



SUDAN UNIVERSITY OF SCIENCE

AND TECHNOLOGY

College of Graduate Studies

**Anomaly detection approach using hybrid algorithm of
Data mining technique**

**طريقة كشف التسلل باستخدام خوارزمية هجين من تقنيات
تنقيب البيانات**

A Thesis Submitted In partial fulfillment of the Requirements for the
Degree of Master of Computer Engineering and Networking

Prepared by:

SAAD MOHAMED Ali MOHAMED GADAL

Supervised by:

Dr. Rania A. Mokhtar

February 2017

DEDICATION

*...to my small family wife,
kids and big family father,
mother because they are
close to my heart ♥!*

ACKNOWLEDGEMENT

Thanks to the Almighty God, for the Grace and Strength for completion. Thanks to Sudan University of science and technology for grate master program. Then I would like to express my special thanks of gratitude to my supervisor Dr. Rania A. Mokhtar as well as my big family and small family, gave me the right atmosphere for studyto seizegolden opportunity to do this wonderful project, also I would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited timeframe.

Saad Mohamed Ali

ABSTRACT

As known that most people in recent years become depend on the Internet in most things in their life, Now-a-days people rely on networks to send and receive emails, banking online system, stock price and online shopping. The excessive use of the communication networks leads to make important and secret information suspected to attacker, and the number of attacks on the important information over the internet is increasing daily. Intrusion is one of the main threats to the internet. Hence security issues had been big problem, so that various techniques and approaches have been presented to address the limitations of intrusion detection system such as low accuracy, high false alarm rate, and time consuming. This research proposed a hybrid machine learning technique for network intrusion detection based on combination of K-means clustering and Sequential Minimal Optimization (SMO) classification. The aim of this research is to introduce novel approach that able to reduce the rate of false positive alarm, to improve the detection rate and detect zero-day attackers and to get high accuracy for classify intrusion. The NSL-KDD dataset has been used to evaluate the proposed technique. In order to improve classification performance, some steps have been taken on the dataset like feature selection. The classification has been performed by using (Sequential Minimal Optimization SMO + K-mean clustering). After training and testing the proposed hybrid machine learning technique, the results have shown that the proposed technique (K-mean + SMO) has achieved a positive detection rate, reduce the false alarm rate and get high accuracy.

المستخلص

كما نعلم أن معظم الناس في السنوات الأخيرة أصبحت تعتمد على الإنترنت في معظم الأشياء في حياتهم، الآن الناس يعتمدون على الشبكات لإرسال واستقبال رسائل البريد الإلكتروني ونظام الخدمات المصرفية عبر الإنترنت وسعر الأسهم والتسوق عبر الإنترنت. الاستخدام المفرط لشبكات الاتصالات يؤدي إلى جعل المعلومات الهامة والسرية عرضة للهجوم، وعدد من الهجمات على المعلومات الهامة عبر الإنترنت يتزايد بصورة يومية والتسلل هو واحدة من التهديدات الرئيسية لشبكة الإنترنت. ولذلك أصبحت القضايا الأمنية مشكلة كبيرة. العديد من مختلف التقنيات والطرق قُدمت لمعالجة أوجه القصور في نظام كشف التسلل مثل انخفاض دقة النظام، وارتفاع معدل الانذارات الكاذبة، والزمن المستغرق للكشف عن التسلل. اقترح هذا البحث أسلوب تعلم الآلة الهجينة للكشف عن التسلل على الشبكة على أساس مزيج من خوارزمية K-mean وخوارزمية (SMO) والهدف من هذا البحث هو تقديم نهج جديد يمكن من تقليص معدل الانذارات الكاذبة، وتحسين معدل الكشف والكشف عن الهجمات الجديدة والتي تظهر لأول مرة وكذلك تحسين دقة التصنيف. وقد استخدمت مجموعة البيانات تسمى NLS-KDD dataset في التقنية المقترحة، ولأجل تحسين الأداء والتصنيف، واتخذت بعض الخطوات على مجموعة البيانات المستخدمة (NLS-KDD dataset) مثل اختيار الصفات ذات التأثير العالي في عملية التصنيف، وقد تم تنفيذ تصنيف باستخدام الخوارزمية الهجين (مزيج K-mean+SMO)، وبعد التدريب واختبار التقنية الهجين المقترحة، وقد أظهرت النتائج أن التقنية المقترحة (K-mean+ SMO) حققت معدل اكتشاف عالي وخفض معدل الإنذارات الكاذبة والحصول على دقة تصنيف عالية.

TABLE OF CONTENT

DEDICATION	ii
ACKNOWLEDGEMENT	iii
ABSTRACT.....	iv
المستخلص	v
TABLE OF CONTENT.....	vi
LIST OF FIGURES.....	ix
LIST OF TABLES.....	x
LIST OF ABBREVIATIONS	xi
chapter one	1
Introduction.....	1
1.1 Preface	1
1.2 Problem Statements:	2
1.3 Aim and Objectives	4
1.4. Methodology	4
1.5. Scope and limitations of the Thesis:.....	5
1.6. Thesis Organization	5
Chapter Two.....	6
Literature review	6
2.1. Introduction to network security:	6
2.2 Attacks in network:.....	8
2.2.1 Passive attacks	8
2.2.2 Active attacks.....	9
2.3 Threats to Network Security.....	11
2.3.1 Accidental Threats	11

2.3.2 Intentional Threats	11
2.4 Network security tools	14
2.4.1 Firewalls.....	15
2.5. Intrusion Detection System.	17
2.5.1 Types of IDS:.....	18
2.6. Organization of A Generalized Intrusion Detection System..	21
2.7. Anomaly Detection as a Process:	22
2.8. Adv/ Dis of Anomaly Detection and Misuse Detection:.....	23
2.9. Data Mining Technology	24
2.10. Data Mining Techniques and Intrusion Detection:	25
2.11. Algorithms build to optimize anomaly detection approach.	27
2.11.1. K-Means Algorithm.....	27
2.11.2. Sequential Minimal Optimization (SMO)	30
2.11.3. Genetic search algorithm.	32
2.12. Related Work:	35
Chapter three	41
Methodology	41
3.1 Overview:.....	41
3.2. Training phase:	43
3.3. Pre-processing:	43
3.3.1. Features Selection:.....	43
3.3.2. Clustering phase:	46
3.4. Testing phase:	47
3.5. Classification phase:	47
Chapter Four.....	48
Experiment implementation and Performance evaluation	48
4.1 Overview:.....	48

4.2. Dataset Description:.....	49
4.2.1. Brief description of the NSL-KDD dataset[48].....	50
4.3. Accuracy Measure of individual algorithms	54
4.3.1. Accuracy Measure of individual algorithms (SMO):.....	55
4.3.2. Secondly: apply K-mean algorithm:.....	57
4.3.3. Thirdly: apply hybrid approach (K-mean + SMO):	58
4.4. Comparison (SOM , K-meam, hybrid (K-mean + SMO))	60
Chapter five	64
Conclusions and Recommendations.....	64
5.1 CONCLUSION.....	64
5.2 .Future work:.....	65
References.....	66

LIST OF FIGURES

Figure (2.1) Intentional Threats [11].....	12
Figure (2.2) IDS categories.....	21
Figure (2.3) General architecture of IDS [23].....	22
Figure (2.4) Flow chart demonstrate k-mean process.....	30
Figure (2.5) Flow chart demonstrate Genetic search algorithm.....	34
Figure (3.1) model diagram.....	42
Figure (3.2) Feature selection.....	44
Figure (3.3) Feature selection Consistency SebsetEvel and GSA...	45
Figure (3.4) Name of Feature selected.....	46
Figure (4.1) SMO result using dataset with 22 attributes.....	55
Figure (4.2) K-mean clustering result using dataset with 22 attributes...	57
Figure (4.3) (K-mean + SMO) result using dataset with 22 attributes....	59
Figure (4.4) Compression of rate (SMO, K-mean, K-mean + SMO).....	61
Figure (4.5) Compare false positive (SMO, K-mean-mean + SMO).....	61
Figure (4.6) Accuracy for (SMO, K-mean, K-mean + SMO).....	62
Figure (4.7) Measurement parameters for (K-mean+SMO).....	62

LIST OF TABLES

Table (4.1) NSL KDD dataset features	49
Table (4.2) Anomaly types that include in NSL KDD dataset	50
Table (4.3) Description of the attributes of the NSL-KDD dataset....	50
Table (4.4) Demonstrate details of accuracy parameters for SMO.....	56
Table (4.5) Demonstrate Confusion Matrix for SMO.....	56
Table (4.6) Demonstrate measurement parameters for SMO.....	57
Table (4.7) Demonstrate Confusion Matrix for K-mean.....	58
Table (4.8) Demonstrate measurement parameters for.....	58
Table (4.9) Demonstrate accuracy parameters for (K-mean+SMO).....	59
Table (4.10) Demonstrate Confusion Matrix for(K-mean+SMO).....	60
Table (4.11) Demonstrate measurement parameters (K-mean+SMO)	60
Table (4.12) Demonstrate Comparison of for (K-mean+SMO).....	60

LIST OF ABBREVIATIONS

ADAM	Audit Data Analysis and Mining
AC	Accuracy
ACK	Acknowledgment
ARFF	Attribute Relation File Format
DoS	Denial of service
DTR	Decoction Rate
FN	False Negative
FP	False Positive
FPR	False Positive Rate
HIDS	Host Based Intrusion Detection
ICMP	Internet Control Message Protocol
IDS	Intrusion Detection System
IP	INTERNET PROTOCOL
KDD	Knowledge Discovery Databases
NATE	Network Analysis of Anomalous Traffic Event
NIDS	Network Based Intrusion Detection System
P-BEST	Production Based Expert System Tool Set
QP	Quadric Programming
R2L	Remote to Local
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine.
TCP	Transmission Control Protocol
TN	True Negative
TP	True Positive
U2R	User to Root
UDP	User Datagram Protocol
WEKA	Waikato Environment for knowledge Analysis

CHAPTER ONE

INTRODUCTION

1.1 Preface

Sensitive data and information stored on systems must be protected. The need for protecting data in computer systems became very necessary with the advent of shared systems. The need became more severe for systems being accessed over a public network, the Internet.

Information transmitting over network system must be protected from unauthorized release and changes. The connection itself must be securely established and maintained. The acceleration and development of computer network technology and Internet has brought huge convenience to people. Internet does not protect the information completely as the Internet is open system for general public.

Due to recent development and advances in network technology, computer systems have become more vulnerable to attacks. Novel attacks are appearing endlessly. In recent years, there are many statistics have shown that number of reported intrusions in the Symantec Global Internet Security Threat Report is growing [1].

Moreover, our dependency on network based systems is growing day by day. On the other hand, protection techniques of such systems do not keep up with the increasing threat. Existing defense mechanisms such as user authentication, data encryption, and firewalls are used as the first line of defense against attacks. Till date no combination of technology can protect the system completely because systems face novel attacks every other day. Therefore, researchers are now attempting to find better and safer approach to minimize and overcome some limitation of existing prevention techniques.

In order to face the problem of attacks, tools such as Firewall, Anti-Virus, Intrusion Detection, Prevention, and Response Systems have invented. Their aim is to monitor the system or network activity and detect, prevent, or counter suspicious incident. The technologies deployed by security managers to protect the enterprise [2].

In this research, we introduce novel approach which combines two algorithms using data mining techniques clustering and classification (k-mean clustering and sequential minimal optimization) to enhance the performance of intrusion detection system parameters (reduce the rate of false positive alarm, to improve the detection rate and accuracy).

1.2 Problem Statements:

With the advance in science and technology, people are getting more dependent on computer networks for news, stock prices, email and online shopping. Due to this increasing dependency, technology is

defending against a number of threats to maintain the integrity, confidentiality and availability of computer system. Now-a-days it has become important to maintain a safe computer system and devices. So, intrusion detection system (IDS) plays a major role while detecting intrusion. In misuse detection method only known attacks are detected and on other side anomaly detection method detects unknown attacks also. Anomaly detection can prevent the system from any kind of new or Zero-day attack. Anomalies are detected on the basis of user behavior. There are so many data mining techniques to classify the normal or abnormal behavior of the user. But these data mining techniques have some limitation so the main objective of the research is to reduce these limitations and improve the accuracy [3,4]. Existing techniques having the following problems:

- One big issue with anomaly detection systems is the efficiency and speed. If the amount of network traffic is high, it might be impossible to use complicated algorithms fast enough to detect intrusions before it's too late. Many advanced algorithms achieve a high detection rate but are too computationally complex for practical use.
- Missing values or missing data have a significant effect on the result while classification. These missing values are extremely important for feature selection while classification. Because of this missing data, conclusions that could be drawn from the data are not accurate.

- Real-time traffic analysis is one of the major concerns that are being faced now-a-days. If detection of real-time traffic is not accurate, then it leads to compromising the system. So, there is a need for the classifiers which are more accurate in detecting intrusion.

1.3 Aim and Objectives

The objectives of the research are as follows:

- To propose a new technique to improve the accuracy and reduce the false alarm rate and to improve the detection rate of anomalies.
- To analyze and compare different data mining techniques for the above stated problem.
- To validate the new proposed technique on the dataset.

1.4. Methodology

In this section, a new anomaly detector approach is proposed based on using k-means clustering algorithm and Sequential Minimal Optimization (SMO) to detect online network anomaly, the proposed approach aims to generate a suitable number of detectors with high detection accuracy. The main idea is based on using feature selection in preprocessing phase to reduce the number of dataset, The ConsistencySubsetEval and Genetic search algorithms have been applied on NLS-KDD dataset to select specific features from the dataset and remove

those features which are irrelevant before clustering and classification phases, after attribute selection, k-means clustering algorithm selected to reduced training dataset in order to decrease time and processing complexity. In classification phase, supervised algorithm Sequential Minimal Optimization (SMO) selected to improve the quality of detection.

1.5. Scope and limitations of the Thesis:

Hybrid of data mining algorithm (K-mean clustering and sequential minimal optimization SMO) proposed to improve parameters of performance of anomaly detection like detection rate, false positive rate alarm and accuracy of the system, WEKA (Waikato Environment for Knowledge Analysis) used to done this task.

1.6. Thesis Organization

This thesis composed of five chapters, the first chapter briefly describes the background study of network security and introduces the problem statement, methodology and aim and objective of whole thesis, chapter two gives a brief background of Intrusion Detection System, it also presents the related works. Chapter three represents the project methodology, Chapter four presents the experimental and results discussion. Chapter five concludes this thesis and shows the future directions as recommendations.

CHAPTER TWO

LITERATURE REVIEW

2.1. Introduction to network security:

Information stored on systems had to be protected since the very introduction of computers. The need for protecting files in computer systems became more evident with the advent of shared systems [5]. The need became even more severe for systems being accessed over a public network, the Internet. The rapid development of computer network technology and Internet has brought huge convenience to people. Internet does not conceal or protect the information completely as the Internet is open system for general public.

Due to dependency on network based systems is growing day by day. On the other hand, protection techniques of such systems have not kept up with the increasing threat. Traditional defense mechanisms such as user authentication, data Encryption, avoiding programming loopholes and firewalls are used as the first line of defense against attacks. Till date no combination of technology can protect the system completely because systems face novel attacks every other day. In order to counter the problem of attacks, tools such as Firewall, Anti-Virus, Intrusion Detection, Prevention, and Response Systems have emerged. Their aim is to monitor the system or network activity and detect, prevent, or counter suspicious incident [2].

The technologies deployed by security managers to protect the enterprise, are useful for defending attacks to some extent only. They have their own limitations. For example, firewalls may be configured to block certain types of traffic, but attackers still find ways to exploit legitimate traffic types to mount their attacks. The following major security objectives of an application have paramount importance to ensure the security of network [6].

- Confidentiality: It means that certain information is only accessible to those who have been authorized to access it.
- Integrity: It guarantees that a message being transferred is never corrupted.
- Availability: Availability guarantees the survivability of network services despite denial of service (DoS) attacks.
- Authenticity: Authenticity ensures that participants in communication are genuine and not impersonators.
- Non-repudiation: It ensures that the sender and the receiver of a message cannot deny that they have ever sent or received such a message.
- Authorization: It is a process in which a trusted certificate authority issues a credential to an entity.

- **Anonymity:** It means that all the information related to owner or current user entity identification should be kept secret and not distributed to other communicating parties [7].

2.2 Attacks in network:

Without security measures and controls in place, your data might be subjected to an attack. Some attacks are passive, meaning information is monitored; others are active, meaning the information is altered with intent to corrupt or destroy the data or the network itself.

Networks and data are vulnerable to any of the following types of attacks if we do not have a security plan in place.

2.2.1 Passive attacks

A passive attack [8] is the one in which the intruder eavesdrops or monitors the transmitted data but does not modify the message stream in any way. The goal of the attacker is to get information in transit. Two types of passive attacks are release of message contents and traffic analysis. The release of message contents is simply reading the contents of a message. It can be troublesome if the message is carrying sensitive or confidential data. A second type of passive attack, traffic analysis is subtler. An intruder makes inferences by observing message patterns. It can be done even if messages are encrypted. In this attack, the eavesdropper analyzes the traffic, determines the location and identifies communicating hosts. Eavesdropper can also observe the frequency and length of message being exchanged.

This information is used to predict the nature of communication. All incoming and outgoing traffic of network is analyzed but not altered. Passive attacks are very difficult to detect because they do alter the data. Neither the sender nor the receiver is aware that a third party has read the messages during their exchange. This can be prevented by means of encryption of data. Encryption is the technique for masking the contents of message.

If one had encryption protection done, an opponent can still observe the pattern of this message [9]. The opponent could determine the location and identity of communicating hosts and could observe the frequency and length of messages being exchanged. The information might be useful in guessing the nature of the communication that was taking place. Thus, the emphases in dealing with passive attacks are on prevention rather than detection.

2.2.2 Active attacks

Active attacks [8] involve some modification of data stream or creation of a false stream. An active attack is one in which the intruder may send messages, replay old messages, change messages in wire, or delete selected messages in trans.

Atypical active attack is one in which an intruder pretends to be one end of the conversation, or acts as a man-in-the-middle. They can be subdivided into four categories:

➤ **Masquerade:**

A masquerade takes place when one entity pretends to be a different one. This attack generally includes one of the other forms of active attacks. For instance, authentication sequences can be captured and replayed after a valid authentication sequence has taken place. This enables an authorized entity with few privileges to obtain extra privileges by impersonating an entity that has those privileges [10].

➤ **Replay:**

Replay involves the passive capture of a data unit and its subsequent retransmission to produce an unauthorized access [10].

➤ **Modification of messages:**

Modification of messages means that some part of a legitimate message is altered or that messages are delayed or reordered, to produce an unauthorized effect. For example, a message meaning “Allow John Smith to read confidential files account” is modified to mean “Allow Fred Brown to read confidential file accounts” [10].

➤ **Denial of service:**

The denial of service prevents or inhibits the normal use of communication devices. This attack may have a specific target, for instance, an entity may suppress all messages meant for a particular destination.

Another form of this attack is disruption of the entire network. This is done either by crippling the network or by overloading it with messages in order to degrade its performance [10].

2.3 Threats to Network Security.

Threats can be defined as potential violations of security. They exist because of vulnerabilities or weaknesses in a system. Basically, there are two types of threats: Accidental threats and intentional threats [11].

2.3.1 Accidental Threats

Accidental threats result in either the exposure of confidential information or occurrence of an illegal system state. Exposures can appear from both hardware and Software failures as well as from user and operational mistakes. This results in the violation of confidentiality. It can also be demonstrated as modification of an object that is the violation of object integrity.

2.3.2 Intentional Threats

Intentional threats can be defined as an action deliberately performed by an entity with the intention of violating security. Examples of such attacks are modification, interception, interruption and fabrication of data as shown in Figure (2.1) [11].

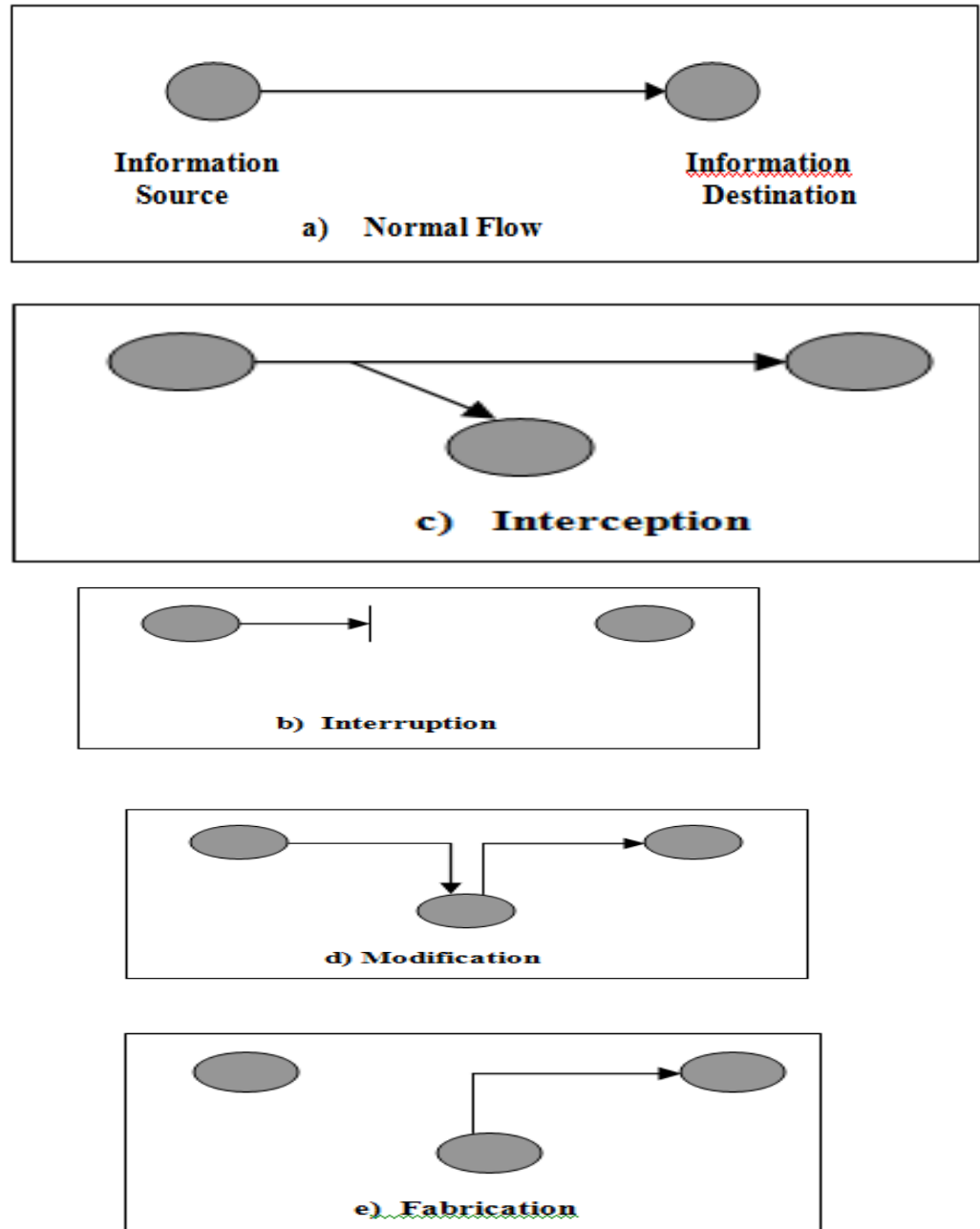


Figure (2.1): Intentional Threats [11]

There are numerous threats to network security. Some of them are as listed below:

➤ **SYN flooding:**

It is a denial-of-service attack in which a large amount of SYN packets are sent to a network or a server. The attack packets usually have spoofed source addresses to hide the real attacking sources and this makes defense much harder. The SYN flooding attacks exploit the TCP's three-way handshake mechanism and its limitation in maintaining half-open connections, when a server receives a SYN request; it returns a SYN/ACK packet to the client. Until the SYN/ACK packet is acknowledged by the client, the connection remains in half open state for a period of up to the TCP connection timeout, which is typically set to 75 seconds. The server has built in its system memory a backlog queue to maintain all half-open connections. Since this backlog queue is of finite size, once the backlog queue limit is reached, all connection requests will be dropped [12].

➤ **Packet sniffers:**

Packet sniffing [13] is a method of tapping each packet flowing across the network. It is a technique in which a user sniffs data belonging to other users of the network. Packet sniffers can be used as an administrative tool or for malicious purposes. Network administrators operate them for monitoring and validating network traffic.

➤ **Viruses:**

A small piece of code that recursively replicates a possibly evolved copy of itself on real programs. They run every time a program runs and multiply to form new generations. Most of them can reproduce and attack other programs. Trojan horse is a program containing hidden functions that can exploit the privileges of the user running the program. It can erase important information; send credit card numbers and password to the intruder [14].

➤ **Spyware:**

Spyware is a new type of potentially unwanted program whose goal is to monitor users' online behavior without user consent. Users infected with spyware generally experience highly degraded reliability and performance such as increased boot time, unresponsive system, and frequent application crashes [15].

2.4 Network security tools

Information security managers have used multiple technologies to keep their network safe from intrusions. However, as an effect of the improvements in technology, networks are now connected to other outside networks – including the Internet. So, the corporations face a wide range of threats. So, security managers are under a lot of pressure to prevent any penetration to the network perimeter. Similar to the physical security, there are numerous security tools to help security managers in setting up complex

protection strategy plans for their computer systems. Most common ones are listed below:

2.4.1 Firewalls

Firewall [16] is a hardware or software solution implemented within the network to enforce security policies by controlling network access. The original function of firewall was protecting a network from unauthorized external access. Now firewalls can also inspect and filter traffic arriving or departing a network by comparing packets to a set of rules and performing the matching rule action, which is accept or deny. A firewall is often seen as the first step toward a network security solution. Firewalls must be installed at the choke points to control network traffic and implement network security policy of the organization for its external network connections, especially for the Internet. Because many Internet-based services are inherently insecure, a firewall is needed to disable some services according to the organizational security policy. A firewall can act as a wall between the two networks. The person want access to the either network has to pass this wall before entering. A firewall is a system that is set up to control traffic flow between two networks. There are various firewall products but they are grouped into three major types based on their mechanisms:

➤ Packet filtering:

Packet filtering [17] is a mechanism that controls the flow of packets in a network by examining their headers. No content-based

decisions are made. The decision is exclusively based on the packet headers which include type of traffic (such as TCP, UDP, ICMP), characteristics of the transport layer communications sessions (such as source and destination ports), source and destination address. Packet filters are coupled with interfaces and the packet that flows through the interface can be restricted.

Packet filter firewall may have rules such as: allowing certain hosts send email via Simple Mail Transfer Protocol or not permitting any outside system to connect to an internal host via Telnet.

➤ **Stateful inspection:**

This technology has evolved from the need to accommodate some features of the TCP/IP protocol suite. Stateful inspection firewalls are packet filter firewalls with the ability of connection status awareness. This awareness is made by making a dynamic list of active connections between hosts which is called state table. Those packets are rejected by the firewall, which do not belong to an active connection or is not a connection request. A packet belonging to an active connection is allowed through bypassing the firewall rules and thus optimizing the investigation process [17].

➤ **Proxying:**

It is a mechanism that provides all internal hosts the untreated external network access while it appears that a single host is accessing outside, since all connections to the external network is made by a single host, deep packet inspection can be done before passing packets to internal

nodes. Proxying inspects source address, destination address, protocol, source port number, destination port number and payload of packets [18].

2.5. Intrusion Detection System.

Intrusion is any kind of unauthorized activity on a computer network. It is achieved passively or actively. In passive, intrusion takes place by information gathering and eavesdropping, whereas in case of active intrusion takes place through harmful packet forwarding, packet dropping and by hole attacks [6]. An IDS is a process or device that monitors events occurring on a network and analyzing it to detect any kind of activity that violate computer security policies. The IDS device can be hardware, software or a combination of both that monitors the computer network against any unauthorized access [7]. The main motive of the IDS is to catch the intruder before a real and serious damage to computer network.

The purpose of an IDS product is to monitor the network system for any type of attacks, an attack might be signaled by something as simple as a program that could modify the user name or could be a complex attack that involves sequence of events spanning multiple systems. IDSs are classified through system monitors because they usually depend on auditing information provided from the systems logs or data gathered by sniffing network traffic [19].

The process of intrusion-analysis can be separated into four stages as below [20]:

- Preprocessing: When data is collected from an IDS sensor, the data is organized for classification. The preprocessing will help us to determine the format the data is put into, which is usually some canonical format or could be a structured database.
- Analysis: The analysis stage begins after the preprocessing stage is completed and it is applied to all the records in the database. The data record is compared with the knowledge base, and the data record will either be logged as an intrusion event or it will be dropped.
- Responses: When the data record has been logged as an intrusion, a response can be initiated. This response contains an alert and passively collection information about the intrusion.
- Refinement: This stage is responsible for the correctness of the intrusion.

2.5.1 Types of IDS:

In software based NIDS approach the IDS are software systems that are specially designed with the aim of identifying and hence help to prevent the malicious activities and security policy violations. IDS can be classified into two main categories: analysis approach and placement of IDS. Analysis approach consists of misuse detection and anomaly detection.

2.5.1.1. Placement approach: consist three types mentions below.

➤ Host based IDS (HIDS).

Host based IDS examines the data from a single host. Host based intrusion detection tries to recognize unauthorized, illegitimate, and malicious behavior on a specific Device. HIDS usually involves an agent installed on each system which monitors and Alerts on local OS and application activity. This installed agent uses signatures and Rules to discover unauthorized activity. Host IDS only collects identifies and alerts the system, so its role is only passive [21].

➤ Network based IDS (NIDS).

NIDS are deployed on strategic point in network infrastructure. The NIDS can capture and analyze data to detect known attacks by comparing patterns or signatures of the database or detection of illegal activities by scanning traffic for anomalous activity.

NIDS are also referred as “packet-sniffers”, because it captures the packets passing through the of communication mediums.

➤ Hybrid based IDS.

The management and alerting from both network and host-based intrusion detection devices, and provide the logical complement to NID and HID -central intrusion detection management.

2.5.1.2. Analysis approach: consist two types:

➤ Misuse Detection:

This approach uses pattern matching algorithm to look for some known misuses. They have very low false positive (IDS generate alarm when no attack has taken place) rate. Since they depend on comparing the incoming traffic with a known set of malicious strings they are unable to identify novel attacks. Hence a high false negative (Failure to detect an actual attack) rate is observed. The number of patterns now has reached the order of the thousands making the computation a rather difficult task.

➤ Anomaly Detection:

This approach makes decisions based on normal network or system behavior using statistical techniques [22]. This approach monitors network traffic and compares it against an established baseline of normal traffic profile. The baseline characterizes normal behavior for the network - such as the normal bandwidth usage, the common protocols used. This approach is able to identify novel attacks that are yet unknown and hence undetectable by signature based NIDS. The main disadvantage of anomaly detection method is that it may generate a large number of false positives.

There are many anomaly detection techniques as follows: statistical methods, data-mining methods and machine learning based methods. In statistical methods, it is assumed that a variation of the traffic in terms of volume of number of packets indicates attack, like bandwidth flooding attack. But if the attacker keeps traffic parameter below a certain level this

method will not work. Incorrect combinations of port numbers and devices indicate attack. Then NIDS should alert the administrator or user regarding detection of anomalous traffic.

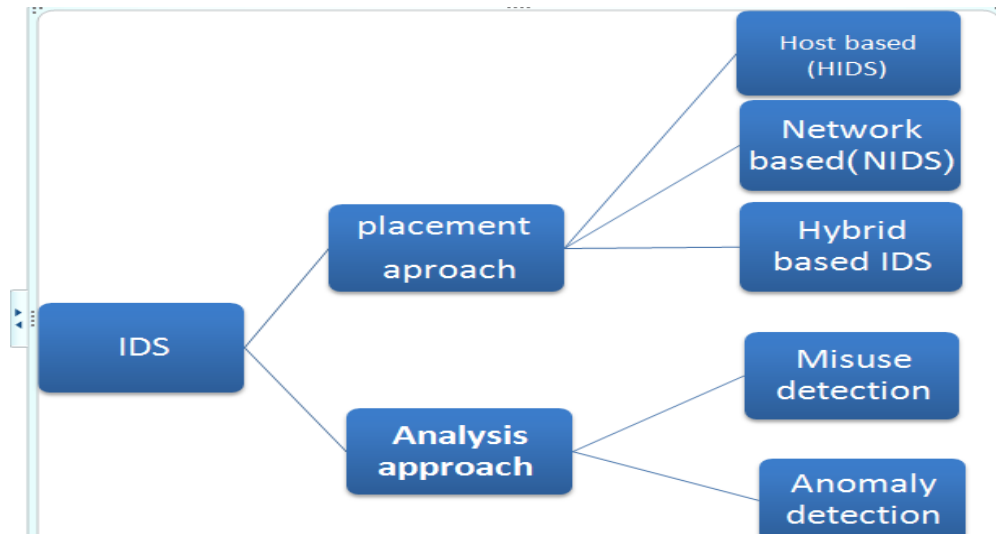


Figure (2.2). IDS categories

2.6. Organization of A Generalized Intrusion Detection System.

Figure 2.3 illustrates the overall architecture of IDS. It has been placed centrally to capture all the incoming packets that are transmitted over the network. Data are collected and send for pre-processing to remove the noise; irrelevant and missing attributes are replaced. Then the preprocessed data are analyzed and classified according to their severity measures. If the record is normal, then it does not require any more change or else it send for report generation to raise alarms. Based on the state of the data, alarms are raised to make the administrator to handle the situation in advance. The attack is modeled so as to enable the classification of network data. All the above process continues as soon as the transmission starts [23].

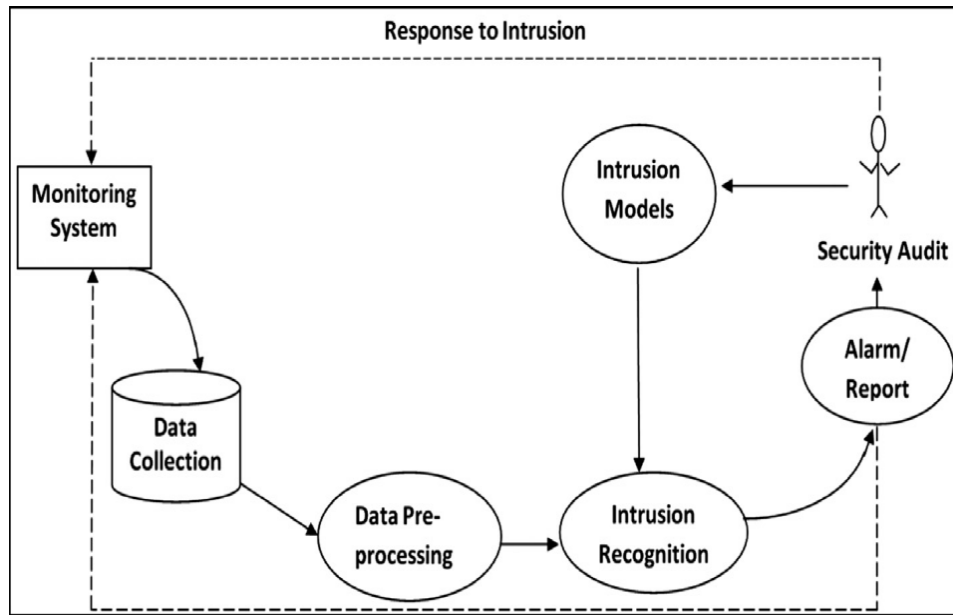


Figure (2.3) General architecture of IDS [23]

2.7. Anomaly Detection as a Process:

As a process, anomaly detection can be divided into two phases. In the first phase a model of normal network traffic is created. This model can be derived or learned from training data using model generation algorithms or mathematical models. In the second phase traffic is monitored for deviations from the normal model. [24]

The model of normal network traffic is created by using features from the traffic. Feature in the context of anomaly detection means a value or symbol which describes the network traffic. These features should represent the traffic behavior and characteristics but in the same time they should not contain any redundant information in order to be as lightweight as possible, the word, feature has numerous synonyms such as variable, parameter and descriptor.

In order to create a model of the normal network traffic, it needs to be clean from malicious activities and at the same time it needs all the variations of the environment it is monitoring.

Once the model of normal network traffic is created, traffic is then monitored for deviations from the model. Some analysis is needed to decide whether the deviation is intrusive or malicious. Normally this analysis is done by a network security guard. As the detected anomalies, might be previously unknown it is difficult to know what is actually, causing the anomaly and whether it is intrusive or not. [25] This anomaly analysis process needs to be supported by as much information as possible, so that the security guard could work efficiently.

2.8. Advantages and Disadvantages of Anomaly Detection and Misuse Detection:

The main disadvantage of misuse detection approaches is that they will detect only the attacks for which they are trained to detect. Novel attacks or unknown attacks or even variants of common attacks often cause high false positive alarm. The main advantage of anomaly detection approaches is the ability to detect novel attacks or unknown attacks against software systems, variants of known attacks, and deviations of normal usage of programs regardless of whether the source is a privileged internal user or an unauthorized external user.

The disadvantage of the anomaly detection approach is that well-known attacks may not be detected, particularly if they fit the established

profile of the user. Once detected, it is often difficult to characterize the nature of the attack for forensic purposes. Finally, a high false positive rate may result for a narrowly trained detection algorithm, or conversely, a high false negative rate may result for a broadly trained anomaly detection approach [26].

2.9. Data Mining Technology

The term data mining is used to describe the process of extracting useful information from the large databases. Data mining analyses the observed sets to discover the unknown relation and sum up the results of data analysis to make the owner of data to understand [27]. Hence data mining problems are considered as a data analysis problem. Data mining framework automatically detect patterns in our data set and use these patterns to find a set of malicious, Data mining techniques can detect patterns in large amount of data, such as byte code and use these patterns to detect future instances in similar data. In intrusion detection system, information comes from various sources like host data, network log data, alarm messages etc. Since the variety of different data sources is too complex, the complexity of the operating system also increases. Also, network traffic is huge, so the data analysis is very hard. The data mining technology have the capability of extracting large databases; it is of great importance to use data mining techniques in intrusion detection. By applying data mining technology, intrusion detection system can widely verify the data to obtain a model, thus helps to obtain a comparison between the abnormal pattern and the normal behavior pattern. Manual analysis is not required for this method. One of the main advantages is that same data

mining tool can be applied to different data sources. Main problem in intrusion detection is effective separation of the attack patterns and normal data patterns from a large number of network data and effective generation of the automatic intrusion rules after collected raw network data. For this purpose, several methods of data mining are used in such type of classification, clustering and association rule mining etc. (23).

2.10. Data Mining Techniques and Intrusion Detection:

Data Mining is used in variety of applications that requires data analysis. Now a day's data mining techniques plays an important role in intrusion detection systems. Different data mining techniques like Classification, Clustering and Association rules are frequently used to acquire information about intrusions by observing network data. Here we describe different data mining techniques that help in detecting intrusions.

Classification: Classification is a form of data analysis which takes each instance of a dataset and assigns it to a particular class. It extracts models defining important data classes. Such models are called classifiers [28].

A classification based IDS will classify all the network traffic into either normal or malicious. Data classification consists of two steps – learning and classification. A classifier is formed in the learning step and that model is used to predict the class labels for a given data in the classification step. In analysis of classification the end-user/analyst requires to know ahead of time how the classes are defined.

Each record in the dataset already has assessment for the attribute used to define the classes. The main aim of a classifier is not only to explore the data to discover different classes, but also to find how new records should be arranged into classes. Classification helps us to categorize the data records in a predetermined set, it can be used as attribute to label each record and for distinguishing elements belonging to the normal or malicious class [29]. Different types of classification techniques are decision tree induction, Bayesian networks-nearest neighbor classifier, genetic algorithm and fuzzy logic. As compared to the clustering technique, classification technique is less efficient in the field of intrusion detection. The main reason for this is the enormous amount of data needed to be collected to use classification. To classify the dataset into normal and abnormal, large amount of data is required to analyze its proximity. Classification method can be useful for both misuse detection and anomaly detection, but it is more commonly used for misuse detection.

Clustering Since the amount of available network data is too large, human labeling is time-consuming, and expensive. Clustering is the process of labeling data and assigning into groups. ie, Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of members from the same cluster are quite similar and members from the different clusters are different from each other. Hence clustering methods can be useful for classifying network data for detecting intrusions. Clustering algorithms can be classified into four groups: partitioning algorithm, hierarchical algorithm, density-based algorithm and grid based algorithm [30].

Clustering techniques can discover complex intrusions over a different time period. Clustering is an unsupervised machine learning mechanism for discovering patterns in unlabeled data with many dimensions. Clustering is the collection of patterns based on similarity. Patterns within a cluster are equivalent to each other, but they are different with other clusters. Therefore, patterns that are far from any of these clusters indicate that an unusual activity happened. That can be part of a new attack. Clustering can be applied on both Anomaly detection and Misuse detection.

2.11. Algorithms used to build optimize anomaly detection approach.

This section discusses the optimization algorithms used to build and optimize anomaly detection approach which can meet all the IDS requirements and objectives.

2.11.1. K-Means Algorithm

The term "k-means" was first used by James Macqueen in 1967. The idea though goes back to Hugo Steinhaus in 1957. The standard algorithm was first introduced by Stuart Lloyd in 1957 but it wasn't published outside Bell labs until 1982 [31].

In data mining, k-means clustering is a method of cluster analysis which aims to partition N observations into K clusters in which each observation belongs to the cluster having the nearest mean. This process

partitions the data space into Voronoi cells. In mathematics, a Voronoi diagram simply divides the space into a number of regions. The regions are called Voronoi cells [32]. Simply speaking it is an algorithm to group your objects based on attributes/features into K number of groups. The clustering is done by minimizing the sum of squares of distances between data and the cluster centroids.

Aim of K-mean clustering is simply to classify the data into K different clusters through the iteration and to converge it to a local minimum. In this process, data objects are grouped into disjoint clusters in such a way that the data in the same cluster is similar and data belonging to a different cluster is different.

So, the generated clusters are compact and independent. Euclidean distance is usually considered to determine the distance between data object and the cluster centroids.

➤ **K-Means Algorithm Process**

- The dataset is grouped into K clusters and the data points are randomly associated to the clusters resulting in clusters having almost same number of data points.
- For each data point: Find the distance between each data point and its respective cluster.

- Leave the data point if it is nearest to its own cluster. Otherwise move it into the closest cluster.
- Repeat the above step until centroids don't change their position anymore and no data point moving from one cluster to another. Now the clusters are stable and it marks the end of clustering process.

This is a very simple and reasonably fast algorithm. It is also efficient in processing large data sets like network traffic. The only difficulty is in comparing the quality of the clusters produced. Another limitation of k-means is that k should be specified in advance. But in Intrusion detection k is set to be two since there are two clusters for normal and anomalous data.

➤ **Advantages of K-Means Clustering Algorithm:**

1. This is a very simple algorithm.
2. It is reasonably fast algorithm.
3. K-Means may form tighter clusters than hierarchical clustering, especially in case of globular clusters are [33].

➤ **Disadvantages of k-means clustering:**

1. It is tough to compare the quality of the clusters produced. For instance, for different initial partitions or different values of K affect the final outcome.
2. Fixed number of clusters can make it problematic to predict the value of K [33].
4. The k-means result depends upon the data set. It works fine on some data sets, while fails on others.

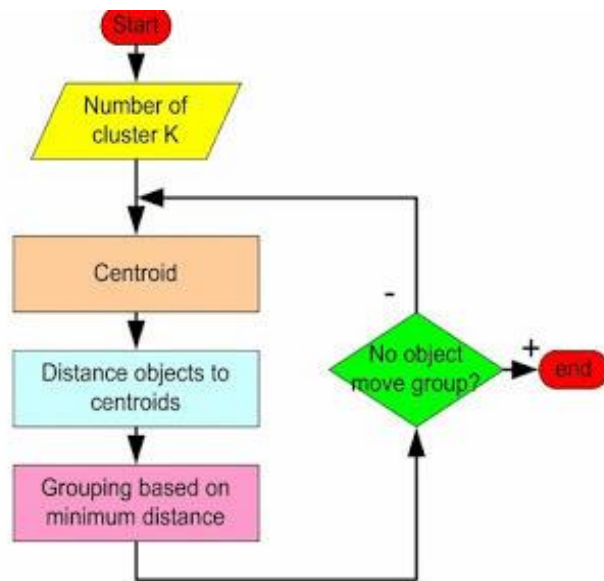


Figure (2.4) Flow chart demonstrate k-mean process

2.11.2. Sequential Minimal Optimization (SMO)

SVMs are starting to increasing adoption in the machine learning and computer vision research communities. However, SVMs have not yet

enjoyed widespread adoption in the engineering community. There are two possible reasons for the limited use by engineers. First, the training of SVMs is slow, especially for large problems. Second, SVM training algorithms are complex, subtle, large memory needed, and sometimes difficult to implement.

So, that new SVM learning algorithm is called Sequential Minimal Optimization (SMO) is invented.

Sequential Minimal Optimization (SMO) is a simple algorithm that quickly solves the SVM quadric programming QP problem without any extra matrix storage and without invoking an iterative numerical routine for each sub-problem. SMO decomposes the overall QP problem into QP sub-problems. SMO chooses to solve the smallest possible optimization problem at every step.

For the standard SVM QP problem, the smallest possible optimization problem involves two Lagrange multipliers because the Lagrange multipliers must obey a linear equality constraint.

At every step, SMO chooses two Lagrange multipliers to jointly optimize, finds the optimal values for these multipliers, and updates the SVM to reflect the new optimal values.

The advantage of SMO lies in the fact that solving for two Lagrange multipliers can be done analytically. Thus, an entire inner iteration due to numerical QP optimization is avoided. The inner loop of the algorithm can be expressed in a small amount of C code, rather than invoking an entire

iterative QP library routine. Even though more optimization sub-problems are solved in the course of the algorithm; each sub-problem is so fast that the overall QP problem can be solved quickly.

In addition, SMO does not require extra matrix storage (ignoring the minor amounts of memory required to store any 2x2 matrices required by SMO). Thus, very large SVM training problems can fit inside of the memory of an ordinary personal computer or workstation. Because manipulation of large matrices is avoided, SMO may be less susceptible to numerical precision problems [34].

2.11.3. Genetic search algorithm.

Genetic algorithms are a class of stochastic search algorithms based on biological evolution, given a clearly defined problem to be solved and a binary string representation for candidate solutions [49].

A GA applies the following major steps

- Step 1:** Represent the problem variable domain as a chromosome of a fixed length, choose the size of a chromosome population N , the crossover probability p_c and the mutation probability p_m .
- Step 2:** Define a fitness function to measure the performance, or fitness, of an individual chromosome in the problem domain. The fitness function establishes the basis for selecting chromosomes that will be mated during reproduction.
- Step 3:** Randomly generate an initial population of chromosomes of size N :

$$x_1, x_2, \dots, x_N$$

Step 4: Calculate the fitness of each individual chromosome:

$$f(x_1), f(x_2), \dots, f(x_N)$$

- Step 5:** Select a pair of chromosomes for mating from the current population. Parent chromosomes are selected with a probability related to their fitness. Highly fit chromosomes have a higher probability of being selected for mating than less fit chromosomes.
- Step 6:** Create a pair of offspring chromosomes by applying the genetic operators – crossover and mutation.
- Step 7:** Place the created offspring chromosomes in the new population.
- Step 8:** Repeat Step 5 until the size of the new chromosome population becomes equal to the size of the initial population, N .
- Step 9:** Replace the initial (parent) chromosome population with the new (offspring) population.
- Step 10:** Go to Step 4, and repeat the process until the termination criterion is satisfied.

GA represents an iterative process. Each iteration is called a generation. A typical number of generations for a simple GA can range from 50 to over 500. The entire set of generations is called a run. At the end of a run, we expect to find one or more highly fit chromosomes.

Because GAs use a stochastic search method, the fitness of a population may remain stable for a number of generations before a superior chromosome appears. This makes applying conventional termination criteria problematic. A common practice is to terminate a GA after a specified number of generations and then examine the best chromosomes in the population. If no satisfactory solution is found, the GA is restarted.

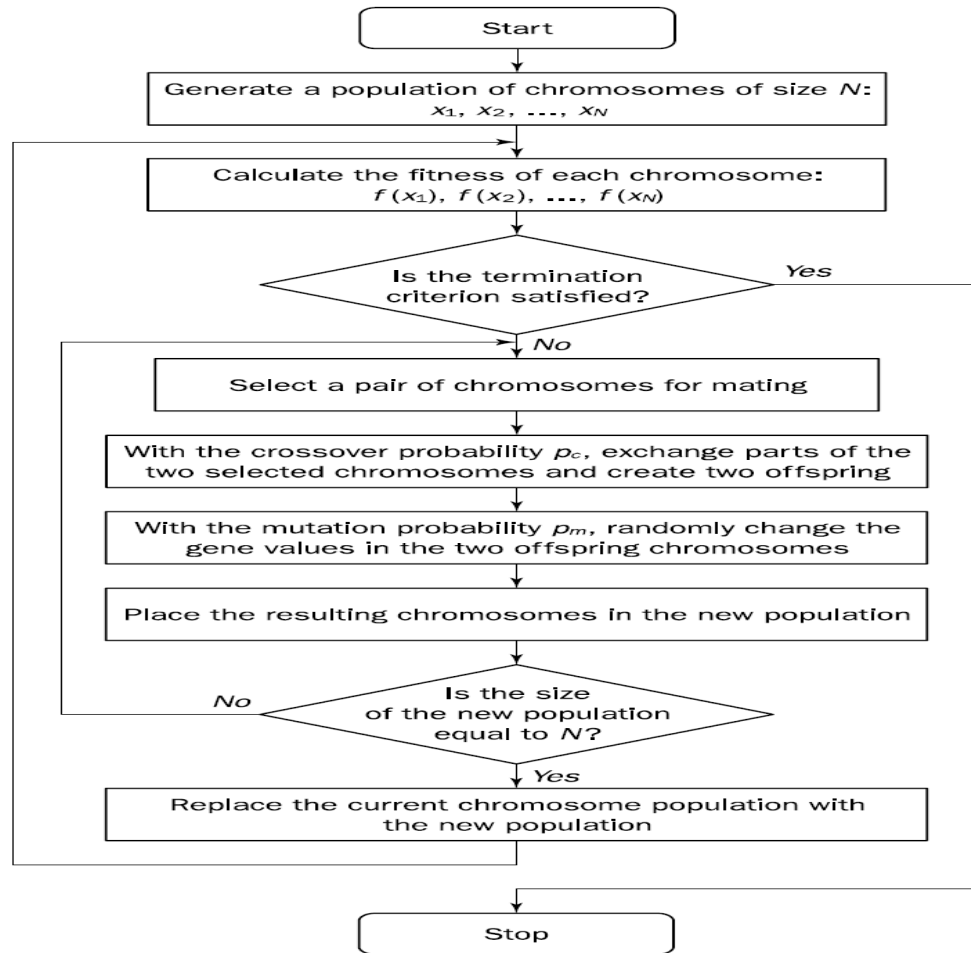


Figure (2.5) Flow chart demonstrate Genetic search algorithm process

2.12. Related Work:

This section presents a literature survey of various models and techniques used to detect the intrusion and demonstrate How IDS developed and various kinds of changes take place in existing and new models

C.Taylor et al. [35] introduce a low-cost approach based on clustering and multivariate analysis known as NATE - Network Analysis of Anomalous Traffic Events. This resolves the problem of those IDS who was not able to handle high volume traffic and real time detection constraint. A purpose solution was like any other light weighted approach with the quality feature of minimal network traffic measurement, limited attack scope and anomaly detection. NATE model performed on MIT Lincoln lab's data. It consists of two phases of operation. In Phase-I data collection and a database creation were performed. In this phase collected data were closely analyzed for possible attack and imagine that only normal data was captured. But in reality, if in Phase-I intrusion was treated as a normal than it was a big problem for further detection. While in Phase-II detects intrusion in real-time environment. This classification of normal and abnormal data was performed on the basis of cluster algorithm. In this paper provide an idea about clustering, which was performed on the real-time traffic so that easily and quickly updating of the new features of the attacks in the database.

U.Lindqvist et al. [36] this paper introduces a tool set known as production based expert system tool set (P-BEST). This tool set works for misuse detection and developed a new signature corresponding to the attacks. Basically, this was the advanced version of the IDS that was used

for research purpose. P-BEST provides a better mechanism and performance to detect intrusion in real time environment. This was integrated with the c-programming which make it easy to use, powerful and flexible. But this tool was some limitation. This was less capable of detecting attacks or intrusion where data were incomplete, inaccurate and uncertain. P-BEST was more feasible in the known environment. Tool set was not able to generate new form of attacks in this never know how that particular attack was performed by an attacker.

Minegishi, T et al. [37]. proposed the use of a data mining framework for building intrusion detection models. This framework consists of classification, association rules, and frequent episodes' programs, which can be used to automatically construct detection models. They used the set of relevant system features to compute inductively learned, process raw audit data from the send mail system, call data and the network TCP dump data and then summarized into connection records attributes.

This approach applies on two general data mining algorithms: association rules algorithm, and the frequent episode's algorithm.

Minegishi, T et al. [38]. Proposed Audit Data Analysis and Mining (ADAM), a real-time anomaly detection system that uses data mining techniques to detect intrusions, ADAM uses a combination of association rules mining and classification to discover attacks in a TCP dump. ADAM uses a classifier which has been previously trained to classify the suspicious connections as a known type of attack, unknown type or a normal connection.

ADAM performs anomalies detection in two phases: the training phase, and the testing phase. In the training phase, it uses a data stream for which we know the types of the attack. The attack-free parts of the stream are fed into a module that performs off-line association rules discovery. The output of this module is a profile of rules that we call “normal”. The profile along with the training data set is also fed into a module that uses a combination of a dynamic, on-line algorithm for association rules, whose output consists of frequent item sets that characterize attacks to the system. These item sets, along with a set of features extracted from the data stream by a feature selection module are used as the training set for a classifier “decision tree”. This whole phase takes place off-line before using the system to detect intrusions. In the testing phase, the actual detection of intrusions is implemented.

Barbará, Det al[39]. proposed the combination of multiple host-based detectors using decision tree. This method uses conventional measures for intrusion detection and modeling methods appropriate to each measure. Statistical Rule-based method is used to model these measures which are combined with decision tree. The proposed detection method has a good performance because it can model normal behaviors from various perspectives, the decision tree used here is the C4.5 algorithm. The result shows that the combined detection method dramatically reduces the false-positive error rate against various types of intrusion.

Zhang et al. [40] in this paper a hybrid approach based on both misuse and anomaly detection was implemented on the NIDS. The main objective of this paper was to reduce the limitation of both detection

techniques by combining with each other. In this paper for detecting the intrusion random forest data mining techniques were implemented. Firstly, random forest implemented for misuse detection in real time. After that it was implemented on the anomaly detection for detecting unknown attacks in an off-line mode. So by combining the two approaches the false alarm ratio and overall performance has been improved.

TG.Dietterich et al. [41] this paper introduces bagging; boosting and randomization technique was used to improve the effectiveness of the decision tree algorithm. Bagging and boosting generate a different range of classifiers by manipulating the training data which provided to the learning algorithm as a base. Bagging shows that if there was a substantial classification noise than randomization technique performs better. This paper describes that bagging performed better than another classifier if there was a noise dataset which is a great advantage of detecting the noise traffic in a real-time environment.

AP.Muniyandi et al.[42] a semi-supervised learning technique was used in this paper. Firstly, unsupervised learning using K-means clustering. In which part of training instances, was trained using the Euclidean distance method. After that supervised learning performed using C4.5 algorithm. By applying the clustering the boundary was refined, it helps the C4.5 algorithm to detect anomalies with more accuracy. This semi-supervised learning technique performs better than unsupervised or supervised learning technique. But limitation was that it takes more time than simple classification or clustering. It was a disadvantage while detecting the real-time traffic.

Panda, M. and Patra, M.R[43] designed a fuzzy logic-based system for effectively identifying the intrusion activists within a network. The fuzzy logic-based system helps to detect an intrusion behavior of the networks, since the rule base contains a better set of rules. The system uses mechanical method for generation of fuzzy rules that are obtained from the definite rules using frequent items. The experiments and evaluations of this intrusion detection system are performed with the KDD CUP 99 Intrusion detection dataset. The experiments results show that fuzzy logic-based system achieved higher precision in identifying whether the records are normal or attack.

S. Ahmed[44] proposed the use of two algorithms: back propagation algorithm and C4.5 algorithm for intrusion detection. Since these algorithms are mainly applicable to misuse detection, it shows that the definition of anomaly detection not only takes into account normal profiles, but also handles known attacks and explores supervised machine learning algorithm; particularly neural networks and decision trees for intrusion detection. In fact, decision trees induction algorithm has proved its efficiency in predicting the different classes of unlabeled data in Knowledge Discovery Databases (KDD CUP99). Test data set contains attacks such as Denial of Service (DoS), probe, User to Root (U2R) and Remote to Local (R2L). Experimental results demonstrate that while neural networks are highly successful in detecting known attacks, decision trees are more interesting to detect new attacks.

Hatimet al [45] proposed a hybrid machine learning technique for network intrusion detection based on combination of K-means clustering

and support vector machine classification. The aim of this research is to reduce the rate of false positive alarm, false negative alarm rate and to improve the detection rate. The NSL-KDD dataset has been used in the proposed technique. In order to improve classification performance, some steps have been taken on the dataset. The classification has been performed by using support vector machine. After training and testing the proposed hybrid machine learning technique, the results have shown that the proposed technique has achieved a positive detection rate and reduce the false alarm rate.

CHAPTER THREE

METHODOLOGY

3.1 Overview:

To provide an appropriate solution in network anomaly detection, we need the concept of normality. The idea of normal is usually introduced by a formal model that expresses relations among the fundamental variables involved in the system dynamics. Consequently, an event or an object is detected as anomalous if its degree of deviation with respect to the profile or behavior of the system, specified by the normality model.

In this section, a new anomaly detector approach is proposed based on using k-means clustering algorithm and Sequential Minimal Optimization (SMO) algorithm to detect online network anomaly detection, the proposed approach aims to generate a suitable number of detectors with high detection rate and accuracy.

The main idea is based on using feature selection in preprocessing phase to reduce the number of dataset, The ConsistencySubsetEval and Genetic search algorithms have been applied to select specific features from the dataset and remove those features which are irrelevant before clustering and classification phases, after attribute selection k-means clustering algorithm selected to reduced training dataset in order to decrease time and processing complexity. In classification phase, supervised algorithm called

Sequential Minimal Optimization (SMO) selected to improve the quality of detection. The main stages are shown in Figure. (3.1), and a detailed description of each stage is presented below.

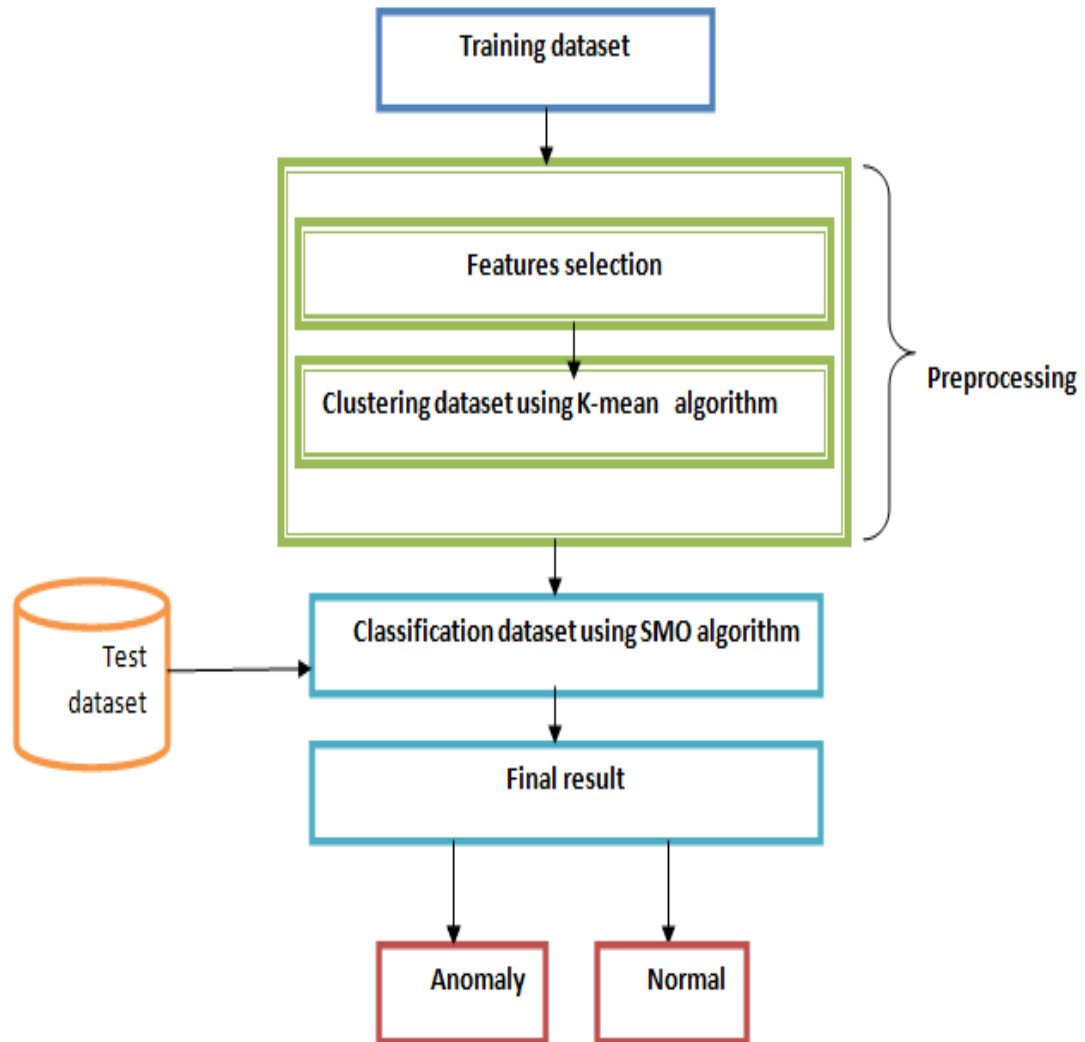


Figure. (3.1) model diagram

3.2. Training phase:

The training data contains both normal and abnormal data, we assume that attack data will not occur frequently as normal data would. Hence, less than $x\%$ of data is anomalous.

3.3. Pre-processing:

Pre-processing of original NSL-KDD intrusion data set is an important phase to make it as an appropriate input for classification phase. The main objective of preprocessing phase is to reduce ambiguity and provide accurate information to detection engine. The preprocessing phase cleans the network data by grouping, labeling and it handles the missing or incomplete dataset. The dataset pre-processing is achieved by applying the following stages sequentially.

3.3.1. Features Selection:

Features selection is the most critical stage in building a hybrid intrusion detection models and is equally important to improve the efficiency of data mining algorithms. In general, the input data to classifiers is in a high dimension feature space but not all of the features are relevant to the classes to be classified. Some of the data includes irrelevant, redundant or noisy features. In this case, irrelevant and redundant features can introduce noisy data that distract the learning algorithm. It decreases the number of attributes, eliminates irrelevant, noisy or redundant features and brings about effects on applications such as speeding up a data mining

algorithm, improving learning accuracy and leading to better model comprehensibility.

Feature selection process is demonstrated in Figure (3.2) on the left there are the features ($F_0 \dots F_N$) that are available from the data monitored, which is, for example, from network traffic. On the right side is the output ($F_0 \dots F_M$) of the selection tool. The number of features in the output varies based on the selection tool used and the inter-correlation of features in the input. Following the basic principles of feature analysis, the number of features in the output (M in Figure 3.2) is in most of the cases less than the number of features in the input (N in Figure 3.2). However, it is possible that the output is equal to the input.

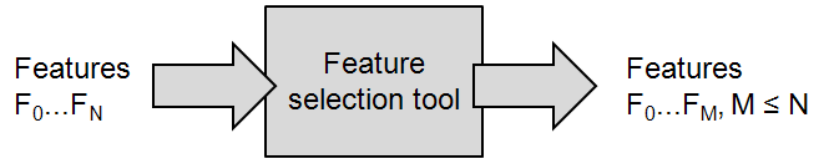


Figure (3.2) Feature selection

During this step, the set of attributes or features deemed to be the most effective attributes which are extracted in order to construct suitable detection system.

The goal of features selection increases the detection rate and decreases the false alarm rate in network intrusion detection. WEKA 3.6 which is a machine learning tool has been used to compute the features selection subsets for hybrid approach (K-mean +SMO) to test the classification performance on each of these feature sets.

The ConsistencySubsetEval and Genetic search algorithms have been applied to select specific features from the dataset and remove those features which are irrelevant before clustering and classification phases, the result shown below:

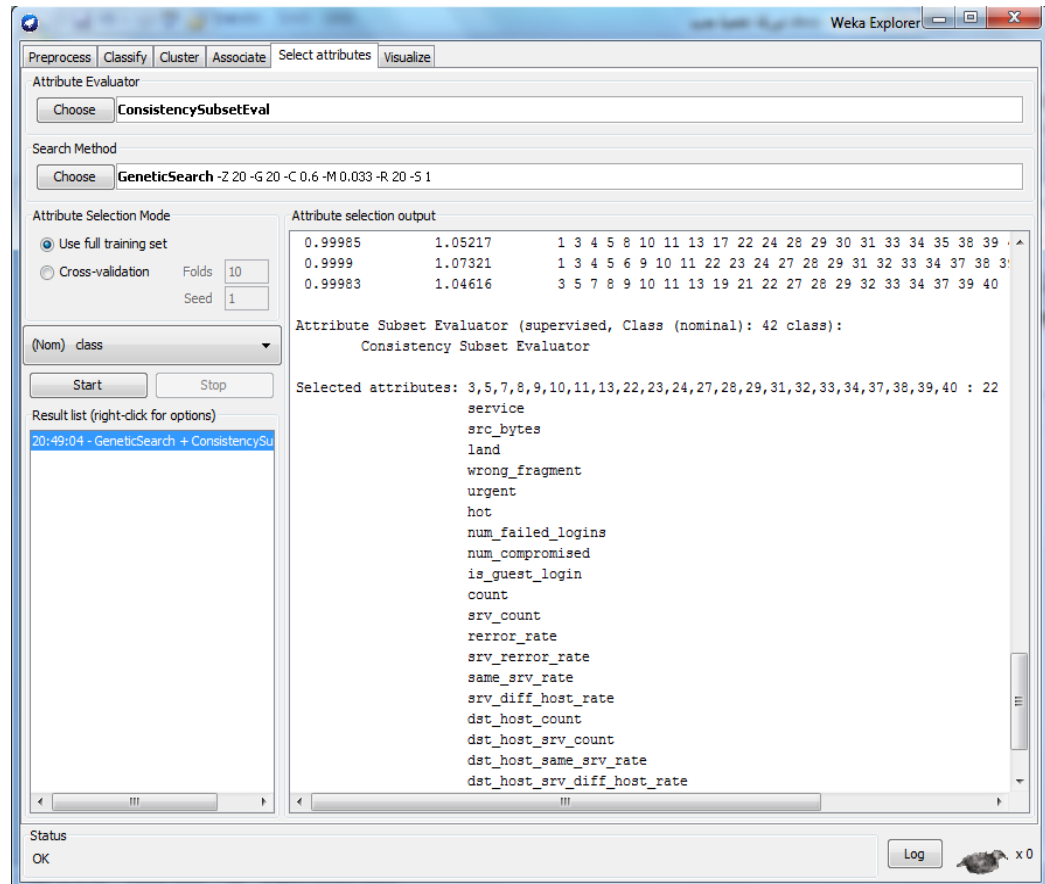


Figure (3.3) Feature selection Consistency SubsetEval and Genetic search algorithms

Selected attributes: 3,5,7,8,9,10,11,13,22,23,24,27,28,29,31,32,33,34,37,38,39,40 : 22
service
src_bytes
land
wrong_fragment
urgent
hot
num_failed_logins
num_compromised
is_guest_login
count
srv_count
rerror_rate
srv_rerror_rate
same_srv_rate
srv_diff_host_rate
dst_host_count
dst_host_srv_count
dst_host_same_srv_rate
dst_host_srv_diff_host_rate
dst_host_serror_rate
dst_host_srv_serror_rate
dst_host_rerror_rate

Figure (3.4) Name of Feature selected

3.3.2. Clustering phase:

The clustering phase done by applying the K-means clustering algorithm, two clusters were specified and created. As the algorithm iterates through the training data, each cluster's architecture transferred to another. The updating of clusters causes the values of the centroids to modify. This change is a reflection of the current cluster elements. When there are no changes to any cluster, the clustering of the K-Means algorithm becomes complete.

3.4. Testing phase:

Analyses the traffic generated on the network based on the information gathered from the testing phase.

3.5. Classification phase:

Last phase is classification in this phase supervised algorithm Sequential Minimal Optimization (SMO) was used to classify dataset to normal or anomaly.

CHAPTER FOUR

EXPERIMENT IMPLEMENTATION AND PERFORMANCE EVALUATION:

4.1 Overview:

The following experiment done by using WEKA (WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS). WEKA is a Tool for Data Mining and Machine Learning which was implemented at the University of Waikato, in New Zealand in the year 1997. WEKA is a set of Machine Learning and Data Mining algorithms. This WEKA software is programmed in JAVA language and it has a GUI Interface to interact with data Files. With 49 data pre-processing tools WEKA tool contains 76 classification algorithms, 15 attribute evaluators and ten search algorithms for feature selection. There are three algorithms to find association rules. It also has three Graphical User Interfaces: "The Explorer", "The Experimenter" and "The Knowledge Flow." The file format to store data in WEKA is ARFF. Meaning of ARFF is Attribute Relation File Format. It also includes tools for visualization. It has a several panels that can be used to perform precise tasks. WEKA has the ability to expand and contain the new algorithms for Machine Learning in it. These expanded algorithms can directly be applied to dataset [46].

4.2. Dataset Description:

Various drawbacks of KDD CUP 99 which was the main cause to decrease in the performance of various IDS led to the invention of NSL KDD dataset. NSL KDD is the refined version and also called the successor of KDD CUP dataset. It contain (KDDTrain.arff = 125973, KDDTest. arff = 22543), It consists of all the needed attributes from KDD CUP dataset. It is an open source data and can be downloaded easily. The advantage of using this dataset is redundant record is removed and sufficient number of records is present for train and test data. It consists of 41 attributes which is classified under Nominal, Binary and Numeric [47]. One attribute is added as class which is 42 and attribute.

Table (4.1): NSL KDD dataset features [47]

Type	Feature
Nominal	protocol_type, service, flag
Binary	land, logged_in, root_shell, su_atempted, Is_host_login, is_guest_login
Numerical	duration, src_bytes, dst_bytes, wrong_fragment, urgent, hot, num_failed_login, num_compromised, num_root, num_file_creations, num_shells, num_access_files, num_outbound_cmds, count, srv_count, serror_rate, srv_serror_rate, rerror_rate, srv_rerror_rate, same_srv_rate, diff_srv_rate, srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_diff_src_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate

There are two types of class called Normal and Anomaly. Anomaly class can be further divided into DOS, PROBE, R2L and U2R as shown in Figure 2. For experiment purpose, only two classes are considered: Normal and Anomaly.

Table NO. (4.2): anomaly types that include in NSL KDD dataset [47]

DOS	PROBE	R2L	U2R
Back, Land, Neptune, Pod, Smurf, Teardrop, Apache2, Udpstorm, Processtable, Worm	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint	Guess_passwd, Ftp_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httpunnel, Sendmail, named	Buffer_Overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps.

4.2.1. Brief description of the attributes of the NSL-KDD dataset[48]:

Table. (4.3):description of the attributes of the NSL-KDD dataset

Attribute No.	Attribute Name	Description
1	Duration	Length of time duration of the connection
2	Protocol_type	Protocol used in the connection
3	Service	Destination network service used
4	Flag	Status of the connection – Normal or Error
5	Src_bytes	Number of data bytes transferred from source to destination in singleconnection
6	Dst_bytes	Number of data bytes transferred from destination to source in singleconnection
7	Land	if source and destination IP

		addresses and port numbers are equal then, this variable takes value 1 else 0
8	Wrong_fragment	Total number of wrong fragments in this connection
9	Urgent	Number of urgent packets in this connection. Urgent packets are packets with the urgent bit activated
10	Hot	Number of „hot“ indicators in the content such as: entering a system directory, creating programs and executing programs
11	Num_failed_logins	Count of failedlogin attempts
12	Logged_in	Login Status : 1 if successfully logged in; 0otherwise
13	Num_compromised	Number of ``compromised' ' conditions
14	Root_shell	1 if root shell isvobtained; 0 otherwise
15	Su_attempted	1 if ``su root' command attempted or used; 0 otherwise
16	Num_root	Number of ``root" accesses or number of operations performed as a root in the connection
17	Num_file_creation s	Number of file creation operations in the connection
18	Num_shells	Number of shell prompts
19	Num_access_files	Number of operations on access control files
20	Num_outbound_cmds	Number of outbound commands in an ftp session
21	Is_hot_login	1 if the login belongs to the``hot" list i.e., root or admin; else 0
22	Is_guest_login	1 if the login is a ``guest" login; 0 otherwise
23	Count	Number of connections to the same destination host as the currentconnection in the past two

		second
24	Srv_count	Number of connections to the same service (port number) as the currentconnection in the past two seconds
25	Error_rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated incount (23)
26	Srv_error_rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in srv_count (24)
27	Rerror_rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in count (23)
28	Srv_rerror_rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in srv_count (24
29	Same_srv_rate	The percentage of connections that were to the same service, among theConnections aggregated in count (23)
30	Diff_srv_rate	The percentage of connections that were to different services, among theConnections aggregated in count (23)
31	Srv_diff_host_rate	The percentage of connections that were to different destination machinesamong the connections aggregated in srv_count (24
32	Dst_host_count	Number of connections having the same destination host IP address
33	Dst_host_srv_count	Number of connections having the same port number
34	Dst_host_same_srv_rate	The percentage of connections that were to the same service, among

		the connections aggregated in dst_host_count (32)
35	Dst_host_diff_srv_rate	The percentage of connections that were to different services, among the connections aggregated in dst_host_count (32)
36	Dst_host_same_src_port_rate	The percentage of connections that were to the same source port, among the connections aggregated in dst_host_srv_count (33)
37	Dst_host_srv_diff_host_rate	The percentage of connections that were to different destination machines, among the connections aggregated in dst_host_srv_count (33)
38	Dst_host_serro r_rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_count(32)
39	Dst_host_srv_s error_rate	The percent of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_srv_count (33)
40	Dst_host_rerror_rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_count (32)
41	Dst_host_srv_rerror_rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_srv_count (33)

4.3. Accuracy Measure of individual algorithms (SMO and K-mean):

Detection of attack can be measured by following metrics:

1. **False positive (FP)**: Or false alarm corresponds to the number of detected attacks but it is in fact normal.

2. **False negative (FN)**: Corresponds to the number of detected normal instances but it is actually attack, in other words these attacks are the target of intrusion detection systems.

3. **True positive (TP)**: Corresponds to the number of detected attacks and it is in fact attack.

4. **True negative (TN)**: Corresponds to the number of detected normal instances and it is actually normal [8].

Here We can use three parameters of measurement (Decoction rate (**DTR**), False positive rate alarm (**FPR**), Accuracy (**AC**)).

Detection rate (DTR), It is defined as the ratio of detecting attacks to total number of attacks. This is the best parameter to measure the performance of the model, is determined using the equation:

$$\text{Decoction rate} = \text{DTR} = \frac{TP}{TP+FN} * 100 \dots \dots (4.1)[45]$$

False Positive Rate (FPR): This is one of the main parameters to find out the effectiveness of various models and also the major concern while network setup. A normal data is considered as abnormal or attack type data. FPR is obtained using the following formula:

$$\text{False positive rate alarm} = \text{FPR} = \frac{FP}{TN+FP} * 100 \dots \dots (4.2) [45]$$

Accuracy (AC) is the proportion of the total number of the correct predictions to the actual data set size, It is determined using the equation:

$$\text{Accuracy} = AC = \frac{TP+TN}{TP+TN+FP+FN} * 100 \dots\dots\dots (4.3) [45]$$

4.3.1. Accuracy Measure of individual algorithms (SMO):

In order to perform data analysis and prediction of algorithms that used to build our model.

Firstly apply SMO algorithm for NLS-Kdd dataset with 22 attributes (with feature selection) using WEKA, the result shown below:

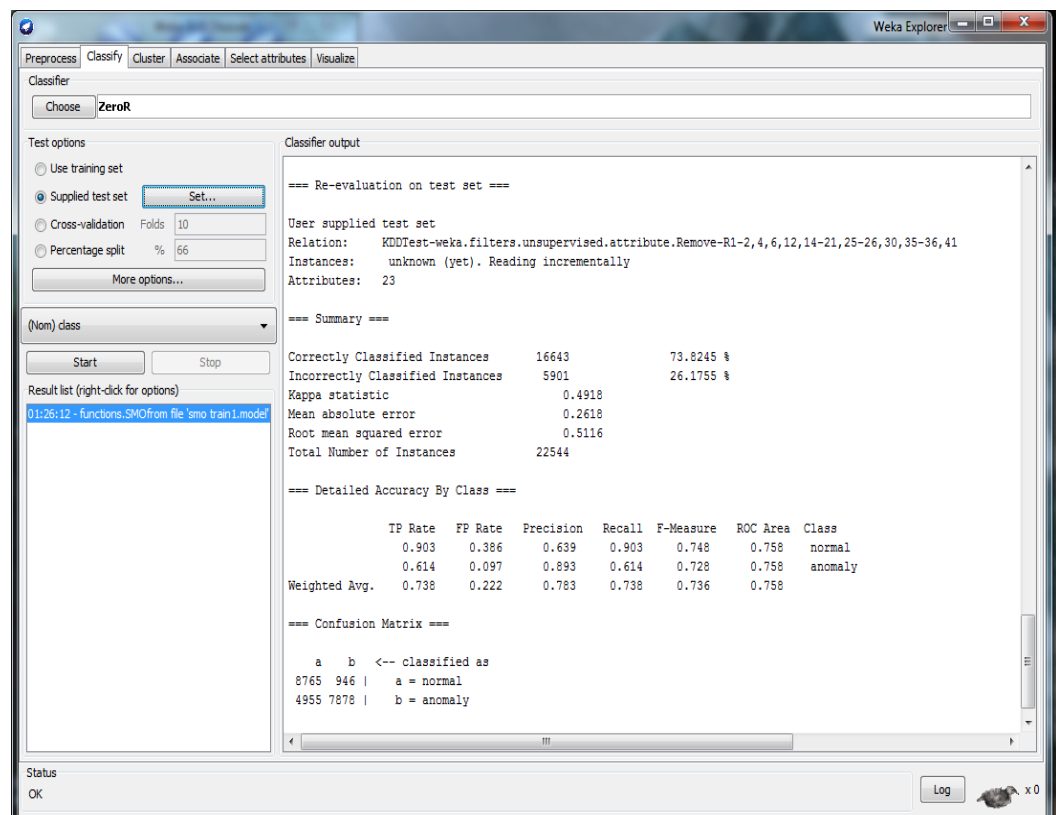


Figure (4.1) SMO result using dataset with 22 attributes

Refer to the results in figure (4.1) the details of accuracy shown in table (4.4) below.

Table. (4.4) demonstrate details of accuracy parameters for SMO

TPRate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.903	0.386	0.639	0.903	0.748	0.758	normal
0.614	0.097	0.893	0.614	0.728	0.758	anomaly
0.738	0.222	0.783	0.738	0.736	0.758	Weighted Avg

Correctly Classified Instances 16643 = 73.8245%

Incorrectly Classified Instances 5901 = 26.1755%

Confusion Matrix :

Also from the results in figure (4.1) the details of confusion matrix shown in table (4.5) below.

Table.(4.5) demonstrate Confusion Matrix for SMO.

a	b	Classified as
TN = 8765	FP =946	a= normal
FN = 4955	TP = 7878	b= anomaly

Now calculate the measurement parameters for **SMO**:

$$\text{Decoction rate} = \text{DTR} = \frac{TP}{TP+FN} * 100$$

$$\text{DTR} = \frac{7878}{7878+4955} * 100 = 61.39$$

$$\text{False positive rate alarm} = \text{FPR} = \frac{FP}{TN+FP} * 100$$

$$\text{FPR} = \frac{946}{8765+946} * 100 = 9.7$$

$$\text{Accuracy} = \text{AC} = \frac{TP+TN}{TP+TN+FP+FN} * 100$$

$$\text{AC} = \frac{7878+8765}{7878+ 8765+946+4955} * 100 = 73.82$$

Table.(4.6) demonstrate measurement parameters for SMO

algorithm	DTR	FPR	AC
SMO	61.39	9.7	73.82

4.3.2. Secondly: apply K-mean algorithm:

for NLS-Kdd dataset with 22 attributes (with feature selection) using WEKA, the result shown below:

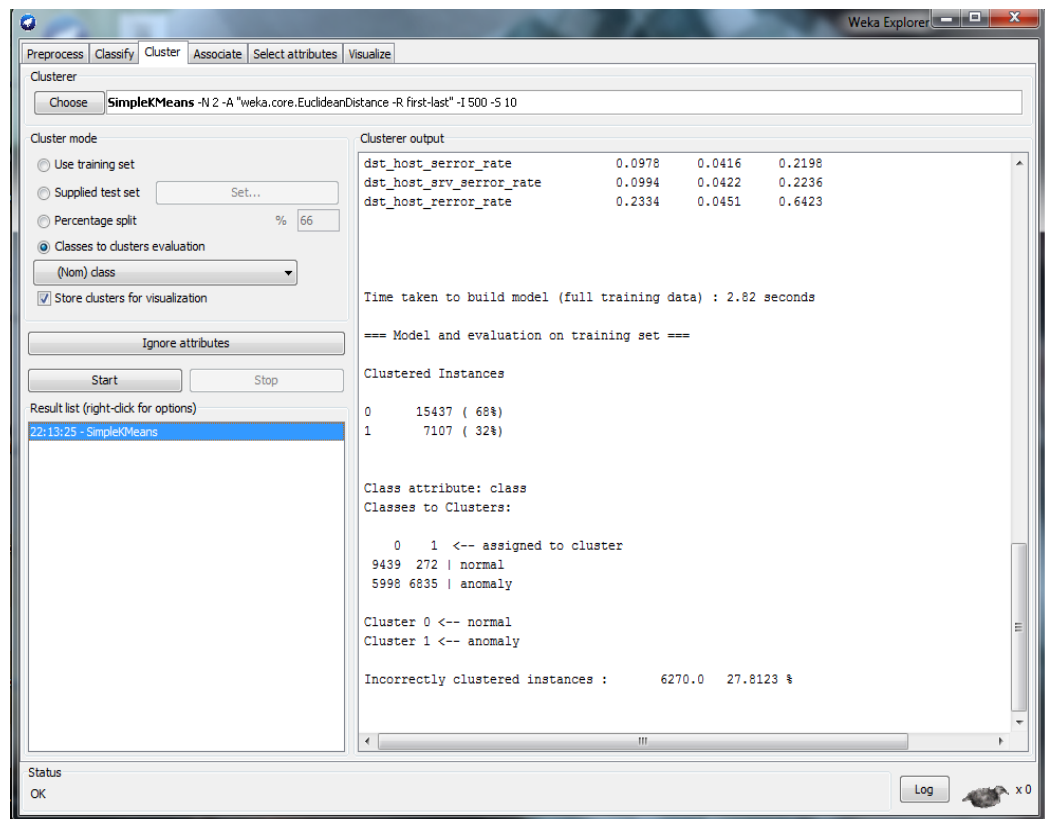


Figure (4.2) K-mean clustering result using dataset with 22 attributes

Correctly Classified Instances = 72.1877%

Incorrectly Classified Instances = 27.8123%

Refer to the results in figure (4.2) the details of confusion matrix shown in table (4.7) below.

Table.(4.7) demonstrate Confusion Matrix for K-mean

0	1	Classified as
TN=9439	FP= 272	0= normal
FN=5998	TP=6835	1= anomaly

$$\text{Decoction rate} = \text{DTR} = \frac{TP}{TP+FN} * 100$$

$$\text{DTR} = \frac{6835}{6835+5998} * 100 = 53.26$$

$$\text{False positive rate alarm} = \text{FPR} = \frac{FP}{TN+FP} * 100$$

$$\text{FPR} = \frac{272}{9439+272} * 100 = 2.8$$

$$\text{Accuracy} = \text{AC} = \frac{TP+TN}{TP+TN+FP+FN} * 100$$

$$\text{AC} = \frac{6835+9439}{6835+9439+272+5998} * 100 = 72.1877$$

Table.(4.8) demonstrate measurement parameters for K-mean

algorithm	DTR	FPR	AC
K-mean	53.26	2.8	72.188

4.3.3. Thirdly: apply hybrid approach (K-mean + SMO):

for NLS-Kdd dataset with 22 attributes (with feature selection) using WEKA, the result shown below:

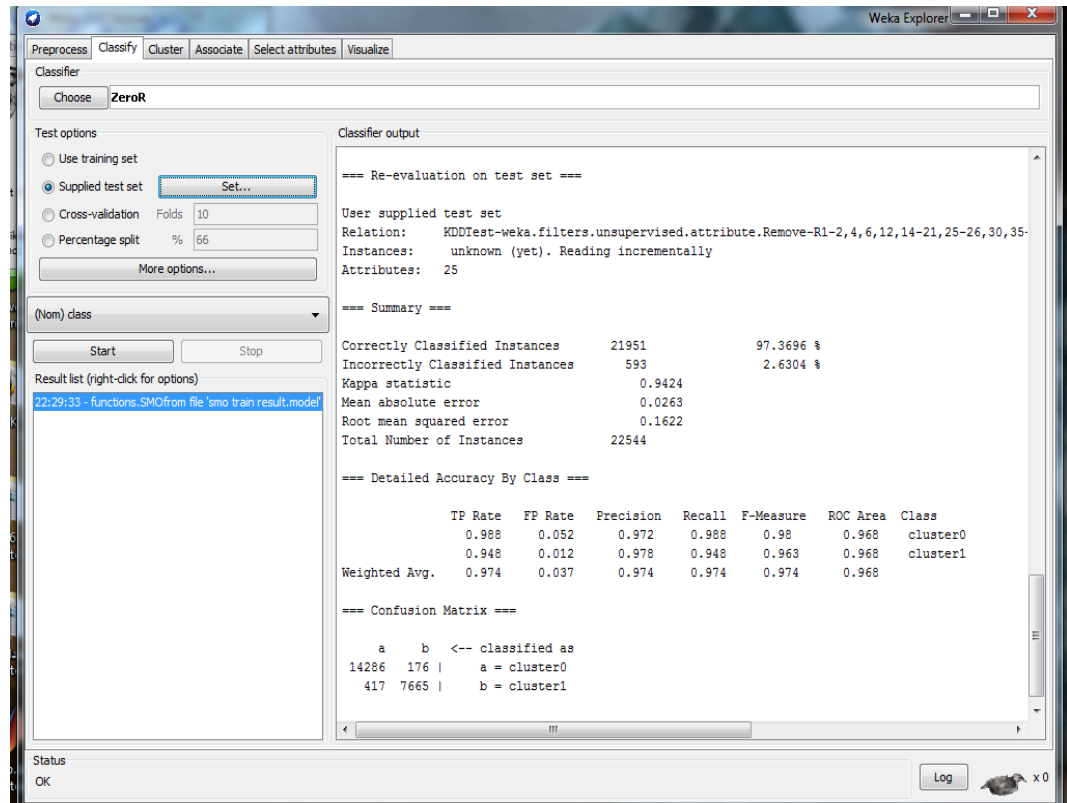


Figure (4.3) (K-mean + SMO) result using dataset with 22 attributes

Refer to the results in figure (4.3) the details of accuracy shown in table (4.9) below.

Table.(4.9) demonstrate details of accuracy parameters for (K-mean+SMO)

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.988	0.052	0.972	0.988	0.98	0.968	normal
0.948	0.012	0.978	0.948	0.963	0.968	anomaly
0.974	0.037	0.974	0.974	0.974	0.968	Weighted Avg

Correctly Classified Instances 21951= 97.3696%

Incorrectly Classified Instances 593 = 2.6304%

Confusion Matrix :

Refer to the results in figure (4.3) the details of confusion matrix shown in table (4.10) below.

Table.(4.10) demonstrate Confusion Matrix for(K-mean+SMO)

a	b	Classified as
TN =14286	FP = 176	a= normal
FN = 417	TP = 7665	B= anomaly

$$\text{Decoction rate} = \text{DTR} = \frac{TP}{TP+FN} * 100$$

$$\text{DTR} = \frac{7665}{7665+417} * 100 = 94.84$$

$$\text{False positive rate alarm} = \text{FPR} = \frac{FP}{TN+FP} * 100$$

$$\text{FPR} = \frac{176}{14286+176} * 100 = 1.2$$

$$\text{Accuracy} = \text{AC} = \frac{TP+TN}{TP+TN+FP+FN} * 100$$

$$\text{AC} = \frac{7665+14286}{7665+14286+417+176} * 100 = 97.3695$$

Table.(4.11) demonstrate measurement parameters for (K-mean+SMO)

algorithm	DTR	FPR	AC
K-mean + SMO	94.48	1.2	97.3695

4.4. Comparison between (SOM , K-meam, hybrid (K-mean + SMO))

Table.(4.12) demonstrate Comparison of measurement parameters for (K-mean+SMO)

algorithm	DTR	FPR	AC
SMO	61.39	9.7	73.82
K-mean	53.26	2.8	72.188
K-mean + SMO	94.48	1.2	97.3695

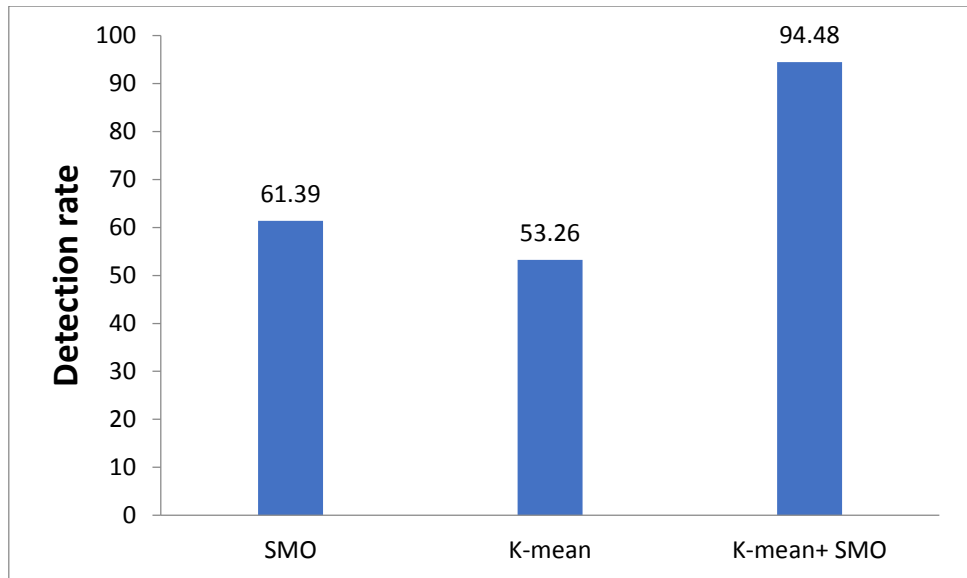


Figure (4.4): compression of detection rate for (SMO, K-mean, K-mean + SMO)

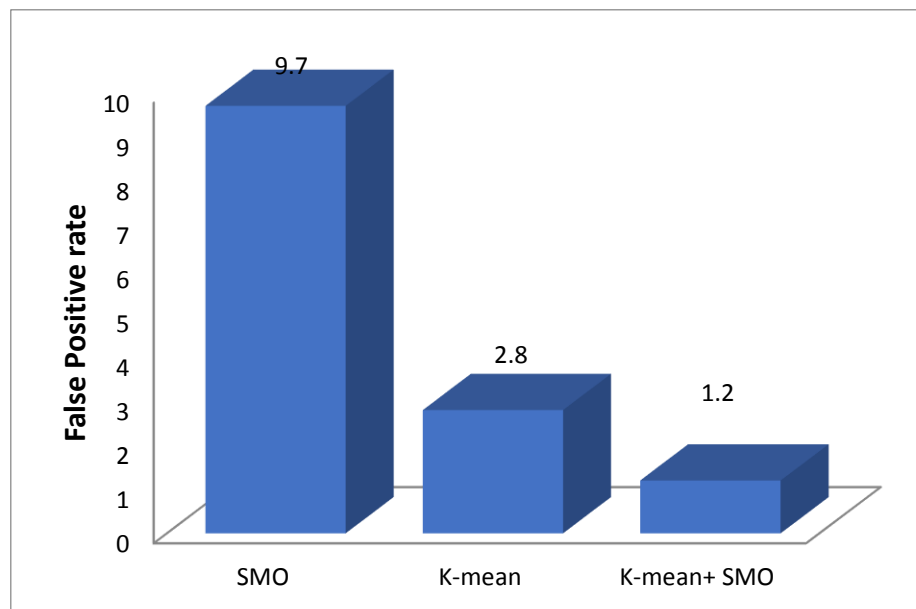


Figure (4.5): compression of false positive rate for (SMO, K-mean, K-mean + SMO)

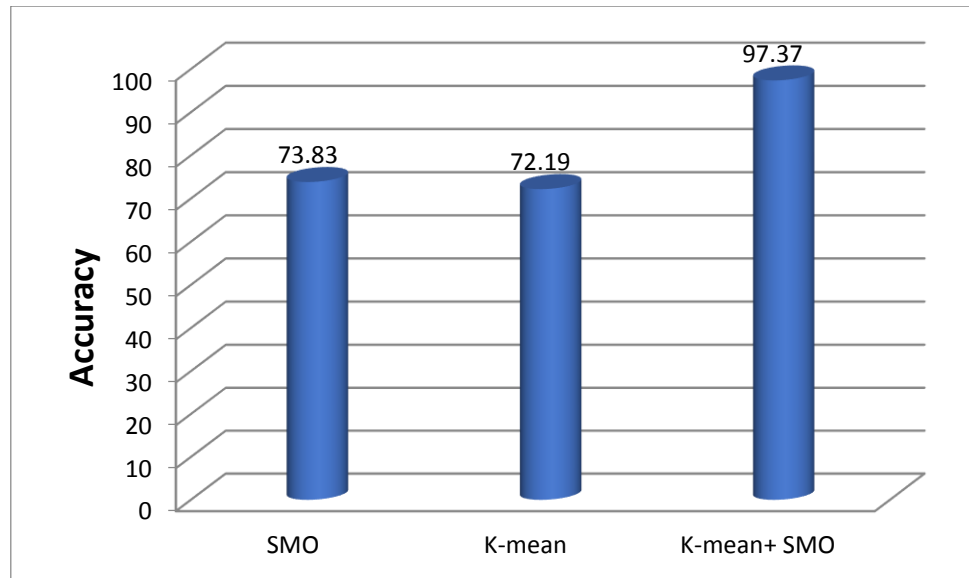


Figure (4.6): compression of accuracy for (SMO, K-mean, K-mean + SMO)

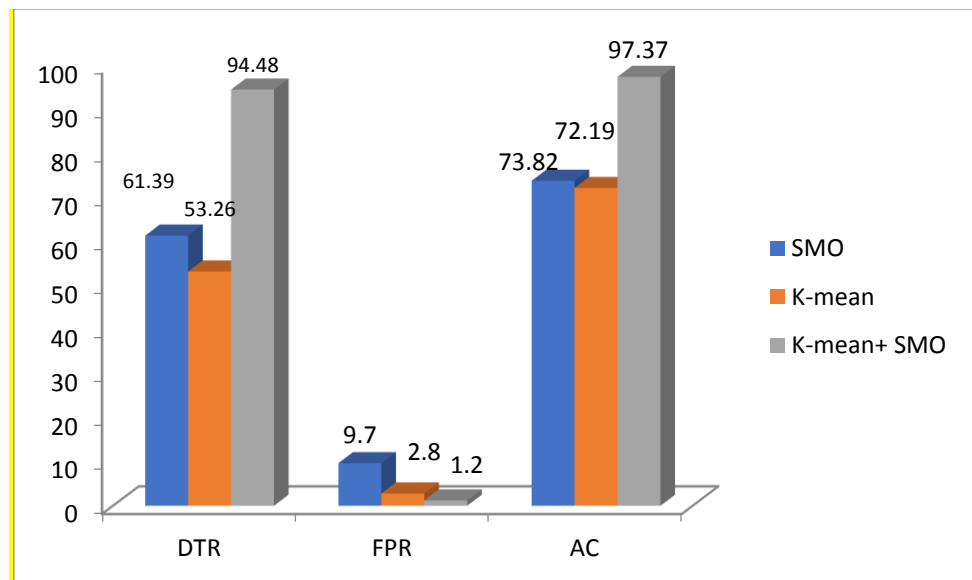


Figure (4.7) demonstrate Comparison of measurement parameters for (SMO, K-mean, K-mean+SMO)

Refer to above Figures the experimentation results show that the proposed model (K-mean+SMO) is more accurate as compare to other individual data mining techniques (SMO= 73.82) and (K-mean =72.188). The accuracy of proposed model (K-mean+SMO) is 97.3695. The proposed

method performs better than individual performance of the (SMO) and (K-mean).

Also in terms of detection rate or Detection ratio (means correctness in a model for detecting intrusion), experimental result shows that the proposed algorithm performs better in term of correctness in detecting intrusion (94.48) while other individual data mining techniques (SMO= 61.39) and (K-mean =53.26).

As shown in the comparison, false positive rate proposed model perform better (1.2) as compared to other models individual data mining techniques (SMO= 9.7 and K-mean =2.8). This parameter is very important measure to evaluate the performance of a model. Hence results show that the proposed model performs better than other models.

From all the above experimentation results, it is shown that after applying all the evaluation parameters, proposed model found to be the best model in all scenarios. By applying the hybrid approach of data mining models on the dataset, the detection rate is improved for anomaly detection. So the main objective to improve the detection rate in anomaly detection has been met, also false positive rate alarm was reduced and the accuracy is highest one.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 CONCLUSION

In recent years, by spread of using the Internet, need of information security has been felt more than ever to prevent personal and confidential information from unauthorized intrusion. The different approaches introduce for intrusion detection.

In this research presents a hybrid approach to anomaly detection using of K-means clustering and Sequential Minimal Optimization (SMO) classification.

The solution specifically addresses issues that arise in the context of large scale datasets. It uses feature selection in preprocessing phase to reduce the number of dataset, The ConsistencySebsetEvel and Genetic search algorithms have been applied to select specific features from the NLS-KDD dataset and remove those features which are irrelevant before clustering and classification phases, and also it used k-means clustering to reduce the size of the training dataset while maintaining. After that in classification phase supervised algorithm called Sequential Minimal Optimization (SMO) selected to improve the quality of detection.

A comparison between the proposed approach (K-mean + SMO) and individual algorithm K-mean clustering and Sequential Minimal

Optimization (SMO) classification was done, and the results show that our approach outperforms other by a positive detection rate (94.48%) and reduce the false alarm rate (1.2%) and high accuracy (97.3695%). A proposed approach which will be considered in future work, online processing time is expected to be minimized. The reason behind this is that a suitable number of detectors will be generated with high detection accuracy and low false positive rate. As a result, a positive effect on online processing time is expected.

5.2 .Future work:

1. In future, the proposed approach will be evaluated on other standard training datasets to ensure its high performance.
2. Some other feature selection algorithm can be used that can select the more significant feature and make system more effective.
3. the proposed method have classifieds dataset in two classes, so future research can have classified dataset to five classes (four for type of intrusion (Dos, U2R, PROBE, R2L) and one for normal).

REFERENCES

- [1] Leu, F.Y., Tsai, K.L., Hsiao, Y.T. and Yang, C.T., 2015. An Internal Intrusion Detection and Protection System by using Data Mining and Forensic Techniques.
- [2] Münz, G., Li, S. and Carle, G., 2007, September. Traffic anomaly detection using k-means clustering. In GI/ITG Workshop MMBnet.
- [3] Bansal, P., 2015. *A Hybrid Approach to improve the Anomaly Detection Rate Using Data Mining Techniques* (Doctoral dissertation, THAPAR UNIVERSITY PATIALA).
- [4] García-Laencina, P.J., Sancho-Gómez, J.L. and Figueiras-Vidal, A.R., 2010. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2), pp.263-282.
- [5] Fossi, M., Egan, G., Haley, K., Johnson, E., Mack, T., Adams, T., Blackbird, J., Low, M.K., Mazurek, D., McKinney, D. and Wood, P., 2011. Symantec internet security threat report trends for 2010. *Volume*, 16, p.20.
- [6] Parneet Kaur, V P Singh, “Applying k-means SVM on live traffic to differentiate the anomalous data”, communicated to International Journal of Computer Science and Software Engineering, July 2012.
- [7] Mishra, B.K., Rath, A., Nayak, N.R. and Swain, S., 2012, August. Far efficient k-means clustering algorithm. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (pp. 106-110). ACM.
- [8] Wang, Y., Attebury, G. and Ramamurthy, B., 2006. A survey of security issues in wireless sensor networks.
- [9] Park, Z.W. and Kim, M.K., 2005. The Extended TCP for Preventing from SYN Flood Do Attacks. *Journal of KIISE: Computer Systems and Theory*, 32(10), pp.491-498.
- [10] William, Stallings, *Cryptography and Network Security*, 4/E. Pearson Education India, 2006.

- [11] Stallings, William. Network Security Essentials: Applications and Standards (ForVTU). Pearson Education India, 1982.
- [12] Wang, H., Zhang, D. and Shin, K.G., 2002, June. Detecting SYN flooding attacks. In INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE (Vol. 3, pp. 1530-1539). IEEE.
- [13] Ansari, S., Rajeev, S.G. and Chandrashekar, H.S., 2002. Packet sniffing: a brief introduction. IEEE potentials, 21(5), pp.17-19.
- [14]Tso, M.M.H. and Bakshi, B.S., Intel Corporation, 2000. System for virus-checking network data during download to a client device. U.S. Patent 6,088,803.
- [15] Wang, Y.M., Roussev, R., Verbowski, C., Johnson, A., Wu, M.W., Huang, Y. and Kuo, S.Y., 2004, November. Gatekeeper: Monitoring Auto-Start Extensibility Points (ASEPs) for Spyware Management. In LISA (Vol. 4, pp. 33-46).
- [16] Nassar, S., El-Sayed, A. and Aiad, N., 2010, May. Improve the network performance by using parallel firewalls. In Networked Computing (INC), 2010 6th International Conference on (pp. 1-5). IEEE.
- [17] Zalenski, R., 2002. Firewall technologies. IEEE potentials, 21(1), pp.24-29.
- [18] Acharya, S., Wang, J., Ge, Z., Znati, T.F. and Greenberg, A., 2006, June. Traffic-aware firewall optimization strategies. In 2006 IEEE International Conference on Communications (Vol. 5, pp. 2225-2230). IEEE.
- [19] Wang, Y., Attebury, G. and Ramamurthy, B., 2006. A survey of security issues in wireless sensor networks.
- [20] Park, Z.W. and Kim, M.K., 2005. The Extended TCP for Preventing from SYN Flood Do Attacks. Journal of KIISE: Computer Systems and Theory, 32(10), pp.491-498.

- [21] Osareh, A. and Shadgar, B., 2008. Intrusion detection in computer networks based on machine learning algorithms. *International Journal of Computer Science and Network Security*, 8(11), pp.15-23.
- [22] William, Stallings, *Cryptography and Network Security*, 4/E. Pearson Education India, 2006.
- [23] Modi, M.U. and Jain, A. 2016, A survey of IDS classification using KDD CUP 99 dataset in WEKA.
- [24] Sundaram, A., 1996. An introduction to intrusion detection. *Crossroads*, 2(4), pp.3-7.
- [25] Fogla, P. and Lee, W., 2006, October. Evading network anomaly detection systems: formal reasoning and practical techniques. In *Proceedings of the 13th ACM conference on Computer and communications security* (pp. 59-68). ACM.
- [26] Abhaya, K.K., Jha, R. and Afroz, S., 2014. Data Mining Techniques for Intrusion Detection: A Review. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(6).
- [27] Wang, X.B., Yang, G.Y., Li, Y.C. and Liu, D., 2008, September. Review on the application of artificial intelligence in antivirus detection system i. In *2008 IEEE Conference on Cybernetics and Intelligent Systems* (pp. 506-509). IEEE.
- [28] Bhatt, C.A. and Kankanhalli, M.S., 2011. Multimedia data mining: state of the art and challenges. *Multimedia Tools and Applications*, 51(1), pp.35-76.
- [29] Lu, C.T., Boedihardjo, A.P. and Manalwar, P., 2005, August. Exploiting efficient data mining techniques to enhance intrusion detection systems. In *IRI-2005 IEEE International Conference on Information Reuse and Integration, Conf, 2005.* (pp. 512-517). IEEE.
- [30] Jianliang, M., Haikun, S. and Ling, B., 2009, May. The application on intrusion detection based on k-means cluster algorithm. In *Information Technology and Applications, 2009. IFITA'09. International Forum on* (Vol. 1, pp. 150-152). IEEE.

- [31] Wang, J., Wu, X. and Zhang, C., 2005. Support vector machines based on K-means clustering for real-time business intelligence systems. *International Journal of Business Intelligence and Data Mining*, 1(1), pp.54-64.
- [32] Agarwal, S., Yadav, S. and Singh, K., 2012. K-means versus k-means++ clustering technique. In *2012 Students Conference on Engineering and Systems*.
- [30] Reddy, P.C. and Reddy, R.S.S., K-Means Algorithm with Different Measurements in Clustering Approach.
- [34] Platt, J.C., 1999. 12 fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, pp.185-208.
- [35] Taylor, C. and Alves-Foss, J., 2001, September. Nate: Network analysis of a nomalous t raffic e vents, a low-cost approach. In *Proceedings of the 2001 workshop on New security paradigms* (pp. 89-96). ACM.
- [36] Lindqvist, U. and Porras, P.A., 1999. Detecting computer and network misuse through the production-based expert system toolset (P-BEST). In *Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium on* (pp. 146-161). IEEE.
- [37] Minegishi, T. and Niimi, A., 2011, February. Detection of fraud use of credit card by extended VFDT. In *Internet Security (WorldCIS), 2011 World Congress on* (pp. 152-159). IEEE.
- [38] Minegishi, T., Ise, M., Niimi, A. and Konishi, O., 2009. Extension of Decision Tree Algorithm for Stream Data Mining Using Real Data. *Journal of IEEE SMC Hiroshima Chapter*, pp.208-212.
- [39] Barbará, D., Couto, J., Jajodia, S. and Wu, N., 2001. ADAM: a testbed for exploring the use of data mining in intrusion detection. *ACM Sigmod Record*, 30(4), pp.15-24.
- [40] Zhang, J. and Zulkernine, M., 2006, April. A hybrid network intrusion detection technique using random forests. In *First International Conference on Availability, Reliability and Security (ARES'06)* (pp. 8-pp). IEEE.

- [41] Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2), pp.139-157.
- [42] Muniyandi, A.P., Rajeswari, R. and Rajaram, R., 2012. Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm. *Procedia Engineering*, 30, pp.174-182.
- [43] Panda, M. and Patra, M.R., 2007. Network intrusion detection using naive bayes. *International journal of computer science and network security*, 7(12), pp.258-263.
- [44] S. Ahmed, "Intrusion Detection System using Data Mining", Applied Science Department, Master Thesis, University of Technology, 2006.
- [45] Mohamad Tahir, H., Hasan, W., Md Said, A., Zakaria, N.H., Katuk, N., Kabir, N.F., Omar, M.H., Ghazali, O. and Yahya, N.I., 2015. Hybrid machine learning technique for intrusion detection system. 5th International Conference on Computing and Informatics (ICOCI) 2015.
- [46] Modi, M.U. and Jain, A. 2016, A survey of IDS classification using KDD CUP 99 dataset in WEKA.
- [47] Murthy, C., Manjunatha, A.S., Jaiswal, A. and Madhu, B.R., 2016. Building Efficient Classifiers For Intrusion Detection With Reduction of Features. *International Journal of Applied Engineering Research*, 11(6), pp.4590-4596.
- [48] Dhanabal, L. and Shantharajah, D.S., 2015. A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, 4(6).
- [49] Negnevitsky, M., 2005. Artificial intelligence: a guide to intelligent systems. Pearson Education.