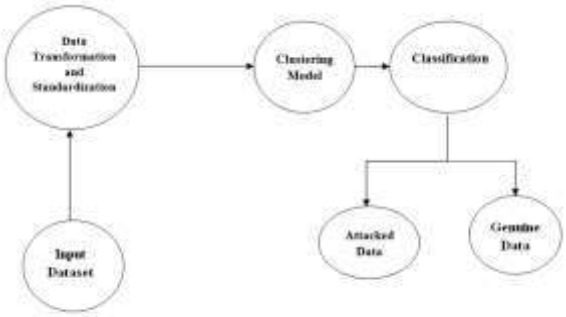

Figure 6: Data Flow Diagram of DDoS Attack Detection

The data set is the input given. In next step the data is converted from .csv format to .text. Features (Attributes) are selected for each data internally.

The attributes of Data Set are

- Time: Describes start time of connection.
- Source IP: Describes source address.
- Destination IP: Shows destination address.
- Source Port: Shows Transport layer port
- Destination Port: shows destination port.
- Protocol: Describes protocol type of TCP/IP suite
- Size: Size of the packet

Then the data with its selected features (attributes) is clustered using different Clustering Algorithms and then it is classified as genuine data and the attacked data.

IMPLEMENTATION

The implementation phase of any project development is the most important phase as it yields the final solution, which solves the problem at hand. The implementation phase involves the actual materialization of the ideas, which are expressed in the analysis document and developed in the design phase. Implementation should be perfect mapping of the design document in a suitable programming language in order to achieve the necessary final product.

Implementation of any software is always preceded by important decisions regarding selection of the platform, language used, etc. These decisions are often influenced by several factors such as real environment in which the system works,the speed that is required, the security concerns, and other implementation specific details.

There are three major implementation decisions that have been made before the implementation of this project. They are as follows:

1. Selection of the platform (Operating System).
2. Selection of the programming language for development of the application.
3. Coding guideline to be followed.

The implementation involves two main modules:

1. **First Module:** Detection of DDoS Attacks.
   a) First sub module imports data, changes the format of data i.e., from .csv (Big file) to .txt (text file).
   b) In second sub module, three different clustering algorithms is used to perform the detection.
2. **Second Module:** Analyzing the Performance of Algorithms Used.
   a) In first sub module, False Alarm Rate and Accuracy level for each algorithm is calculated.
   b) The second sub module displays the comparison of different algorithms in the form of graph.

K-Means Clustering Algorithm

Algorithmic steps for k-means clustering:

Let X = {x_1, x_2, x_3,……,x_n} be the set of data points and V = {v_1, v_2, …..,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) Re-calculate 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) Re-calculate the distance between each data point and new obtained cluster centers.
6) If no data point is reassigned then stop, otherwise repeat from step 3.

FUZZY C-MEANS CLUSTERING ALGORITHM

Algorithmic steps for fuzzy-c means clustering:

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

1) Randomly select 'c' cluster centers.
2) Calculate the fuzzy membership 'μ_{ij}' using:

$$\mu_{ij} = 1\Big/\sum_{k=1}^{c}(d_{ij}/d_{ik})^{(2/m-1)}$$

where,

'c' represents the number of cluster center.

'm' is the fuzziness index m€[1,∞].

'μ_{ij}' represents the membership of i^{th} data to j^{th}cluster center.

'd_{ij}' represents the Euclidean distance between i^{th} data and j^{th}cluster center.

'd_{ik}' represents the Euclidean distance between i^{th} data and k^{th}cluster center.

3) Compute the fuzzy centers 'v_j' using:

$$v_j = \left(\sum_{i=1}^{n}(\mu_{ij})^m x_i\right)\Big/\left(\sum_{i=1}^{n}(\mu_{ij})^m\right), \forall j = 1, 2, \ldots . c$$

where, 'n' is the number of data points.

'v_j' represents the j^{th} cluster center.

4) Repeat step 2 and 3 until the minimum 'J' value is achieved or $||U^{(k+1)} - U^{(k)}|| < \beta$.
Where, 'k' is the iteration step.

'β' is the termination criterion between [0,1].

'$U = (\mu_{ij})_{n*c}$'is the fuzzy membership matrix.

'J' is the objective function.

K-MEDOIDS CLUSTERING ALGORITHM

Algorithmic steps for k-medoids clustering:

The most common realization of k-medoid clustering is the Partitioning AroundMedoids (PAM) algorithm and is as follows:

1. Initialize: randomly select *k* of the *n* data points as the medoids.

2. Associate each data point to the closest medoid.
   For each medoid*m*
   1. For each non-medoid data point *o*.
   2. Swap *m* and *o* and compute the total cost of the configuration.
3. Select the configuration with the lowest cost.
4. Repeat steps 2 to 5 until there is no change in the medoid

INTREPRATATION OF RESULTS

This Chapter introduces the evaluation metric along with the results obtained from all the data sets in Fuzzy c-means, K-medoids as well as K-means. A comparison is drawn between the results in the table.

The main Window of the project includes a text field to display the path of the selected dataset and two buttons, one button to browse the file and another button to perform DDoS Attack Detection. A pop up window appears after clicking the browse button, here we select any one dataset to test.

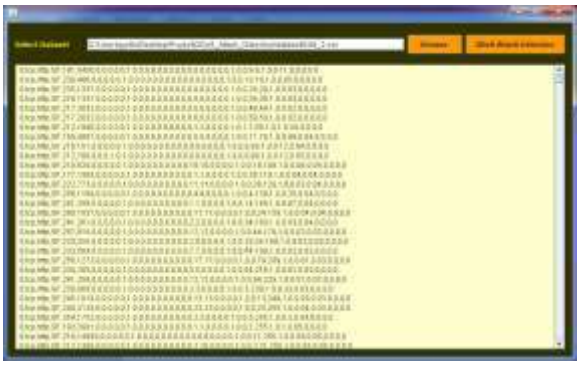

Figure 7: File contents Displayed

After selected the input dataset, the contents of it are read and its path is displayed in the text field and the contents are shown in the text area.

After clicking the DDoS Attack Detection button the second window opens. Here we show the working of 3 different algorithms and display their accuracy level and false alarm rate.

The rectified and injured packets are displayed in two separate text areas and the False Alarm Rate and Accuracy level of k-means,Fuzzy c-means,k-medoids algorithms is calculated and displayed in the right side.

Figure 8: Algorithms Executed

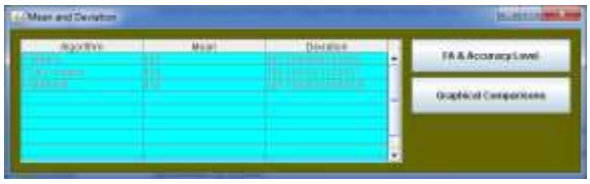

Figure 9: Mean and Deviation window

When the FA and Accuracy Level button is clicked (shown in Figure 9) new window opens which displays the False Alarm Rate and Accuracy Level of the three different algorithms in a table.

When you click the False Alarm Rate button in the Graph window it displays the False Alarm Rate of three algorithms in a Column graph .

When you click the Accuracy Level button in the Graph window it displays the Accuracy Level of three algorithms in a Column graph.

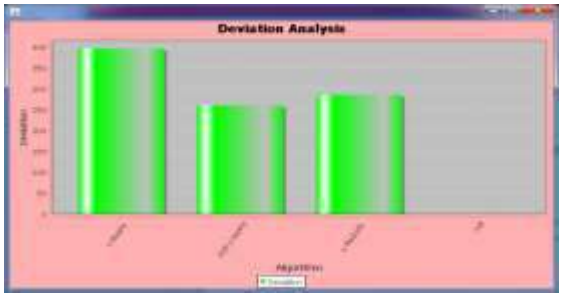

Figure 10: Plot of Deviation

When you click the Deviationbutton in the Graph windowit displays the Deviationof three algorithms in a Column graph (Figure 10).

Result:

The result of the K-means, Fuzzy c-means and K-medoids execution on the selected data set gives the total number of Injured/Attacked Data in that particular Data set. Graphs and Reports are generated of Accuracy Level and False Alarm Rate for the different algorithm. These Reports and Graphs show a comparative study between the different algorithms. We can see that K-medoids is able to detect most of the injured packets and is more efficient compared to Fuzzy c-means and K-means Algorithms. Accuracy level of K-medoids is greater than Fuzzy c-means, which in turn is greater than K-means.

CONCLUSION

Three different clustering algorithms i.e. K-Means clustering algorithm, Fuzzy-c means clustering algorithm and K-medoids clustering algorithm are implemented for the detection of DDoS attack detection. The performance of each algorithm is

been analyzed by processing the data sets. The accuracy and false alarm rate of each algorithm is calculated and displayed in the form of graphs.

FUTURE ENHANCEMENT

The following are few of the Enhancements that can be focused on:

1. The DDoS detection has been made using the basic clustering algorithm. Hence, there is a need to evaluate the performance of advanced clustering algorithms.
2. The serialized execution of the clustering algorithms needs to be parallelized for the improvement in the performance of the project.
3. The detected data needs to be further analyzed for identifying the intruders.

REFERENCES

1. *WesamBhaya and Mehdi EbadyManaa, "A Proactive DDoS Attack Detection Approach Using Data Mining Cluster Analysis", Journal of Next Generation Information Technology (JNIT), vol.5, No.4, November 2014.*
2. *SaravananKumaraswamy and Dr. R. Asokan, "Distributed Denial of Service (DDoS) Attacks Detection Mechanism", in International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), vol.1, No.5 December 2011.*
3. *Longlong Li, Qin Chen, Shuiming Chi, Xiaohang Liu, "Unsupervised Intrusion Detection Based on FCM and Vote Mechanism", Information Technology Journal, Science Alert, vol.13, No.1, pp.133-139, 2014.*
4. *RuiZhong, GuangxueYue, "DDoS Detection System Based on Data Mining", in Proceedings of the Second International Symposium on Networking and Network Security (ISNNS'10), pp.062-065, 2010.*
5. *Laura Feinstein, Dan Schnackenberg, RavindraBalupari, Darrell Kindred, "Statistical Approaches to DDoS Attack Detection and Response", in Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX'03), vol.1, pp.303-314,2003.*
6. *The CAIDA UCSD "DDoS Attack 2007" Dataset http://www.caida.org/data/passive/ddos-20070804_dataset.xml*
7. *The CAIDA UCSD Anonymized Internet Traces 2008 <insert dates used here>http://www.caida.org/data/passive/passive_2008_dataset.xml*