إقرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Detecting DDoS Attack Using A Multilayer Data Mining techniques

أقر بأن ما اشتملت عليه هذه الرسالة إنما هي نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وإن هذه الرسالة ككل، أو أي جزء منها لم يقدم من قبل لنيل درجة أو لقب علمي أو بحثي لدى أية مؤسسة تعليمية أو بحثية أخرى.

DECLARATION

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification

Student's name :	اسم الطالب: هبة البلتاجي
Signature	التوقيع: هبة
Date:	التاريخ: 7-8-2015

The Islamic University of Gaza
Graduate Studies
Faculty of Information Technology

Detecting DDoS Attack Using A Multilayer Data Mining techniques

Prepared by

Heba S. Albiltaje

Supervised by

Dr. Tawfiq S. Barhoom

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science In Information Technology

2015-1436H

بِسْمِ اللهِ الرَّحْمَنِ الرَّحِيمِ

الجامعة الإسلامية – غزة
The Islamic University - Gaza

الجامعة الإسلامية غزة
The Islamic University Gaza

مكتب نائب الرئيس للبحث العلمي والدراسات العليا   هاتف داخلي: 1150

الرقم ج س غ/35/ Ref

التاريخ 2015/02/17م Date

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحثة/ **هبة صقر محمود البلتاجي** لنيل درجة الماجستير في **كلية *تكنولوجيا المعلومات*** **برنامج تكنولوجيا المعلومات** وموضوعها:

كشف هجمات حجب الخدمة الموزعة باستخدام تقنية الطبقات المتعددة في تنقيب البيانات

Detecting DDos Attack Using a Multilayer Data Mining Techniques

وبعد المناقشة التي تمت اليوم الثلاثاء 28 ربيع الآخر 1436هـ، الموافق 2015/02/17م الساعة العاشرة صباحاً بمبنى اللحيدان، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

د. توفيق سليمان برهوم	**مشرفاً و رئيساً**	....
د. إياد محمد الأغا	**مناقشاً داخلياً**	الأغا ....
د. تامر سعد فطاير	**مناقشاً خارجياً**	....

وبعد المداولة أوصت اللجنة بمنح الباحثة درجة الماجستير في **كلية *تكنولوجيا المعلومات*/ برنامج تكنولوجيا المعلومات.**

واللجنة إذ تمنحها هذه الدرجة فإنها توصيها بتقوى الله ولزوم طاعته وأن تسخر علمها في خدمة دينها ووطنها.

والله ولي التوفيق،،،

مساعد نائب الرئيس للبحث العلمي والدراسات العليا

شئون البحث العلمي والدراسات العليا
الجامعة الإسلامية غزة
The Islamic University Gaza
Research & Graduate Affairs

أ.د. فؤاد علي العاجز

بسم الله الرحمن الرحيم

Dedication

To my lovely Mather and Father,

Who always picked me up on time

And encouraged me to go on every time

Especially this one

To my kindness Grandfather

To my sisters and Brothers

To all my Best Friends

Acknowledgements

First of all, I would thank Allah for guiding me to complete this research to the fullest

I want to thank everyone help and participated in making this research

This work would not have been possible without the constant encouragement and support I received from Dr. Tawfiq S. Barhoom, my advisor and mentor. I would like to express my deep and sincere gratitude to him. His understanding and personal guidance have provided a good basis for the present thesis.
Special thanks to my friends Mrs. Hanaa Qeshta to all for here provide assistance and guidance.

Table of Contents

LIST OF FIGURES

LIST OF TABELS

LIST OF ABBREVIATIONS

IDS	Intrusion Detection System
HIDS	Host-Based Intrusion Detection System
NIDS	Network-Based Intrusion Detection System
DoS	Denial of service Attack
DDoS	Distributed Denial of service Attack
MCS	Multi-cluster system
AID	Anomaly Detection
MID	Misuse Detection
KM	K-mean
KFM	K-fast mean
KD	K-mididod
MCDDM	**M**ulti **C**luster **D**DoS **D**etection **M**ethod

طريقة لمنع هجمات الحرمان الموزعة باستخدام آليات التنقيب عن البيانات

تعتبر استمرارية وجود المعلومة أو الخدمة من أهم عناصر أمن المعلومات بالإضافة إلى التكاملية والسرية بالنسبة للمصارف في حقل البنوك الإلكترونية او الخدمات المصرفية الإلكترونية بجانب السرية أو الموثوقية و التكاملية .

يستخدم قراصنة الإنترنت أساليب عديدة لاختراق أو تعطيل شبكات الحاسوب المستهدفة، ومن أبرز أساليب القراصنة لتعطيل شبكات الحاسوب ما يعرف بحجب الخدمة أو حجب الخدمة الموزع وهي هجمات تستهدف عادة مؤسسات حكومية أو شركات كبرى كالبنوك مثلا، ومبدأ هذا الأسلوب يتلخص في أن المهاجم يقوم بزرع وكيل في جهاز الضحية او موزعه ومن ثم تقوم هذه الوكلاء بإغراق الأجهزة المزودة بسيل من الطلبات والأوامر التي تفوق قدرة الجهاز المزود على المعالجة .

في الآونة الأخير اهتم الباحثون بإيجاد الحلول لمنع هذه الهجمات اعتمادا على آليات تنقيب البيانات فقدموا العديد من الأبحاث في هذا المجال ومن أهم أنواع التنقيب عن البيانات التنقيب الاستشرافي والتنقيب الوصفي ومن أهم طرق التصنيف الوصفي هو التجميع ويقصد به تجميع البيانات في مجموعات بناء على خصائصها مع عدم العلم مسبقاً عن الخصائص التي سيتم التجميع على اساسها.

في هذا البحث قدمنا طريقة لمنع هجمات الحرمان من الخدمة الموزعة باستخدام العديد من طرق التجميع (Km, KD, KFM) بشكل التجميع المتعدد لكي تكون قادرة على الكشف عن الهجمات الموزعة الجديدة والغير معروفة . وقد أوضحت نتائجنا أن طريقتنا حصلت على نتائج أعلى من نتائج طرق التجميع الاخرى حينما استخدمت بشكل فردي , حيث حصلت على davies_bouldin index (-0.666) .

الكلمات المفتاحية : نظام كشف التسلل , تنقيب البيانات , أمن المعلومات , التجميع المتعدد, هجمات الحرمان من الخدمة ,هجمات الحرمان الموزعة

Abstract

Availability is one of the three main components of computer security, along with confidentiality and integrity. One of the major threats to network security is Denial of Service (DDoS),which is a relatively simple, but very powerful technique to attack internet resources as well as system resources. Distributed multiple agents consume some critical resources at the target within the short time and deny the service to legitimate clients .

Most current network intrusion detection systems employ signature-based methods or supervised-based methods which rely on labelled training data. This training data is typically expensive to produce, these methods have difficulty in detecting new types of attack, Using unsupervised anomaly detection techniques , the system can be trained with unlabelled data and is capable of detecting previously “unseen" attacks.

In this research we multi-clustering method using data mining techniques by combination of clustering method (K-Mean(Km) ,K-Medoid(KD),K-Fast Mean(KFM)) as a multi clustering to be able for detecting anew DDoS attacks from unlabelled dataset depend on unsupervised behavior-anomaly detection approach, Davies_Bouldin index(DB) is used to evaluate the proposed method . The results show that the proposed method has lower davies_bouldin index.

Keywords: Distributed Denial of service attacks(DDoS), Multi Clustering, Data Mining , unsupervised anomaly detection(UAD)

Chapter 1:

Introduction

Chapter 1:Introduction

As Internet is increasingly being used in almost every aspect of our lives, it is becoming a critical resource whose disruption has serious implications. Blocking availability of an Internet service may imply large financial losses, as in the case of an attack that prevented users from having steady connectivity to major ecommerce Web sites such as Yahoo, Amazon, eBay, E*Trade [1]. It may also imply threat to public safety, as in the case of taking down of Houston port system in Texas [2] or national security, as in the case of White House Web site becoming the target of Code Red worm attack [3]. Such attacks that aimed at blocking availability of computer systems or services are generally referred to as denial of service (DoS) attacks DoS attack has gradually developed into a method of using various attack paths as DDoS (distributed denial-of-service) attack and of attacking the entire network to which target belongs.

Distributed Denial of service attacks : is a DoS attack utilizing multiple distributed attack sources. the attackers use a large number of controlled zombies distributed in different locations to launch a large number of DoS attacks against a single target or multiple targets [25].

Multi Cluster DDoS Detection Method "MCDDM": is researcher proposed method which a multi-clustering method using data mining techniques by combination of clustering method (K-Mean(Km) ,K-Medoid (KD),K-Fast Mean(KFM)) as a multi clustering to be able for detecting anew DDoS attacks

In recent years, Many researchers have been developing to detect this kind of attack which results in not only the advance of network security system, but also constantly attack tools improved adept attackers in order to evade these security mechanisms[26] Most of this approaches is supervised depending on labelled dataset, labelled data or purely normal data is not readily available since it is time consuming and expensive to manually classify it. Purely normal data is also very hard to obtain in practice, since it is very hard to guarantee that there are no intrusions when we are collecting network traffic[41], We try to counter this drawback in on our research we proposed ”MCDDM” unsupervised multi-clustering method DDoS detection based on data mining as an efficient way to improve the security of networks ,we use the davies_bouldin index to evaluate the proposed method.

1.2 Goals of DDoS Detection systems

- Protecting networks from DDoS and decreasing the effect of damage caused by This attacks.
- To overcome the shortage of traditional approach of DDoS detection.
- To enhancement of standalone clustering that cannot provide acceptable accuracies in real-world deployments
- To introduce an adaptive detection method that has the ability of detecting DDoS attacks in early stages.
- Introducing a new way of integration of clustering methods by combining them in order to achieve an acceptable accuracy in real world.

1.3 Network Security and DDoS Detection

Distributed Denial of Service attacks (DDoS) overwhelm network resources with useless or harmful packets and prevent normal users from accessing these network resources. These attacks jeopardize the confidentiality, privacy and integrity of information on the internet

Network security is one of the most important issues that can be considered by commercial organizations to protect its information from malicious risk. The problems of detection malicious traffics have been widely studied and still as a hot research topic in the recent decades. Many researches have been designed and implemented an Intrusion Detection System (IDS) to analyses, detect and prevent the DDoS activities .

1.4 Research Motivation

Information has become an organization’s most precious asset upon which they have increasingly become dependent. The widespread use of internet and e-commerce has increased the necessity of protecting the system as they can .

DDoS attacks have become a hot research topic, because they can lead to a loss of confidence and privacy and could lead to illegal actions taken against an organization.

Data mining approach comes to help into DDoS detection, In our research we use multi-clustering approach to distinguishing attack traffic from the common legitimate traffic with high accuracy .

1.5 Problem Statement

In order to reduce the risk of DDoS, a variety of defiance magnesium have been proposed, but the problem caused by new DDoS attack is still not counter.

This research propose new unsupervised multi-layer method for DDoS attack detection by clustering technique .

1.6 Research Objectives

1.6.1 Main Objective

The main objective of this research is to propose A multi-layer system for DDoS attack detection. The proposed solution tries to counter new DDoS attack by using unsupervised multi-clustering methods ".

1.6.2 Specific Objectives

These are specific objectives which could be extracted from the main objective:

- Survey and examine the current techniques and solutions of prevent DDoS and gain further knowledge through the understanding of these techniques.
- Collect the proposed method requirements such as Wirshark as .pcap file reader .
- Design the proposed method's architecture .
- Applying DDoS detection method based on anomaly-behavior detection using unsupervised learning machine technique and by combination of multi- clustering in data mining.
- Testing "MCDDM " method by a various data the percentage of the DDoS attacks on it is, to observe the system's ability to detect the attacks, so that we can prove that this method is able to detect DDoS attacks.
- Evaluate "MCDDM " method by using davies_bouldin index .
- Reduce the davies_bouldin index to achieve best performance for "MCDDM " method .

1.7 Research Scope and Limitation

This research aims to propose new multi-layer clustering DDoS detection method which is able to detect DDoS with lower davies_bouldin index. This work is applied with some limitations and assumption such as:

1.7.1 Research scope

- The proposed method based on network intrusion detection system (NIDS).
- The datasets used in this research is combination of the CAIDA UCSD "DDoS Attack 2012" Dataset [15],and "anonymized Internet trace2013" Dataset [31].
- Using combination of multi clusters techniques in data mining to detect DDoS.
- The cases of experiment is on datasets merged by x% of attacks .

1.7.1 Research Limitation

- The proposed method will be limited to using behavior anomaly detection technique to build a multi-layer DDoS detection system .
- The proposed method is limited for unsupervised learning .
- The proposed method is experimented by passive dataset.
- The proposed method will be evaluated by the davies_bouldin index.

1.8 Significance of the research

- Add a significant contribution to scientific research in the field of finding effective solutions in DDoS detection.
- Helping concerned people working in various DDoS detection domains to get a better prediction for clustering.
- Using more cluster techniques as combination to reduce davies_bouldin index.
- Approve that unsupervised anomaly detection is efficient in DDoS detection domain.

1.10 Outline of the Thesis

This dissertation has been divided into six major chapters, which are structured around the objectives of the research. The dissertation is organized as follows:

Chapter 2, Presents Literature Review of DDoS and DDoS detection approaches.

Also, this chapter presents details about machine learning and data mining techniques, clustering methods, and clustering algorithms used on multi-clustering DDoS detection method.

Chapter 3, Presents some related work of DDoS detection, and highlights its main shortages which are to be avoided and solved in our work.

Chapter 4, Includes the methodology steps and the architecture of the multi-clustering DDoS detection method.. An explanation about the data sets used in the experiments, preprocessing of these data set, and the experiment cases is included as well. Also, this chapter presents the baseline experiments to choose the optimal clustering algorithms, analyze the experimental results. Also discussion for each set experiments.

Chapter 5, Will draw the conclusion and summarize the research achievement of experiments and suggests future work.

CHAPTER 2:

Theoretical Background

CHAPTER 2: **Theoretical Background**

In this chapter, we will identify DDoS, types of DDoS, characteristics of DDos, DDoS strategy. Then we will describe and compare various approaches of DDoS detection. Finally, we will explain the use of machine learning and data mining, especially clustering techniques, and clarify their effectiveness in the detection of DDoS attacks.

2.1. Distributed Denial of service attacks :

Distributed denial of service(DDoS) attacks which are intended attempts to stop legitimate users from accessing a specific network resource, have been known to the network research community since the early 1980s. In the summer of 1999, the Computer Incident Advisory Capability (CIAC) reported the first Distributed DoS (DDoS) attack incident [32], and most of the DoS attacks since then have been distributed in nature. In the next sub sections we will recognize DDoS attacks closely

2.1.1 DDoS attacks Definition

DDoS attacks make the resources of host occupied largely via sending many malicious packets, which results in the failure of normal network services. DDoS attack the target host through constructing a lot of illegal packets, this kind of attacks changed traditional peer to peer attack mode and used distributed attack mode instead that causes the extent of hosts participating in attack wider, data flow generated by attack present irregular status. All of this make DDoS attacks launched easily, prevented and tracked difficultly and so forth. So far DDoS attacks have become one of the essential threats to network security.[14][25]

2.1.2 Types of DDoS attack

There is no general DDoS classification method because there is no theory of DDoS attack. Some researchers are classified DDoS attack in a broadly scheme as below

- **Attack on Bandwidth**

  DDoS attacks of this type send mass junk data messages to cause an overload, leading to the depletion of network bandwidth or equipment resources. Often the attacked routers, servers and firewalls processing resources are limited. Overload attacks lead to their failure in handling normal legal access, resulting in either a sharp decline in the quality of service or a complete denial of service - in either case it means your customers, users, etc cannot access the systems they need to.

- **Attack of Host Resource**

the most common forms are traffic flooding attacks, which send large number of seemingly legitimate TCP, UDP, ICPM packets to target hosts; some attacks may also evade detection system monitoring through source address forging technology. Legitimate requests get lost in noise. These attacks can also be devastating if combined with other illegal activity, such as malware exploitation to cause information leakage: while you are fighting off the DDoS, your sensitive data is slipping out the backdoor.

- **Attack on System/Application Weakness**

  attacks of this type often send application-layer data messages according to business-specific features (using seemingly legit functions, like a DB call, etc), resulting in the depletion of certain resources in the application layer (such as the number of users, connections, etc.) and the system's services are no longer available. Such attacks are usually not particularly large in volume; but even such low-rate traffic can often lead to a serious declination or even paralysis of business system performance.

2.1.3 DDoS Strategy

The DoS attack strategy is depend on sending many harmful packets to the victim system or device this cause overloaded and resource consuming .

The DDoS strategy is depends on start generating as many packets as they can toward the victim. A large number of agents enable the attacker to overload resources of very highly provisioned victims, it is hard to prevent DDoS attacks because it mixing up of legitimate and illegal traffic Figuer 1 show the DDoS strategy .

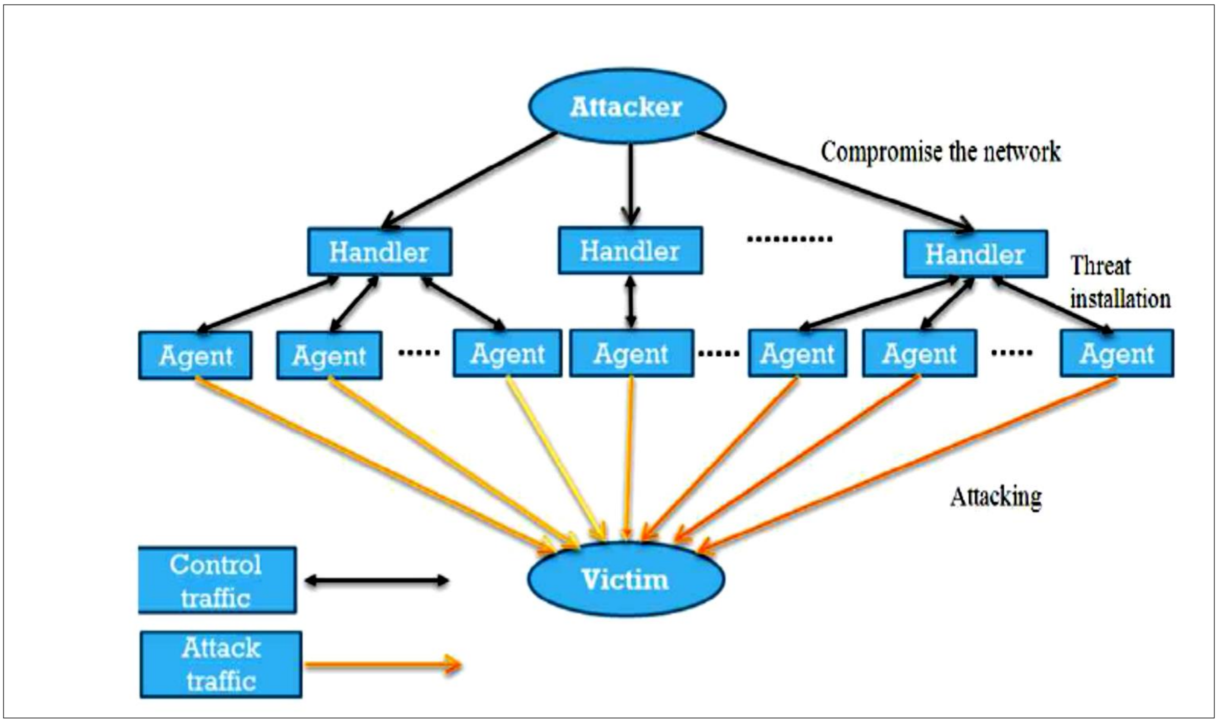


Figure. 2.1 Architecture of Distributed Denial of Service (DDoS) attack[25]

2.1.4 Characteristics of DDoS attacks

The characteristics of DDoS Attack are as follows after the analysis of it:

- Abnormal traffic. A lot of useless packets transmitted by the attacker in order to occupy the resources of the victims(bandwidth or host resources).Such a large number of packets would cause the victims system-halted and fail to provide external services.[30]
- Most DDoS attacks take the three times handshake mechanism and use “SYN” status flag to send the victim connection requests . However, this does not mean to build a real connection, which makes the victim maintain a great deal of half-opened connection and consume the resources of the victims.[30]
- The attacker makes use of one of the characters of TCP/IP protocol that some non-compliant packets could be used so as to launch DDoS attack.[30]

2.2 Generic architecture of DDoS attack defense mechanisms

Based on the locality of deployment, DDoS defense schemes can be divided into three classes: victim end, source-end, and intermediate router defense mechanisms.

2.2.1. Victim-end defense mechanism

Victim-end detection approaches are generally employed in the routers of victim networks, i.e., networks providing critical Web services. A generic architecture of such schemes is shown in Figure 2.2. Here the detection engine is used to detect intrusion either online or offline, using either misuse based intrusion detection or anomaly based intrusion detection. The reference data stores information about known intrusion signatures or profiles of normal behavior. This information is updated by the processing elements as new knowledge about the observed behavior becomes available.[35]

The security manager often updates the stored intrusion signatures and also checks for other critical events such as false alarms. The processing element frequently stores intermediate results in the configuration data.

Detecting DDoS attacks in victim routers is relatively easy because of the high rate of resource consumption. It is also the most practically applicable type of defense scheme as Web servers providing critical services always try to secure their resources

for legitimate users. But the problem with these approaches is that, during DDoS attacks, victim resources, e.g., network bandwidth, often gets overwhelmed and these approaches cannot stop the flow beyond victim routers. Another important disadvantage is that, these approaches detect the attack only after it reaches the victim and detecting an attack when legitimate clients have already been denied is not useful[36].

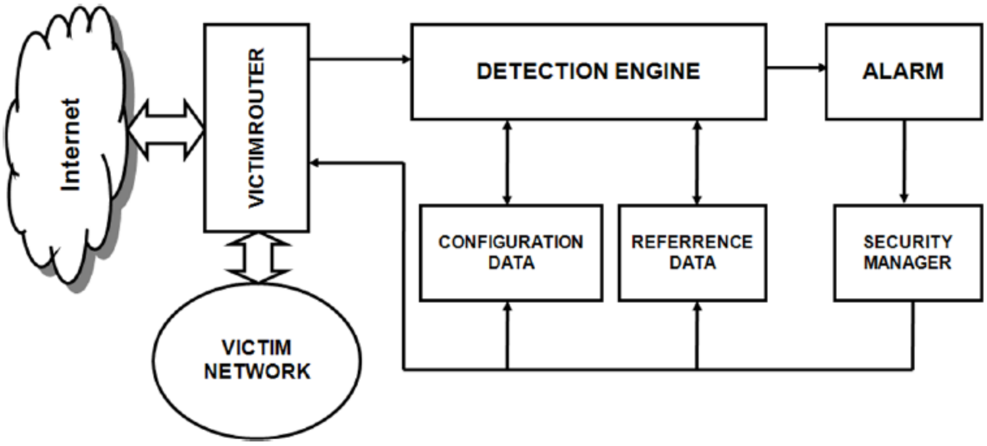


Figure2.2. Generic architecture for victim-end DDoS defense mechanism[36]

2.2.2. Source-end defense mechanism

As DDoS defense is pushed from the victim to the source, detection capability become less.

A source-end defense system can no longer easily observe the effect of incoming traffic on the victim. The defense system has difficulties in detecting anomalies. On the other hand, response effectiveness increases with proximity to the sources.

A small attack volume enables an effective response as it is unlikely to overwhelm the defense system. The small volume and degree of aggregation also facilitates complex profiling that, in turn, minimizes the damage. A generic architecture of such schemes is shown in Figure 2.3.

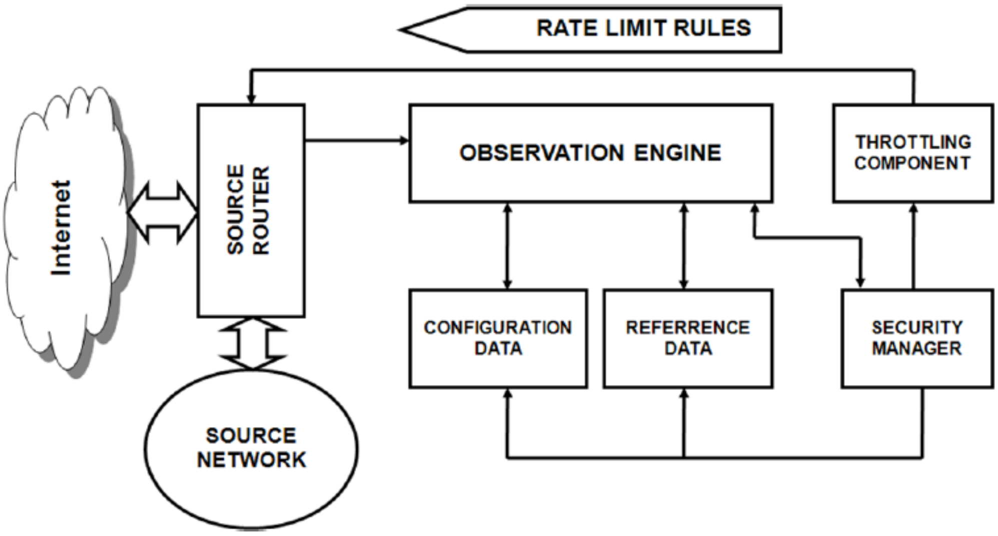


Figure2.3. Generic architecture for Source-end DDoS defense mechanism[36]

2.2.3. Intermediate network defense mechanism

The intermediate network defense scheme balances the trade-offs between detection accuracy and attack bandwidth consumption, the main issues in source-end and victim-end detection approaches. Figure2.4 shows a generic architecture of the intermediate network defense scheme, one that can be employed in any network

router. Such a scheme is generally collaborative in nature and the routers share their observations with other routers. Like a source-end scheme, these schemes also impose rate limits on connections passing by the router after comparing with stored normal profiles.[35]

Detection and traceback of attack sources are easy in this approach due to collaborative operation. Routers can form an overlay mesh to share their observations

[18]. The main difficulty with this approach is deploy ability. To achieve full detection accuracy, all routers on the Internet will have to employ this detection scheme, because unavailability of this scheme in only a few routers may cause failure to the detection and

traceback process. Obviously, full practical implementation of this scheme is extremely difficult reconfiguring all the routers on the Internet.[36]

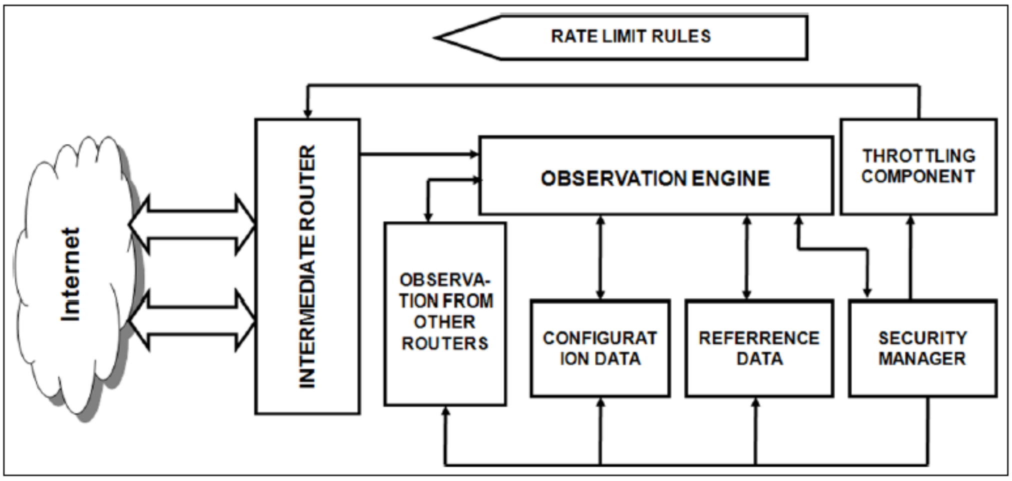


Figure2.4. Generic architecture for Intermediate-end DDoS defense mechanism[36]

2.3 Intrusion Detection Systems based on Data mining

The Purpose of Intrusion Detection Systems (IDS) is to monitor network in order to detect misuse or abnormal behavior, that is statistically analyzing input data (e.g., network traffic) for the purpose of detecting whether an intrusion has occurred or is not occurring [34]. The types of IDS can be divided into two categories: network based (NIDS) and host based (HIDS). Network based (NIDS) tries to detect any abnormal behavior of the system by analyzing the network traffic. Host based (HIDS) to act as the last line of defense, which detect intrusions by analyzing the events on the local system while the IDS is running. The host based IDSs classified into two categories: anomaly detection(AID)and misuse detection(MID).[35] In the misuse detection approach is signature-based detection systems are based on known Database of signatures . MID detection detect what is known. It does not detect any unknown signatures.

AID Supervised anomaly detection depends learning on some known feature and try to learned the system to detect anew abnormal behavior depending on systems learning

The supervised anomaly detection detects what is different from what is known[8] It requires strong knowledge about what is seen “normally” that is about the basic behavior. It is difficult to maintain up to date normal operation profile.

Unsupervised Anomaly Detection [9] uses data mining techniques to extract patterns and uncover similar structures “hidden” in unlabeled traffic of unknown nature. The unsupervised detection of network attacks is based on clustering techniques and outliers detection.

2.4 Data Mining

It is considered as one of the applications of supervised machine learning, and it plays an important role in the process of retrieving the lost information. Data mining refer to the analysis of large quantities of data that are stored in computers [10], and is defined as knowledge discovery, which is the process of extracting useful patterns from large volumes of data using special algorithms [11][12]. Many terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging [13]. Data Mining is essentially a process of data drive extraction of not so obvious but useful information from large databases that is interactive and iterative. Knowledge discovery as a process consists of an iterative sequence of the following steps:

1) **Data Cleaning:** is removing the noise and inconsistent data.
2) **Data Integration:** where multiple data sources may be combined. These sources may include multiple databases, data cubes, or flat files
3) **Data Selection:** where data relevant to the analysis task are retrieved from the database. So, irrelevant, weakly relevant or redundant attributes may be detected and removed.
4) **Data Transformation:** where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance
5) **Data Mining:** an essential process where intelligent methods are applied on data to extract data patterns for decision making.
6) **Pattern Evaluation:** to identify the truly interesting patterns based on some interestingness measures. A pattern consider interesting if it is: Valid, Novel, Actionable, Understandable.
7) **Knowledge Presentation:** is the framework that converts a large amount of data into a particular data or procedure that human being can figure out based on an intention. In Knowledge representation visualization tools and knowledge representation techniques are used to present the mined knowledge to the user.

Figure 2.5, illustrates data mining as a step in the process of knowledge discovery.

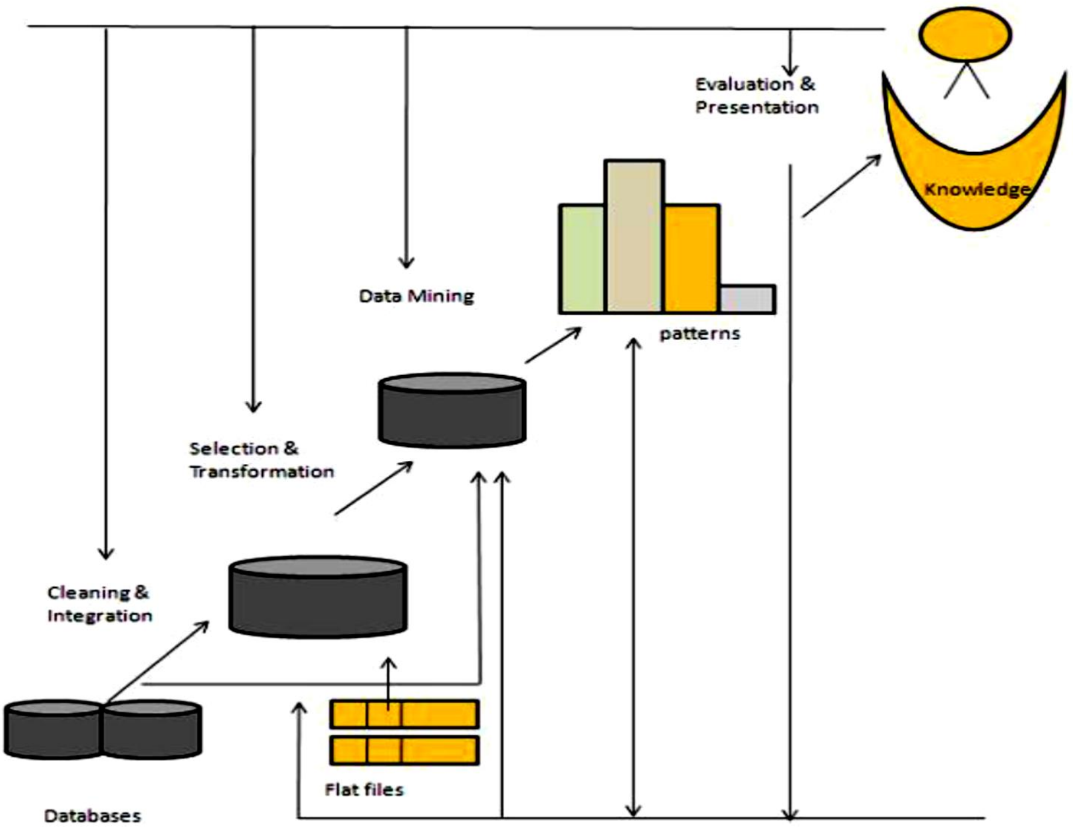


Figure 2.5: Data mining as a step in the process of knowledge discovery [48].

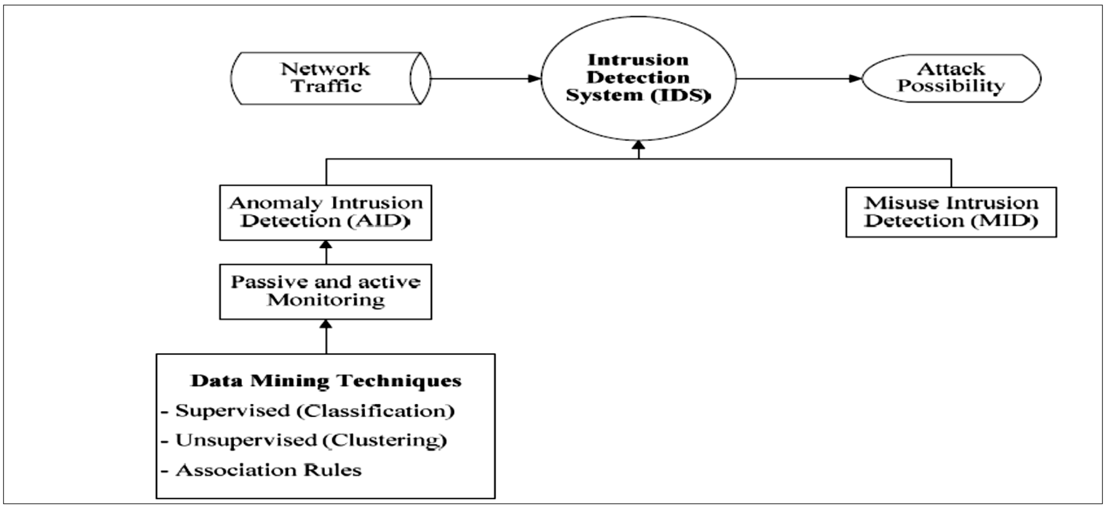
Data Mining functionalities are used to specify the type of patterns to be found in the data mining tasks. In general data mining tasks can be classified into two main categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data. Predictive mining tasks perform inferences on the current data in order to make predictions [13]. Most of data mining tasks can be one or combination of the following:

1) **Classification:** used for predictive mining tasks. This methods is intended for learning different functions that map each item of the selected data into one of a predefined set of classes. Given the set of predefined classes, a number of attributes, and a “learning (or training) set,” the classification methods can automatically predict the class of other unclassified data of the learning set[13].
2) **Prediction:** used for predictive mining tasks**.** Analysis is related to regression techniques. The key idea of prediction analysis is to discover the relationship between the dependent and independent variables. For example, by using historical data from both sales and profit, either linear or nonlinear regression techniques can

produce a fitted regression curve that can be used for profit prediction in the future [4].

3) **Association Rules:** used for descriptive mining tasks. It aims to find out the relationship among valuables in database, and produce a set of rules describing the set of features that are strongly related to each other's, so that the relationship of a particular item in a data transaction on other items in the same transaction is used to predict patterns [13].
4) **Clustering:** used for descriptive mining tasks. It is unsupervised, and does not require a learning set. It shares a common methodological ground with Classification. It ungrouped data and uses automatic techniques to put this data into groups [4]. the data points that belong to one cluster are more similar to each other than to data points belonging to different cluster.
5) **Outlier Analysis:** used for predictive mining tasks**.** Discovers data points that are significantly different than the rest of the data. Such points are known as exceptions or surprises. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable. So that very important identify the outliers [13].

Data mining techniques play major roles in detection the malicious on the network traffic such as DDoS attacks. IDS's can be approached by data mining machine learning techniques **Figuer**.2.6 show the IDS strategy with data mining


Network Traffic
Intrusion Detection System (IDS)
Attack Possibility
Anomaly Intrusion Detection (AID)
Misuse Intrusion Detection (MID)
Passive and active Monitoring
Data Mining Techniques
- Supervised (Classification)
- Unsupervised (Clustering)
- Association Rules

Figuer 2.6 IDS detection strategies with data mining[13]

2.4.1 Clustering

Clustering is a way of grouping together data samples that are similar in some way according to some criteria that is pick, it’s a form of unsupervised learning that is generally we don’t kwon how the data should be grouped together So, it’s a method of data exploration , a way of looking for patterns or structure in the data that are of interest. [4]

Clustering algorithms are classified in five main categories [25].

1) The hierarchical clustering are methods start with each point in its own cluster. Clusters are combined based on their closeness, using one of many possible definitions of “close.”
2) The partitioning clustering: Initial points are chosen randomly or in some order and each point in a state space is assigned to the cluster into which it best fits based on similarity distance.
3) The density-based methods: are developed based on the notation of density. The key idea is to continue growing the given cluster as long as the density (the number of objects or data points) in the “neighborhood” exceeds some threshold.
4) The Grid-based methods are performed in a fast processing time, where the object space quantizes into a finite number of cells that form a grid structure (on the quantized space).

We will describes some of The partitioning clustering algorithms in order to be used in our research such as K-mean ,K-mididod , K-fast mean

2.4.1.1 k-mean

One of cluster algorithm in data mining that preset a data set which contain n data object and k cluster that needs to create. The main idea of k-means algorithm is to split the data object set into k cluster($k \le n$) that could make a standard measure function optimization and make high similarity of data object in the same cluster.

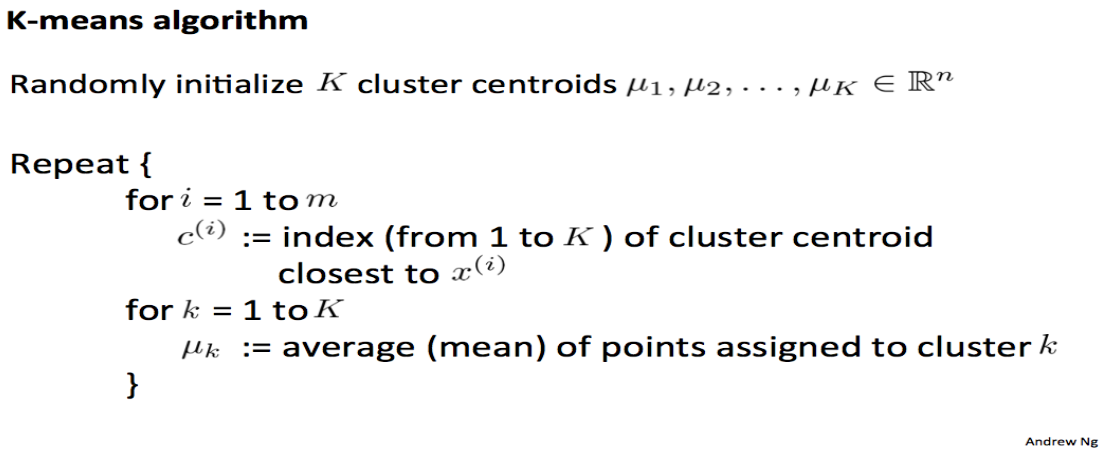
The particular algorithm procedure is as follows:

- **Step1** Select at random K initial cluster center $k_1, k_2, k_3, \ldots, k_n$ in m time window

- **Step2** Calculate the distance between each network traffic data i x and initial cluster center through Dj = min{|| xi − Kv ||} , the sample point that is the nearest to cluster center would be assigned to the cluster whose center is v K
- **Step3** Move every w K to its cluster center and recalculate the cluster center according to new data added in cluster. Then calculate the deviation including sample value in each cluster domain through formula:

$$D = \sum_{i=1}^{n} [\min_{r=1,\dots,k} d(x_i, K_r)^2] \cdot$$

- **Step4** The repetitive execution of step3 and step4 until the convergence of D value and all the cluster center will not move. After that the cluster center is the traffic mean value in different time window.



K-means algorithm

```
Randomly initialize $K$ cluster centroids $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

Repeat {
    for $i$ = 1 to $m$
        $c^{(i)}$ := index (from 1 to $K$ ) of cluster centroid
                closest to $x^{(i)}$
    for $k$ = 1 to $K$
        $\mu_k$ := average (mean) of points assigned to cluster $k$
}
```

Andrew Ng

2.4.1.2 K- medoid

The *k*-medoids algorithm is a clustering algorithm related to the *k*-means algorithm and the medoid shift algorithm. Both the *k*-means and *k*-medoids algorithms are partitioned (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the *k*-means algorithm, *k*-medoids chooses data points as centers (medoids or exemplars) and works with an arbitrary matrix of distances between data points instead of l_2. This method was proposed in 1987[1] for the work with l_1 norm and other distances.

k-medoid is a classical partitioning technique of clustering that clusters the data set of *n* objects into *k* clusters known a priori. A useful tool for determining *k* is the silhouette.

It is more robust to noise and outliers as compared to *k*-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances.

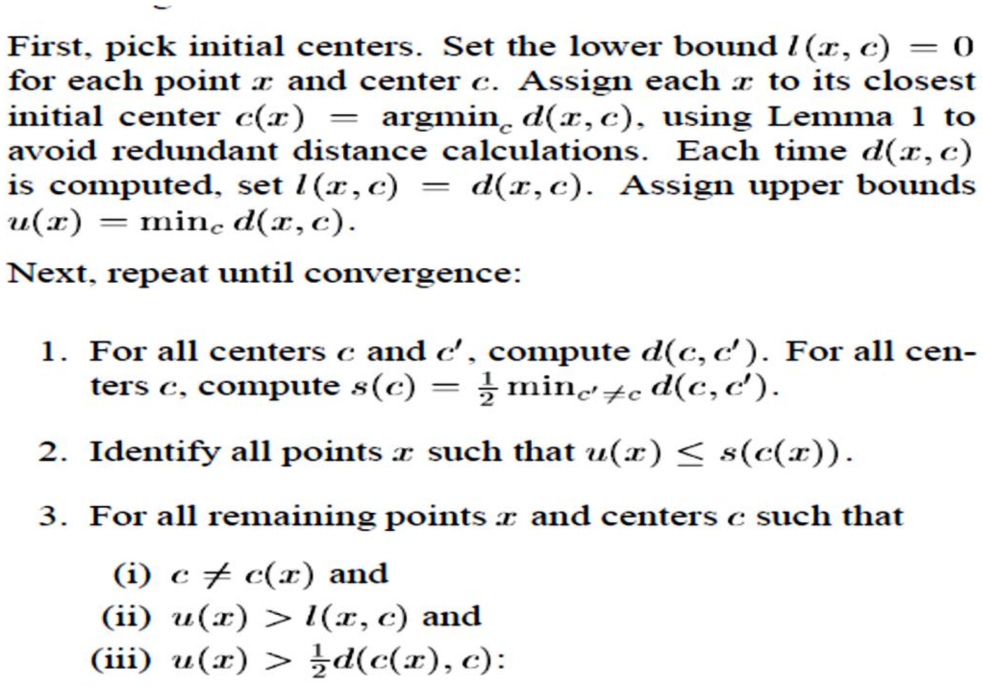
A medoid can be defined as the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal. i.e. it is a most centrally located point in the cluster.

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

1. Initialize: randomly select (without replacement) *k* of the *n* data points as the medoids
2. Associate each data point to the closest medoid. ("closest" here is defined using any valid distance metric, most commonly Euclidean distance, Manhattan distance or Minkowski distance)
3. For each medoid *m*
    1. For each non-medoid data point *o*
        1. Swap *m* and *o* and compute the total cost of the configuration
4. Select the configuration with the lowest cost.
5. Repeat steps 2 to 4 until there is no change in the medoid.

2.4.1.2 K- Mean (Fast)

In contrast to the standard implementation of k-means, this implementation is much faster in many cases, especially for data sets with many attributes and a high k value, but it also needs more additional memory.



First, pick initial centers. Set the lower bound $l(x,c) = 0$ for each point x and center c. Assign each x to its closest initial center $c(x) = \text{argmin}_c\, d(x,c)$, using Lemma 1 to avoid redundant distance calculations. Each time $d(x,c)$ is computed, set $l(x,c) = d(x,c)$. Assign upper bounds $u(x) = \min_c d(x,c)$.

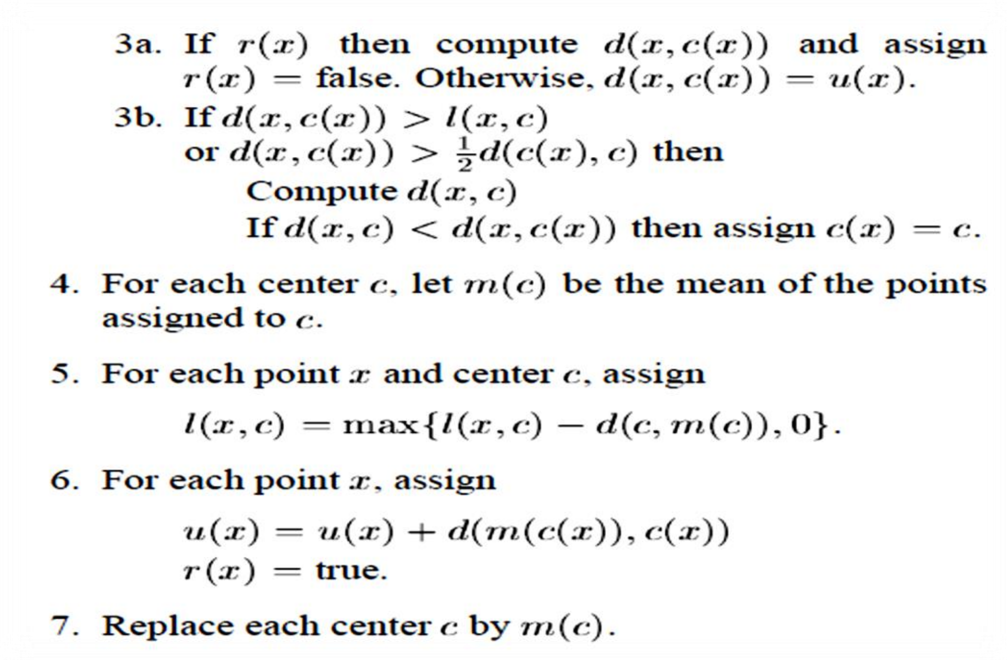
Next, repeat until convergence:

1. For all centers c and c', compute $d(c,c')$. For all centers c, compute $s(c) = \frac{1}{2}\min_{c' \neq c} d(c,c')$.

2. Identify all points x such that $u(x) \leq s(c(x))$.

3. For all remaining points x and centers c such that

    (i) $c \neq c(x)$ and
    (ii) $u(x) > l(x,c)$ and
    (iii) $u(x) > \frac{1}{2}d(c(x),c)$:



```
    3a. If $r(x)$ then compute $d(x,c(x))$ and assign
        $r(x)$ = false. Otherwise, $d(x,c(x)) = u(x)$.
    3b. If $d(x,c(x)) > l(x,c)$
        or $d(x,c(x)) > \frac{1}{2}d(c(x),c)$ then
            Compute $d(x,c)$
            If $d(x,c) < d(x,c(x))$ then assign $c(x) = c$.
4. For each center $c$, let $m(c)$ be the mean of the points
   assigned to $c$.
5. For each point $x$ and center $c$, assign
        $l(x,c) = \max\{l(x,c) - d(c,m(c)), 0\}$.
6. For each point $x$, assign
        $u(x) = u(x) + d(m(c(x)), c(x))$
        $r(x)$ = true.
7. Replace each center $c$ by $m(c)$.
```

2.5 Multi Cluster System (MCS)

Some methods for unsupervised detection of network attacks have been proposed in the past [37][38]; the majority of them are based on clustering techniques and outliers detection. The objective of clustering is to partition a set of unlabeled elements into homogeneous groups of “similar” characteristics, based on some similarity measure. Different from other techniques for unsupervised data analysis, clustering permits to work with multiple-classes problems without modifying the characteristics of the analyzed traffic, it represents an attractive means for unsupervised detection of attacks., even if hundreds of clustering algorithms exist [39], it is very difficult to decide which algorithm would be the best one for DDoS detection. Different clustering algorithms produce different partitions of data, and even the same clustering algorithm provides different results when using different initializations and/or different algorithm parameters. This is in fact one of the major drawbacks in current cluster analysis techniques: the lack of robustness.

To achieve higher robustness, in DDoS detection systems the paramount advantage of unsupervised, knowledge-independent detection algorithms based on clustering used, a multi-clustering methods is combined to perform robust unsupervised detection of DDoS attacks. The combination of multiple clustering adds robustness to the process of separating clustering .[47]

Summary

In this chapter, we presented the details of DDoS attacks, and approaches used in DDoS detection system. Data mining techniques and its use in DDoS detection have been explained as well. Furthermore, a brief description has been proposed about clustering algorithms (k-mean, k-medoid-Fast Mean) to be used in applying multi-clustering DDoS detection method. Finally we explained the importance of multi clustering system for detecting DDoS attack with high performance .

CHAPTER 3:
Related Work

CHAPTER 3: Related Work

Many recent researches in the last few years have been proposed and presented about "DDoS Detection" domain based on data mining as an efficient way to improve the security of networks, Two different approaches are by far dominant in current research community and commercial detection systems: signature-based detection and anomaly detection. The anomaly detection is supervised Anomaly Detection and Unsupervised Anomaly Detection.

3.1. Supervised Anomaly Detection:

In 2011 , Yang et al. [16] propose to detect DDoS attacks using decision trees and grey relational analysis. The detection of the attack from the normal situation is viewed as a classification problem. They use 15 attributes, which not only monitor the incoming/outgoing packet/byte rate, but also compile the TCP, SYN, and ACK flag rates, to describe the traffic flow pattern. The decision tree technique is applied to develop a classifier to detect abnormal traffic flow. They also use a novel traffic pattern matching procedure to identify traffic flow similar to the attack flow and to trace back the origin of an attack based on this similarity.

This technique has one advantage and one limitations, Their system could detect DDoS attacks with the false positive ratio about 1.2–2.4%, false negative ratio about 2–10% as an advantage , and find the attack paths in traceback with the false negative rate 8–12% and false positive rate 12–14% as a limitation .

In 2014 , Thw et al.[43] proposed system presents a classification scheme based on extracted features by using UCLA data set. The various packet features which exhibit DDoS attack natures in traffic are extracted from traffic data. Then, a data mining capability based on K-Nearest Neighbour approach combined with the proposed detection algorithm and classification algorithm is developed for attack detection. the system can correctly detect 94.87% for normal traffic and 98.87% for attack traffic. It incorrectly classified traffic in 5.13% for normal class and 1.13% for attack class.

In 2010, Nguyen et al. [17] develop a method for proactive detection of DDoS attacks by classifying the network status. They break a DDoS attack into phases and select features based on an investigation of DDoS attacks. Finally, they apply the k-nearest neighbor (KNN) method to classify the network status in each phase of DDoS attack.

In 2013, Selvakumar et al.[18] proposed a DDoS classification algorithm" NFBoost", it differs from the existing methods in weight update distribution strategy, error cost minimization, and ensemble output combination method, but resembles similar in classifier weight assignment and error computation. Their proposed NFBoost algorithm is achieved by combining ensemble of classifier outputs and Neyman Pearson cost minimization strategy, for final classification decision. Publicly available datasets such as KDD Cup, CONFICKER worm, UNINA traffic traces, and UCI Datasets were used for the simulation experiments. NFBoost was trained and tested with the publicly available datasets and their own SSE Lab SSENET 2011 datasets. Detection accuracy and Cost per sample were the two metrics used to analyze the performance of the NFBoost classification algorithm and were compared with bagging, boosting, and AdaBoost algorithms. From the simulation results, it is evident that NFBoost algorithm achieves high detection accuracy (99.2%) with fewer false alarms. Cost per instance is also very less for the NFBoost algorithm compared to the existing algorithms. NFBoost algorithm outperforms the existing ensemble algorithms with a maximum gain of 8.4% and a minimum gain of 1.1%.

This technique has the advantages the detection accuracy is high and the false alarm is fewer .but its limitation come from it use an old public dataset to test their method.

In 2011, Karimazad et al.[19] proposed propose an anomaly-based DDoS detection method based on the various features of attack packets, obtained from study the incoming network traffic and using of Radial Basis Function (RBF) neural networks to analyze these features. they evaluate the proposed method using their owne simulated network and UCLA Dataset. The results show that the proposed system can make real-time detection accuracy better than 96% for DDoS attacks.
This technique has an advantages the system can filter the attack traffics quickly and forward the normal traffics simultaneously. and one limitations as this is shown that the proposed method can successfully identify DDoS attacks but in low detection rates.

In 2008, Mihui et al.[20] proposed a combined data mining approach for the DDoS attack detection of the various types, that is composed of the automatic feature selection module by decision tree algorithm and the classifier generation module by neural network. For proving the practical detection performance of their approach, they gathered the real network traffic in the normal case and the attack case. they mounted the most powerful

DDoS attack changing attack types, so they could get the attack traffic of various types. This technique has an advantages they used the NetFlow data as the gathering data, because the analysis per flow is useful in the DDoS attack detection. Because the NetFlow provides the abstract information per flow, we don't need the extensive pre-processing, different with the tcpdump . And the limitations they couldn't gather the many attack runs because the DDoS attack could severely affect their network.

In 2012, Khamruddin et al.[21] proposed approach routers collectively try to mitigate the DDoS attack on the server. There are three steps in the proposed approach, initially, for attack detection and classification destination router (which is attached to the victim) monitors continuously the traffic pattern.
Second, once the attack is detected destination router tries to balance the load using the NAT (Network Address Translator). Third, whenever the attack is detected to mitigate different types of attacks, the signature is pushback to upstream routers so that the upstream routers start monitoring the traffic and apply the mitigation mechanism depending on type of attack detected.
This technique has an advantages they reduce the traffic on the victim machine so that the legitimate users get the services from destination machine.

3.2. Unsupervised Anomaly Detection:

In 2010, Zhong et al.[22] presents a DDoS attack detection method based on data mining algorithm. FCM cluster algorithm and Apriori association algorithm used to extracts network traffic method and network packet protocol status method. The threshold is set for detection method , From the analysis of DDoS attacks in the experiment, it is found that this system has a high detection efficiency, the detection rate reach more than 97%.
This technique has an advantages This method could receive the currently normal network traffic method with data mining algorithm. Once network traffic appears abnormal, this method could detect the packets maintaining in abnormal traffic duration. In this way the system load will be greatly reduced and its real-time can be improved. this system is able to effectively detect DDoS attacks in real time.

In 2008, Lee et al. [24] propose a method for proactive detection of DDoS attacks by exploiting an architecture consisting of a selection of handlers and agents that communicate, compromise and attack. The method performs cluster analysis. The authors experiment with the DARPA 2000 Intrusion Detection Scenario Specific Dataset to evaluate the method. The results show that each phase of the attack scenario is partitioned well and can detect precursors of a DDoS attack as well as the attack itself.

In 2014 , Meera et al.[30] alternative clustering approach is presented to perform robust unsupervised detection of attacks. The main idea is to combine the clustering results provided by multiple independent partitions of the same set of flow. The combination of multiple evidence on flow groupings adds robustness to the process of separating malicious from normal operation traffic. Automatic characterization and updation of attacks is used to find out the variation of flow.

In 2012 , Pedro et al.[40] presented a robust multi-clustering-based detection method and evaluated its ability to detect and characterize standard network attacks without any previous knowledge, using packet traces from two real operational networks. In addition, they have shown detection results that outperform previous proposals for unsupervised detection of attacks, providing more evidence of the feasibility of an accurate knowledge-independent detection system.

a new approach in unsupervised anomaly

In 2005 , Kingsly et al.[41] proposed approach in unsupervised anomaly detection in the application of network intrusion detection. The new approach, fpMAFIA, is a density-based and grid-based high dimensional clustering algorithm for large data sets. It has the advantage that it can produce clusters of any arbitrary shapes and cover over 95% of the data set with appropriate values of parameters. they provided a detailed complexity analysis and showed that it scales linearly with the number of records in the data set. They have evaluated the accuracy of the new approach and showed that it achieves a reasonable detection rate while maintaining a low positive rate.

In 2007 , YANG et al.[42] proposed Another unsupervised detection mechanism is where normal anomaly patterns are built over the network traffic dataset that uses subtractive clustering, and at the same time the built Hidden Markov Method correlates the observation sequences and state transitions to predict the most probable intrusion state sequences. The unsupervised anomaly detection approach proposed in should be capable of reducing false positives by classifying intrusion sequences into different emergency levels

In 2009 , Cuixiao et al.[44] a mixed intrusion detection system (IDS) method is designed. First, data is examined by the misuse detection module, then abnormal data detection is examined by anomaly detection module. In this method, the anomaly detection module is built using unsupervised clustering method, and the algorithm is an improved algorithm of K-means clustering algorithm and it is proved to have high detection rate in the anomaly detection module.

3.3. Semi-Supervised Anomaly Detection

In 2012, Hari et al.[23] presented A hybrid intrusion detection system that combines k-Means and two classifiers: K-nearest neighbor and Naïve Bayes for anomaly detection is presented , The presented method selects the important attributes and removes the irredundant attributes based on entropy based feature selection. This algorithm has been used on the KDD-99 Dataset; the system detects the intrusions and further classify them into four categories: Denial of Service (DoS), User to Root (U2R), Remote to Local (R2L) and probe and the experimental results reduce the false alarm rate.

In 2013 , Palnaty et al.[45] proposed and developed an algorithm called JCADS. The JCADS works based on the text similarities using Jaccord's Coefficient. Initially the dataset tuples are categorized based on the protocol and service used by the session. Because the attributes are categorical, the method is able to distinguish the protocol, service based clusters.The process improved the classification accuracy at the first stage. In the second stage, value similarities are measured on the Euclidian distance measure to form the clusters. The proposed two stage process, highly improved system to get the high accuracy. The experimental results show that, the use of two stage approach is the best way to cluster the intrusion attacks. The categorical clustering (semi-supervised),and the numerical distance in two stage clustering process is the essential for the intrusion clustering. The JCADS proved that multi-level attribute clustering improves the accuracy for intrusion detection systems. We conclude that the protocol,and services attribute values plays major role in the clustering process intrusion datasets.

Summary

In this chapter we presented an overview about some of researches conducted in DDoS detection based on Dataminig , we focus on the anomaly detection with its two category the supervised and the unsupervised detection , We explained the drawbacks of the existing methods used in previous researches In most circumstances, labelled data or purely normal data is not readily available since it is time consuming and expensive to manually classify it. Purely normal data is also very hard to obtain in practice, since it is very hard to guarantee that there are no intrusions when we are collecting network traffic[41], We try to counter this drawback in our research we proposed ”MCDDM” unsupervised multi-clustering method DDoS detection based on data mining as an efficient way to improve the security of networks .

CHAPTER 4:
The Proposed Method “MCDDM”

CHAPTER 4: The Proposed Method “MCDDM”

In this chapter, we present and explain the proposed method and methodology which we followed in this research. This chapter organized into four sections.

Section 4.1, presents methodology steps of proposed method, given description of the collecting data sets and description of their attributes, and integrate the data sets according to the proportion of the attack case1, 2, and 3. Section 4.2, contains the process of building the proposed method including the baseline experiments to select the optimal clustering algorithms in order to be used to build the proposed method. An explanation about the parameters for each algorithm has been mentioned as well.. Section 4.3, Section 4.4, present the measures to evaluating the performance of clustering with explained the equations used. Section 4.6, explained proposed method.

To achieve the objective of this research, we propose the following steps shown in Figure 4.1:

Step I: Collecting datasets normal dataset and DDoS attacks dataset .

Step II: Merge datasets according to attacks percentage , The purpose of this merger is to evaluate the performance of “MCDDM” methods and experimented the method performance to detected the huge amount of malicious packets.

Step III: For each case, we apply the “MCDDM” methods as follows :

a) Apply K-mean cluster in the first step to build KM method , and tested it. This step will produce output (cluster 1,cluster 0) (attacks/normal)

b) Apply K-fast Mean cluster in the second step on the same dataset to build KFM method , and tested it , this step will produce output (cluster 1,cluster 0) (attacks /normal).

c) Apply K-Mididod cluster in the second step on the same dataset to build KD method , and tested it , this step will produce output (cluster 1,cluster 0) (attacks /normal).

Step IV: We combined the three outputs from previous steps to generate the final output for all methods .

Step V: Extraction results to evaluate clusters performance by using the final Davies–Bouldin index .

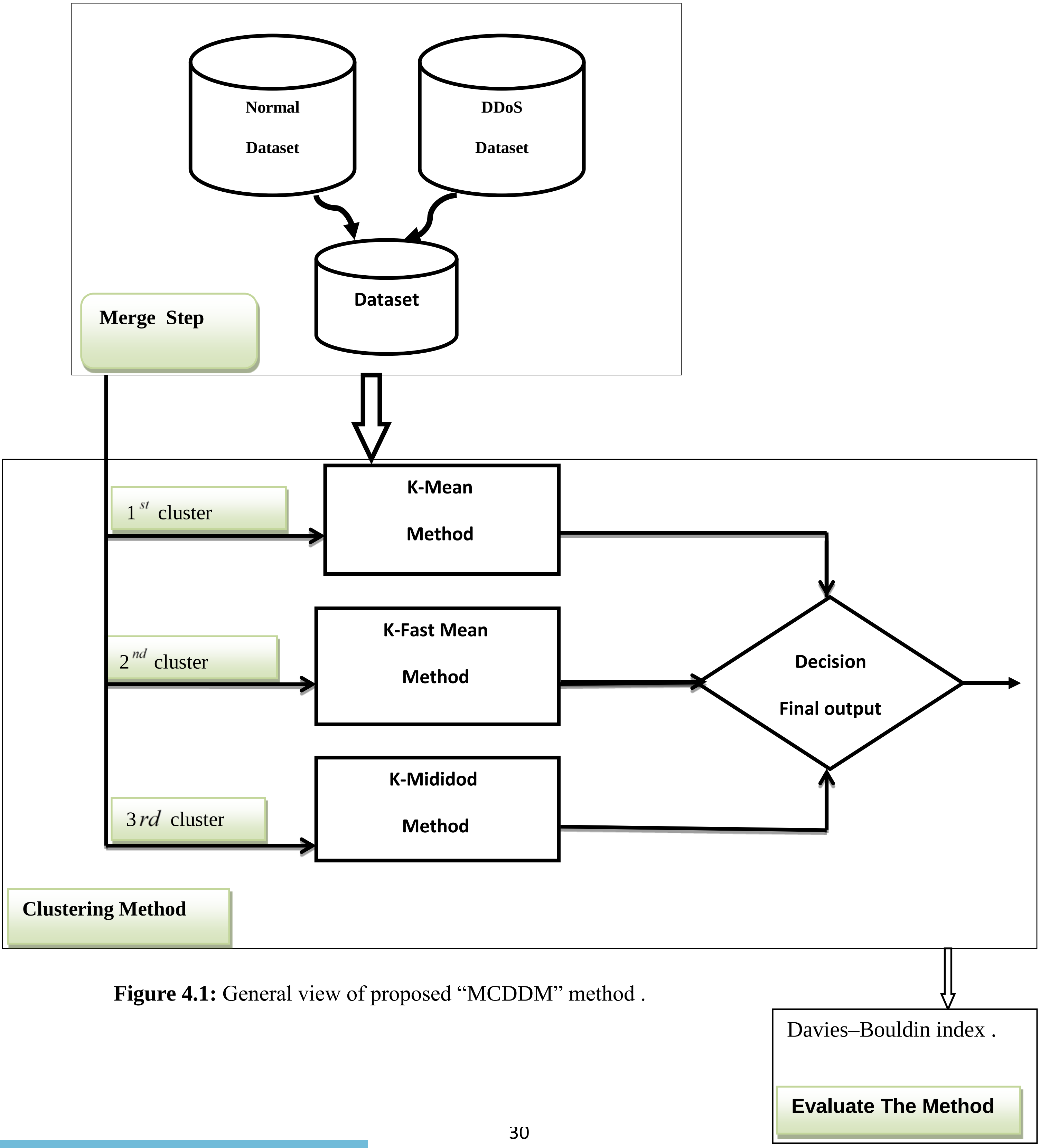


Figure 4.1: General view of proposed “MCDDM” method .

4.1 Methodology Steps

To apply and evaluate proposed method, we use the following methodology steps as

presented in Figure 4.1: Collection data sets: the collection of data sets from “The CAIDA DDoS Attack 2012 Dataset”. [15] ,Preprocessing data sets :For the purpose of applying the proposed method, Data sets converted to excel format , integrated according to attacks ,percentage should be done. Applying the method: By using three clustering algorithms: K-mean(KM), Fast k-mean(FKM), and K-Medoid(KD) as multi clustering, Evaluate the method: To evaluate the clustering performance of our method, we used davies_bouldin index. **Figuer.4.2** show the structure of the proposed method

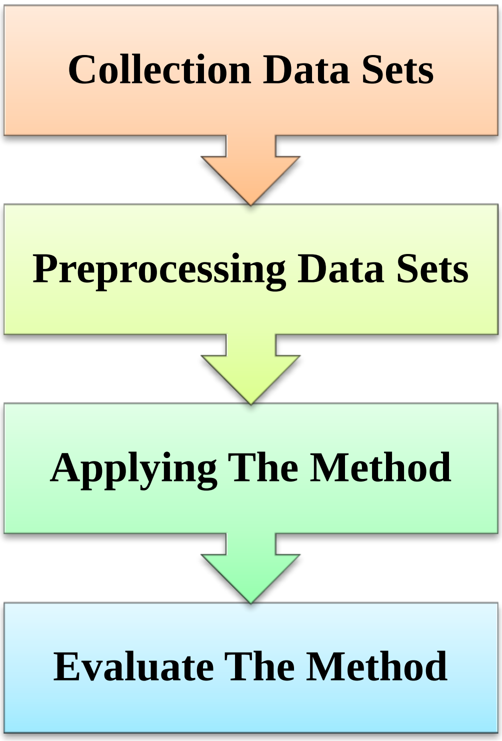


Figure 4.2 Methodology Steps

4.2 Data Collection

The real-world DDoS attacks are collected from [15]“The CAIDA DDoS Attack 2012 Dataset”. In this data set, the anonymized traffic were included a Distributed Denial of Service (DDoS) attack on August 04, 2012 for one hour time and size 21 GB [20].

Anonymized traffics was collected as DDoS attack traffic to-victim (including the attack traffic) and from-victim (including responses to the attack from the victim). DDoS traces block the victim (target server) by consuming the computing resources on the server and all of the bandwidths of the network connecting the server to the internet. On the other hand, the normal traffic traces are collected from “The CAIDA Anonymized Internet Traces 2014 Dataset”. This dataset contains anonymized passive traffic from “Equinix-Chicago’ OC192 link [31]

4.3 Data Preprocessing: We use the datasets from [15] [31],

- Open each data set using Wireshark version 1.10.6.
- convert the dataset to .xlsx to be suitable for rapidminar.
- Merge the datasets according to the parentage of the attacks on the normal dataset in three cases

4.4 Apply the “MCDDM” method

This section describes the types of clusters algorithms used in “MCDM” method: K-Mean(Km) ,K-Medoid(KD),K-Fast Mean(KFM) , which are provided by RapidMiner [46] program. We present these clustering algorithms and their settings which are used during experiments results by our model as the following:

4.4.1 K-Mean(Km)

K-Mean(Km) cluster used in “MCDM” method is one of the most widely used clustering algorithm , k-means clustering is an exclusive clustering algorithm i.e. each object is assigned to precisely one of a set of clusters. Objects in one cluster are similar to each other. The similarity between objects is based on a measure of the distance between them. Table 4.6 explain the setting of K-Mean(Km) cluster [46].

4.4.2 k-Medoids (KD)

In case of the k-medoids algorithm the centroid of a cluster will always be one of the points in the cluster. This is the major difference between the k-means and k-medoids algorithm. In k-means algorithm the centroid of a cluster will frequently be an imaginary point, not part of the cluster itself, which we can take as marking its center. Table 4.6 explain the setting of K-Mean(Km) cluster [46].

Input	Output	parameter
example set input *(Data Table)*	**cluster model** *(Centroid Cluster Model)*	**add cluster attribute** If enabled, a new attribute with *cluster* role is generated directly in this operator, otherwise this operator does not add the *cluster* attribute. In the latter case you ha operator to generate the *cluster* attribute. **Range:** *boolean* **add as label** If true, the cluster id is stored in an attribute with the *label* role instead of *cluster* role **Range:** *boolean* **remove unlabeled** If set to true, unlabeled examples are deleted. **Range:** *boolean* **k** This parameter specifies the number of clusters to form. There is no hard and fast rule of number of clusters to form. But, generally it is preferred to have small number of clusters with examples scattered (not too scattered) around them in a balanced way. **Range:** *integer* **max runs** This parameter specifies the maximal number of runs of k-Means are performed. **Range:** *integer* **max optimization steps** This parameter specifies the maximal number of iterations performed for one run of k-Means **Range:** *integer* **use local random seed** Indicates if a *local random seed* should be used for randomization. Randomization may be used for selecting *k* different points at the start of the algorithm as potential centroids. **Range:** *boolean* **local random seed** This parameter specifies the *local random seed*. This parameter is only available if the *use local random seed* parameter is set to true. **Range:** *integer*

Table 4.5 K-Mean, k-Medoids (KD) Setting

4.4.2 k-Means (fast)

In contrast to the standard implementation of k-means, this implementation is much faster in many cases, especially for data sets with many attributes and a high k value, but it also needs more additional memory.

Input	Output	parameter
example set: expects: ExampleSetMetaData: #examples: = 0; #attributes: 0 , expects: ExampleSet	cluster model cluster set	**add cluster attribute** If enabled, a cluster id is generated as new special attribute directly in this operator, otherwise this operator does not add an id attribute. In the latter case you have to use the Apply Model operator to generate the cluster attribute. Default value: true **add as label** If true, the cluster id is stored in an attribute with the special role 'label' instead of 'cluster'. Default value: false **remove unlabeled** Delete the unlabeled examples. Default value: false **k** The number of clusters which should be detected. Default value: 2 **determine good start values** Determine the first k centroids using the K-Means++ heuristic described in "k-means++: The Advantages of Careful Seeding" by David Arthur and Sergei Vassilvitskii 2007 Default value: false **Expert parameter** **measure types** **The measure type** Default value: NumericalMeasures mixed measure **Select measure** Default value: MixedEuclideanDistance Depends on: • measure types = MixedMeasures **nominal measure**

		Select measure **numerical measure** Select measure Default value: EuclideanDistance **max runs** The maximal number of runs of k-Means with random initialization that are performed. Default value: 10 **max optimization steps** The maximal number of iterations performed for one run of k-Means. Default value: 100 **use local random seed** Indicates if a local random seed should be used. Default value: false Expert parameter *local random seed* Specifies the local random seed Default value: 1992

Table 4.6 Fast K-Mean Setting(FKM)

4.4.4 Final "MCDDM" Output

Final "MCDDM" method output by use three clusters algorithms which are K-Mean(Km) ,K-Medoid(KD),K-Fast Mean(KFM), and combined the tree outputs to

generate the final output for all models, as the final output relies on equality the output of two model as follows:

(a) If any two method clustered the instant in cluster1 ("attacks"), and the third was cluster the instant as("normal"), so that the general output for "MCDDM" was "attacks".

(b) If any two method clustered the instant in cluster1 (“normal”), and the third was cluster the instant as(“attack”), so that the general output for “MCDDM” was “attacks”.

Finally, Davies_Bouldin Index will produce to evaluate “MCDDM” method.

4.5 Building the proposed method

To build the proposed method which is DDoS detection model based on multi clustering , we have conducted the following steps:

4.5.1 Cases of experiments

For 20000,5000,10000 dataset record three cases is done by increasing the percentage of DDoS attacks to experimented the capability of the "MCDDM" method to detected the huge amount of the malicious packets as the DDoS attacks is incremented , as follow

- Case 1: (10% attacks,90% normal).
- Case 2: (20% attacks,80% normal).
- Case 3: (30% attacks,70% normal).

4.5.1.1 First case (10% attacks,90% normal)

Dataset used in this case is composed of 20000,5000,10000 profiles, where contained

- 1000 attacks profile and 19000 normal profile .
- 500 attacks profile and 4500 normal profile.
- 1000 attacks profile and 9000 normal profile

4.5.1.2 Second case (20% attacks,80% normal)

Dataset used in this case is composed of 20000,5000,10000 profiles, where contained

- 200 attacks profile and 18000 normal profile .
- 1000 attacks profile and 4000 normal profile.
- 2000 attacks profile and 8000 normal profile

4.5.1.3 Third case (30% attacks,70% normal)

Dataset used in this case is composed of 20000,5000,10000 profiles, where contained

- 4000 attacks profile and 16000 normal profile .
- 1500 attacks profile and 3500 normal profile.
- 3000 attacks profile and 7000 normal profile .

Tabel 4.1 Cases of experiments

Case #	20000 Record		5000 Record		10000 Record		Output
	Normal 90%	Attack 10%	Normal 80%	Attack 20%	Normal 70%	Attack 30%	
1 (4 Exp)	19000	100	18000	2000	16000	4000	2 custer “attack, or Normal”
2 (4 Exp)	3500	1500	4000	1000	4500	500	
3 (4 Exp)	9000	1000	8000	2000	7000	3000	

4.3.1 The Base Line Experiments

4.3.1.1 Experimental Environment and Tools

Applied to experiments on a machine with properties that are Intel (R) Core(TM) 5i-4200 CPU @ 2.50 GHz processor and 4.00 GB of RAM. To carry out our thesis (including the experimentation), special tools and programs were used:

- **RapidMiner application program:** used to build our method, and Conduct experiments practical and extracting the required results.
- **Wireshark application program** :used to read .pcap Dataset and convert it to .xlsx format.
- **Microsoft Excel:** used excel to partition, organize and store datasets in tables, do some simple preprocessing and analyze the results.

To select the clusters to be used in building the proposed method , we marge datasets into 3 datasets, and apply the 3 partitioning clustering method K-Mean(Km) ,K-Medoid(KD),K-Fast Mean(KFM) .

Experiment Scenario I (10% attacks 90%normal)

In this experiment the datasets is merging as 10 % attacks and 90% normal and the clusters method , we perform 12 experimentation presented in section 4.2.1.1 ,Table 4.2 and Figure 4.3 illustrates experiments results in this case, which show that “MCDDM” method has achieved the best lowest Davies–Bouldin index,

Table 4.2: Experiments results of case 1

Method	Davies–Bouldin index		
	20000 Record	5000 Record	10000 Record
KFM	-0.347	-0.252	–0.542
KM	-0.347	-0.252	–0.542
KD	-0.322	-0.243	-0.494
MCDDM	-0.356	-0.274	-0.551

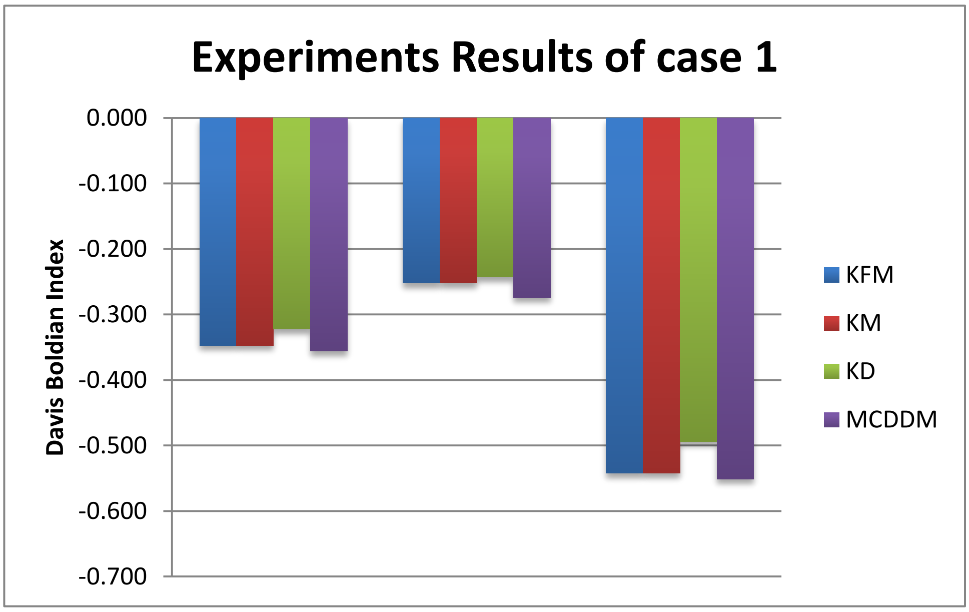


Figure 4.3: Experiments Results of case 1

Experiment Scenario II (20% attacks 80%normal)

In this experiment the datasets is merging as 20 % attacks and 80% normal and the clusters method , we perform 12 experimentation presented in section 4.5.1.2 ,Table 4.3 and Figure 4.4 illustrates experiments results in this case, which show that "MCDDM" method has achieved the best lowest Davies–Bouldin index.

Table 4.3: Experiments results of case 2

Method	Davies–Bouldin index		
	20000 Record	5000 Record	10000 Record
KFM	-0.441	-0.301	-0.565
KM	-0.441	-0.301	-0.565
KD	-0.387	-0.339	-0.544
MCDDM	-0.443	-0.342	-0.570

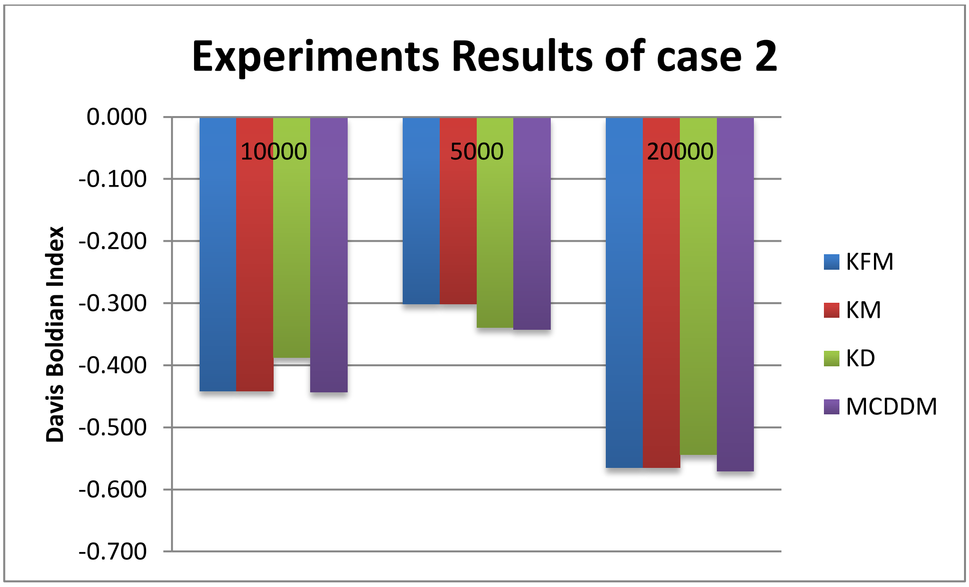


Figure 4.4: Experiments Results of case 2

Experiment Scenario III (30% attacks 70%normal)

In this experiment the datasets is merging as 30 % attacks and 70% normal and the clusters method , we perform 12 experimentation presented in section 4.2.1.3 ,Table 4.4 and Figure 4.5 illustrates experiments results in this case, which show that "MCDDM" method has achieved the best lowest Davies–Bouldin index.

Table 4.4: Experiments results of case 3

Method	Davies–Bouldin index		
	20000 Record	5000 Record	10000 Record
KFM	-0.336	-0.315	-0.634
KM	-0.336	-0.315	-0.634
KD	-0.312	-0.342	-0.644
MCDDM	-0.339	-0.355	-0.666

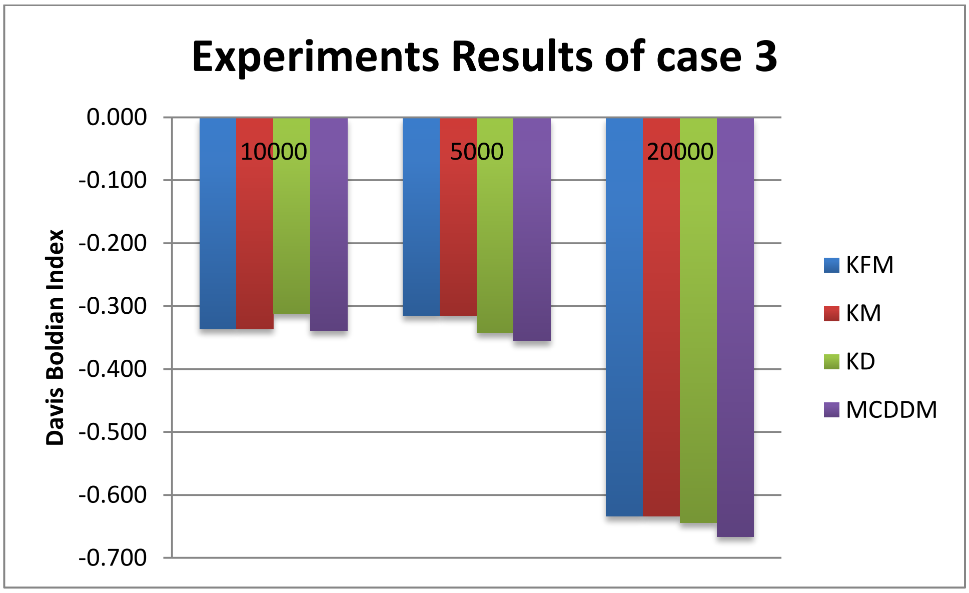


Figure 4.5: Experiments Results of case 3

4.6 Evaluate the “MCDDM” method

Performance evaluation of the “MCDDM” model is one of the most important tasks in our research. When a clustering result is evaluated based on the data that was clustered itself, this is called internal evaluation. These methods usually assign the best score to the algorithm that produces clusters with high similarity within a cluster and low similarity between clusters. One drawback of using internal criteria in cluster evaluation is that high scores on an internal measure do not necessarily result in effective information retrieval applications.[30] we use davies_bouldin index that the commonly evaluation measures for clustering method that can be defind as follow

Davies_Bouldin: The algorithms that produce clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies–Bouldin index, the clustering algorithm that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best algorithm based on this criterion.

The Davies–Bouldin index can be calculated by the following formula:

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

where n is the number of clusters, c_x is the centroid of cluster x, σ_x is the average distance of all elements in cluster x to centroid c_x, and $d(c_i, c_j)$ is the distance between centroids c_i and c_j. Since algorithms that produce clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies–Bouldin index, the clustering algorithm that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best algorithm based on this criterion.[30] Because the objective of the Davies-Bouldin index and its derivatives is to be minimized, a high negative value indicates a good performance of the index. Those values which are highlighted indicate when the Davies-Bouldin index had the best performance.[49]

4.7 Detecting DDoS Attack Using A Multilayer Data Mining techniques “MCDDM” (The Our Proposed Method)

The main objective of this research is to propose a new method of DDoS detection. To achieve this, we used combination of clusters as integration to be able to adapt with new DDoS attacks , and to achieve better Davies–Bouldin index.

Also, we try to overcome the drawbacks of the existing methods used in previous and related researches. For that, we propose “MCDDM” methods for DDoS detection based on multi clustering the experimental result show that the result of case3 is the best result that the "MCDDM" method give better method when the DDoS attack percentage is higher

Summary

In this chapter we explorer the methodology of “MCDDM” method, and the steps of building the “MCDDM” method collecting dataset and how we build “MCDM” method ,which cluster is used and its parameter , we explorer the base line experiment and results and how to evaluate the “MCDDM” method. **We can summarize our experiments results as follows:**

a) The experiments on datasets of case 1 achieved the lowest Davies–Bouldin index(-0.374), were in our method.

b) The experiments on datasets of case 2 achieved the lowest Davies–Bouldin index-0.570), were in our method.

c) The experiments on datasets of case 3 achieved the lowest Davies–Bouldin index-0.666), were in our model.

d) In general, we can say that our model has achieved good results from the all experiments on datasets of case 1, 2, and 3 where lowest Davies–Bouldin index was (-0.666) in case 3 which is 30% attacks and 70% normal .

CHAPTER 5:

Conclusion and Future work

CHAPTER 5: Conclusion and Future work

Today, the number of attacks against large computer systems or networks is growing at a rapid pace, one of the major threats to cyber security is Distributed Denial-of-Service (DDoS) attack As Intrusion detection becomes an integral part of any defense system within commercial organizations. Intrusion detection can be used well within the packages of the computer network devices ,two main types to intrusion detection are broadly used, which are Anomaly Intrusion Detection (AID) and Misuse Intrusion Detection (MID). Data mining techniques come to play a major role to detect and prevent the malicious. In the literature, data mining clustering methods have been considered for intrusion detection, especially for anomaly detection as an efficient ways to increase the security of networks. This chapter concludes the work, its results and discussion. Finally the future work directions were remarked.

5.1 Conclusion

In our research, we use three efficient clustering techniques in data mining, which are K-Mean **(KM)**, K-Mididod **(KD)**, and Fast K-Mean **(FKM)**.

These techniques were used in applying the proposed method. We proposed method which is an adaptive method based on multi clustering that able to be detecting DDoS attacks. The purpose of used multi clustering was to obtain reduce

Phase 1: collection of data sets from [15][31] Open each data set using Wireshark version 1.10.6,,convert the dataset to .xlsx to be suitable for RapidMiner, merge the datasets according to the parentage of the attacks on the normal dataset in three cases as Case 1: (10% attacks,90% normal),Case 2: (20% attacks,80% normal), Case 3: (30% attacks,70% normal).

Phase 2: we used RapidMiner program to apply our method, we have conducted a series of experiments to determine the three clustering used in our method which are
K-Mean (KM), K-Mididod (KD), and Fast K-Mean (FKM).

Phase 3: we used Davies–Bouldin index to evaluate “MCDDM” method,

We can concluded that “MCDDM” method achieved the best results for performance measurements which are Davies–Bouldin index.

5.2 Future Work Directions:

The future work direction of this dissertation extracted from the scope and limitation of the dissertation itself and from the experimental results. These directions can be summarized on the following points:

- Evaluate the proposed method with other attacks such as (DoS, Worm).
- Try to build method by a hyper of clustering methods and classification method to build the method to detect DDoS attacks.
- Evaluate the “MCDDM” method by real-time datasets.
- Try to Build the method with used of other cluster methods .

References

[1]. Sandoval, G. and Wolverton, T.." Leading Web sites under attack". 2000, CNET News. http: //news.cnet.com/Leading-Web-sites-under-attack/2100-1017_3-236683.html.

[2]. McCue, A. "Revenge' hack downed US port systems",. 2003. ZD- Net News. http://www.zdnet.co.uk/news/security-management/2003/10/07/revenge-hack-downed-us-port-systems-39116978/.

[3]. Lemos, R.." Web worm targets white house", 2001. CNET News. http://news.cnet.com/2100-1001-270272.html.

[4]. Gligor, V." A note on denial-of-service in operating systems",1984. Software Engineering, IEEE Transactions on 10, 3 (may), 320-324.

[5]. Hussian,H" An intelligent approach for dos attacks detection", , 2012, pp.1-4.

[6]. Kang, H. et al, " Traffic Deflection Method for DOS Attack Defense using a Location-Based Routing Protocol in the Sensor Network", ComSIS Vol. 10, No. 2, Special Issue, 2013, pp.1-17.

[7]. K Gar,M,." Detection of DDoS attack using data mining. International Journal of Computing and Business Research" (IJCBR) volume 2 Issue 1,2011.

[8]. Georgios,L " Protection against Denial of Service Attacks: A Survey" The Computer Journal Vol. 00 No. 0, 2009, pp. 1-19.

[9]. Dey Ch.; Master's Thesis in " Reducing IDS false positives using Incremental Stream Clustering (ISC) Algorithm ", Dept of Computer and Systems Sciences, Royal Institute of Technology, Sweden, 2009.

[10]. Olson D., and Delen D., “Advanced data mining techniques”, Springer-Verlag Berlin Heidelberg, 2008.

[11]. Law K.; and Kwok L.; "IDS False Alarm Filtering Using KNN Classifier", Lecture Notes in Computer Science, Springer Berlin, Heidelberg, 2005.

[12]. Source: http://en.wikipedia.org/wiki/Intrusion_detection_system, (2013, December), [Online].

[13]. Han J., and Kamber M., “Data Mining: concepts and techniques”, (2nd Edition), the Morgan Kaufmann Series in Data Management Systems, 2006.

[14]. Zhong R., and Guangxue Y., “DDoS Detection System Based on Data Mining”, (2nd Edition), Proceedings of the Second International Symposium on Networking and Network Security, 2010.

[15]. http://www.caida.org/data/passive/ddos-20130804_dataset.xml. [Online].

[16]. Tseng, H., Yang, W., and Jan, R. "DDoS detection and traceback with decision tree and grey relational analysis", International Journal of Ad Hoc and Ubiquitous Computing, 7, 121–136,2011.

[17]. Nguyen, H. and Choi, Y." Proactive detection of DDoS attacks utilizing k-NN classifier in an Anti- DDoS framework", International Journal of Electrical,Computer, and Systems Engineering, 4, 247–252, 2010.

[18]. Selvakumar S. and Arun Raj Kumar P.," Detection of distributed denial of service attacks using an ensemble of adaptive and hybrid neuro-fuzzy systems",Computer Communications Jornal, Volume 36 Issue 3, February, Pages 303-319,2013.

[19]. Reyhaneh K. and Faraahi A." An Anomaly-Based Method for DDoS Attacks Detection using RBF Neural Networks",International Conference on Network and Electronics Engineering,2011.

[20]. Mihui K. and Hyunjung H, "A Combined Data Mining Approach for DDoS Attack Detection",2008.

[21]..Khamruddin Md, Rupa Ch. "A Rule Based DDoS Detection and Mitigation Technique", In Proc :nirma university international conference on engineering, nuicone,2012.

[22]. Zhong , Yue." DDoS Detection System Based on Data Mining", Proc.the Second International Symposium on Networking and Network Security,2010.

[23]. Hari, O. and K. Aritra." A hybrid system for reducing the false alarm rate of anomaly intrusion detection system". Proc, the 1st International Conference on Recent Advances in Information Technology, 2012.

[24]. Lee, K., Kim J., Kwon K. , Han, Y., and Kim,S, " DDoS attack detection method using cluster analysis. Expert Systems with Applications", 2008.

[25]. Wesam, B.," Review clustering mechanisms of distributed denial of service attacks". 2014.

[26]. David, Z.," Peer to peer botnet detection based on flow intervals and fast flux network capture". MSc Thesis, University of Victoria, Heritage, Canada.2012.

[27]. Eskin E, "A geometric framework for unsupervised anomaly detection: detecting intrusions in unlabeled data," Applications of Data Mining in Computer Security, Kluwer Publisher, 2002.

[28]. Lakhina A. Crovella M, and. Diot C,"Mining Anomalies Using Traffic Feature Distributions," Proc. ACM SIGCOMM, 2005.

[29]. Leung K and. Leckie C, "Unsupervised Anomaly Detection in Network Intrusion Detection Using Clustering," Proc. 28th ACSC, 2005.

[30]. Meera R ," Detection & Deletion of DDOS Attacks Using Multi-clustering Algorithm",2014.

[31]. http://www.caida.org/data/passive/passive_2013_dataset.xml. [Online].

[32]. Criscuolo P, "Distributed Denial of Service, Tribe Flood Network" Department of Energy Computer Incident Advisory Capability (CIAC), UCRL-ID-136939, Rev. 1., Lawrence Livermore National Laboratory, February 14, 2000

[33]. Mehdi, E. and A. Amphawan, "Review of synflooding attack detection mechanism".Int. J. Distributed Parallel Syst., 3: 99-117. DOI: 10.202/1202.1761.pdf,2012.

[34]. Douligeris Ch.; and Serpanos D.; "Network Security Current Status and Future Directions", IEEE Press, 2007.

[35]. Pietro R.; and Mancini L.; "Intrusion Detection Systems", Springer Science and Business Media, LLC., 2008.

[36]. Monowar H. et al, "Detecting Distributed Denial of Service Attacks: Methods, Tools and Future Directions",Department of Computer Science, University of Colorado at Colorado Springs, CO 80933-7150,2013.

[37]. Portnoy. L, Eskin E., and Stolfo S., "Intrusion Detection with Unlabeled Data Using Clustering", in Proc. ACM DMSA Workshop, 2001.

[38]. Lakhina, Crovella M., and Diot C., "Mining Anomalies Using Traffic Feature Distributions", in Proc. ACM SIGCOMM, 2005.

[39]. Jain K, "Data Clustering: 50 Years Beyond K-Means", in Pattern Recognition Letters, vol. 31 (8), pp. 651-666, 2010.

[40]. Pedro H, Johan M, Philippe O. "Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge.", Computer Communications, Elsevier, 2012.

[41]. Kingsly L.," Unsupervised Anomaly Detection in Network Intrusion Detection Using Clusters", NICTA Victoria Laboratory Department of Computer Science and Software Engineering The University of Melbourne,2005.

[42]. Yang, C., Deng, F., Yang, H.,” An Unsupervised Anomaly Detection Approach using Subtractive Clustering and Hidden Markov Method”. Communications and Networking in China, 2007. CHINACOM '07. Second International Conference on , vol., no., pp.313-316, 22-24 Aug. 2007.

[43]. Thw T, Thandra P," Analysis of DDoS Detection System based on Anomaly Detection System", International Conference on Advances in Engineering and Technology (ICAET'2014) March 29-30, 2014.

[44]. Cuixiao Z ; Guobing Z ; Shanshan S," A Mixed Unsupervised Clustering-based Intrusion Detection Method ",3rd International Conference on DOI: 10.1109/WGEC.2009.72, 2009 .

[45]. Palnaty, R.; Rao, A. "JCADS: Semi-supervised clustering algorithm for network anomaly intrusion detection systems", Advanced Computing Technologies (ICACT), 15th International Conference on, On page(s): 1 – 5,2013.

[46]. Rapid Miner 6.003, http://www.rapidminer.com , (2014, October), [Online].

[47]. Fred A. and Jain A. K., “Combining Multiple Clustering Using Evidence Accumulation,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.27, no. 6, pp. 835–50, 2005.

[48]. Davies J. , Bouldin D.,.“A cluster separation measure”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1 :224-227, 1979.

[49]. Carlos J., Thomas R.,” New Version of Davies-Bouldin Index for Clustering Validation Based on Cylindrical Distance”, in Proc. Chilean Computer Conference 2013.