# Analysis of Data Mining Methodology and Techniques for Intrusion Detection

Muhammad Tayyab

## ABSTRACT

*An intrusion is an intentional, unauthorized access or manipulates information from the internal or external network. Intrusion detection system monitors all inbound and outbound network traffic and classifies suspicious patterns to a specified system or network attack. Intrusion detection system detects only the malicious attack but does not prevent it from happening. Data mining is the act of providing meaningful and useful information from large bulk of data. In this paper we are comparing some popular techniques and methodologies of data mining used for the detection of such malicious activities.*

**Key Words:** Data Mining, Intrusion, System Attacks and Computer Security.

## INTRODUCTION

Any kind of attack on computer system is a severe problem. Threats to confidentiality, integrity and availability and its protection are known as computer security (Summers, 1997) . Capturing and identifying serious threats and attacks in network such as Distributed Denial of Service (DDoS) or worm distribution is the major challenge (Chen, Gao & Kwiat, 2003); (Douligeris & Mitrokotsa, 2004) .The challenges and issues of the future aid towards the construction of reliable intrusion detection system. Over the years, designers and researchers proposed many intrusion detection systems by using different techniques but there are many issues which must be resolved in order to design a system which should be optimal and finest.

There are some methods that can be used to detect massive kind of viruses and worms such as Firewalls and Routers but it is possible when the virus or worm is already defined in signatures and it can also be detected when the virus is already spread (Rasheed, Ghazali, Norwawi & Kadhum, 2011) .Without using Data mining techniques like clustering, association rules or classification, the detection and finding of intrusions are very hard because we would ask: How much data would we get? How would we display data? What kind of data did we want to see, and what queries are best to highlight that data? In the absence of these properties we suspect that our system is inadequate for detecting the most serious attack (Bloedorn et al., 2001) . Data mining (the analysis step of the Knowledge discovery in Databases process or KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), is the process that discovers patterns in large data sets. This process or step is used to extract useful knowledge from data sets. Further this knowledge is transform in to human understandable layout (Chakrabarti et al., 2004). Data mining techniques can be differentiated by their different model functions and representation, preference criterion, and algorithms (Fayyad, et al., 1996).

The objective of the model is to classify attack, as normal, or malicious, or as a particular type of attack (Ghosh, Schwartzbard & Schatz, 1999) . Data mining systems provide the means to easily perform data summarization and visualization, aiding the security analyst in identifying areas of concern (Lee & Stolfo, 2000).

This paper makes comparison of the four methods on the basis of state of art. We have selected these four methods because of their efficiency and speed for intrusion detection. We first describe four existing methods with analysis and experiments/illustrations and then we move towards their comparison.

## EXISTING METHODS

In this section we explain that how was data collected or generated and how was it analyzed in our selected methods.

Mining data in real time is a big challenge (Naveen, Srinivasan & Natarajan, 2011). For capturing packets in real time the same author carried out experiment on a real data stream called "intrusion data set" which is collected from the server in real time using JPCAP and WINCAP tools. According to authors, JPCAP provides facility to capture and save unprocessed packets to an offline file, identifies packet types, filters packets according to user-defined rules and sends these packets to the network as shown in the figure 1 (Naveen, et al., 2011). This offline file stores data that contains the details of the network connections, such as type of protocol, source IP, destination IP, source port and destination port and number of bytes in the source etc. The same authors used WEKA tool as shown in the figure 3 to analyze audit data, examine it as a normal, abnormal or aggressive behavior and report to the manager in a comprehensive form.

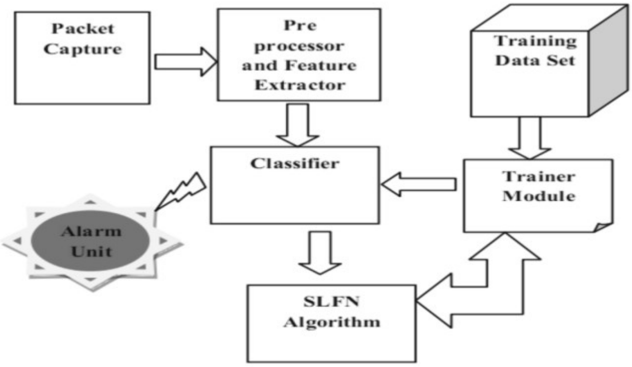
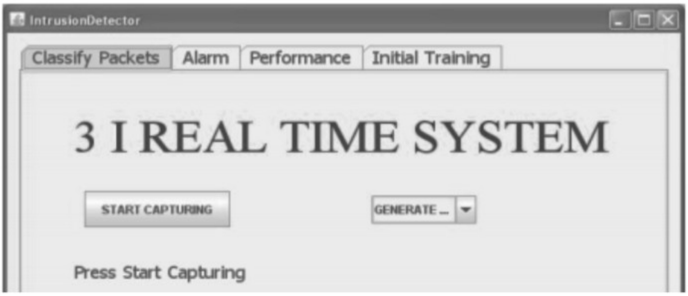


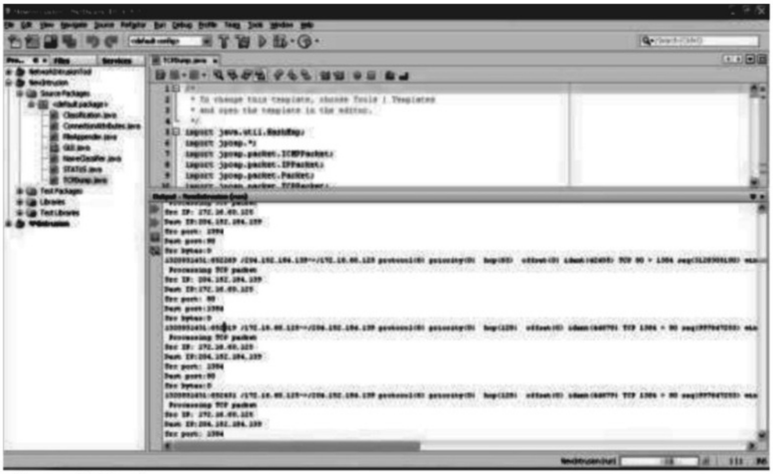
Figure 1. Real Time System Architecture



3 I REAL TIME SYSTEM

Press Start Capturing

Figure 2. Initial Screen



```
/*
 * To change this template, choose Tools | Templates
 * and open the template in the editor.
 */
import java.util.HashMap;
import jpcap.*;
import jpcap.packet.ICMPPacket;
import jpcap.packet.IPPacket;
import jpcap.packet.Packet;
import jpcap.packet.TCPPacket;
```

```
 Processing TCP packet
Src IP: 172.16.65.125
Dest IP:204.152.184.139
Src port: 1394
Dest port:80
Src bytes:0
1320951451:652269 /204.152.184.139->/172.16.65.125 protocol(6) priority(0)  hop(50)  offset(0) ident(42495) TCP 80 > 1394 seq(3128900180) [illegible]
 Processing TCP packet
Src IP: 204.152.184.139
Dest IP:172.16.65.125
Src port: 80
Dest port:1394
Src bytes:0
1320951451:652319 /172.16.65.125->/204.152.184.139 protocol(6) priority(0)  hop(128)  offset(0) ident(64079) TCP 1394 > 80 seq(897947203) [illegible]
 Processing TCP packet
Src IP: 172.16.65.125
Dest IP:204.152.184.139
Src port: 1394
Dest port:80
Src bytes:0
1320951451:652451 /172.16.65.125->/204.152.184.139 protocol(6) priority(0)  hop(128)  offset(0) ident(64079) TCP 1394 > 80 seq(897947203) [illegible]
 Processing TCP packet
Src IP: 172.16.65.125
Dest IP:204.152.184.139
Src port: 1394
```

Figure 3. Packet capture screen

Detecting, learning rate and false positive alert of intrusions show that how well the system is designed and meant to be performed well. Therefore the same authors (Naveen, et al., 2011) have proposed SLFN Algorithm to improve performance of the learning rate and new attacks found are updated to the training set accordingly for future processing. Main attacks which are detected by the model are Probing and Denial of service (DOS). These techniques detect attacks faster as compared to other methods.

A bi-Clustering technique is proposed by (Lappas & Pelechrinis, 2007) as a tool to examine the network packets and enhance the services of IDS. Bi-Clustering performs two operations; first it finds the partition of the vectors in each cluster and other function calculates the subset of dimensions. There is a need of clustering of vectors as well as dimensions at the same time. The easiest way suggested by the authors is to represent data in matrix. This data includes objects and features. Each row of the table represents an object (e.g. packets generated by the user) and each column specifies a feature (e.g. the source port). The relationship between processes and features offers the valuable information. This system only detects normal worms or intrusions but not all fundamental intrusions.

Table 1. Bi-Clustering Technique

| | Feature A | Feature B | Feature C | Feature D | Feature E |
|---|---|---|---|---|---|
| Process 1 | A1 | B2 | C3 | D2 | E2 |
| Process 2 | A3 | B3 | C2 | D1 | E1 |
| Process 3 | A1 | B2 | C3 | D2 | E2 |
| Process 4 | A1 | B2 | C3 | D3 | E2 |
| Process 5 | A1 | B2 | C1 | D2 | E2 |
| Process 6 | A1 | B2 | C2 | D2 | E2 |
| Process 7 | A3 | B1 | C1 | D1 | E1 |

In Table 1 each row indicates process and each column represents a feature (e.g. the source port). If we find the malicious activity or abnormality in the processes, these processes can give us the set of malicious traces and provides the facility to cluster them. The obtained biclusters could be an effective way to summarize and separate similar processes and analyzed them as a group. These biclusters can provide valuable knowledge on the relationship between processes and features. This method is better and more effective intrusion detection system.

Another method applied by (Bridges & Vaughn, 2000) who integrated data mining techniques with fuzzy logic to provide new techniques for intrusion detection. The same author applied this method on two levels i.e. workstation level and network level. According to the authors, the association technique was used in order to detect both misuse and normal intrusions. The system has the tendency to support both fuzzy and non fuzzy rules. The same authors also utilized genetic algorithm to adjust the association functions for fuzzy variables and select the most effective set of features for particular types of intrusions.

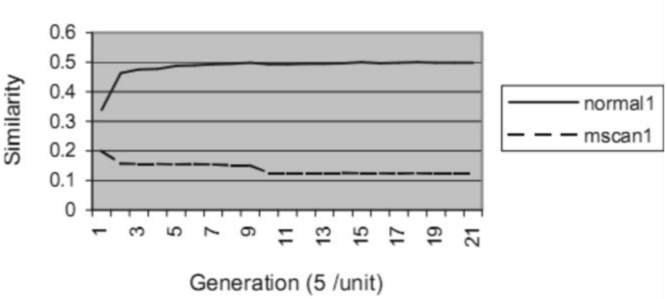


Figure 4. Training Process in Terms of Similarity

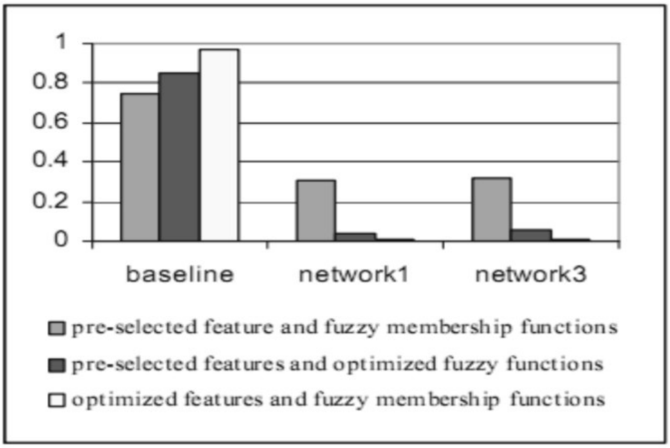


Figure 5. Comparison of Similarity Results

False Positive Alert Reduction Technique for reducing the false positive alarm rate is discussed in (Kumar, Hanumanthappa & Kumar, 2011). According to author, network intrusion detection system (IDS) is very susceptible for identifying and detecting network attacks but whenever intrusion detection system predict any intrusion, it activates alarm for security alert but there are thousands of activities running across the networks; Intrusion Detection System (IDS) generates thousands of alarms on scents of suspiciousness of any particular activity, it is fact that most of them are false (Kumar, et al., 2011). To avoid these same authors has discussed problems and proposed alert reduction technique. In this technique it has been recommended to update the desired patches and updates the security signatures. Until and unless all possible threats and signatures or behaviors are incorporated in intrusion detection system the elimination of false positives cannot be achieved so therefore other methods are essential to address false positives or false negatives.

## COMPARISON OF THE EXISTING METHODS

All existing methods analyze traffic, examine type of attack, and test data according to the data mining technique used. But the system proposed by (Naveen, et al., 2011) is used for detecting malicious activity or normal activity in real time on host systems. The authors used classification data mining technique. Main attacks which are detected by the model are Probing and Denial of service (DOS). These techniques detect attacks faster compared to other methods. Other method which presented by (Lappas & Pelechrinis, 2007) uses bi-clustering technique and proposed for host based systems. Bi-Clustering can provide valuable knowledge on the relationships between processes and features. This method is better but not more useful for detection of abnormal intrusions.

Association and genetic algorithm based technique proposed by (Bridges & Vaughn, 2000) for workstation as well as the network level intrusions detections. This method is very good for large networks but not optimum for small networks. Finally (Kumar, et al., 2011) discussed false positive alert reduction technique for reducing the false positive alarm rate. In this technique all updates and desired patches are recommended to be updated. As discussed all these methods detect intrusions but the type of intrusions and service used on the type of system make it

**Summary of the comparison:**

In this section we summarize above discussed methods in tabular form in table 2 that illustrates the type of data mining technique and system used, threats/attacks categorization and whether method is implemented or not. The table explains the exact system working with the technique used by the author and for the system the desired technique were designed.

Table 2. Summary of Data Mining Techniques Discussed

| Author | DM Technique | System/Network type | Type of attack | Implemented |
|---|---|---|---|---|
| Naveen, N., Srinivasan, D. R., & Natarajan, D. S | Classification | Real time host based | Denial of Service, Probing, Normal | Yes |
| Lappas, T., & Pelechrinis, K | Bi-Clustering | Host based | Normal | Yes |
| Bridges, S. M., & Vaughn, R. B | Association, Genetic | Workstation, network level | Normal, misuse | Yes |
| Kumar, M., Hanumanthappa, M., & Kumar, T. V. S | Association/ False positive alert reduction | Host based | Denial of Service | Yes |

## CONCLUSION AND FUTURE WORK

With the advent of new technologies building intrusion detection system has become more complex. Different techniques and methodologies are proposed by different researchers and some of them are very reliable but without data mining techniques Intrusion or Misuse detection is quite hard. Different methodologies were tried to compare and discus in this paper. Data mining is a hot field of now a days and also for future so a technique could be used to combine multiple data mining techniques to detect suspicious attacks.

## REFERENCES

Bloedorn, E., Christiansen, A. D., Hill, W., Skorupka, C., Talbot, L. M., & Tivel, J. (2001). Data mining for network intrusion detection: How to get started: MITRE Technical Report.

Bridges, S. M., & Vaughn, R. B. (2000). Intrusion detection via fuzzy data mining. Paper presented at the 12th Annual Canadian Information Technology Security Symposium.

Chakrabarti, S. Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., et al. (2004). Data mining curriculum: A proposal (Version 0.91).

Chen, Z., Gao, L., & Kwiat, K. (2003). Modeling the spread of active worms. Paper presented at the INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies.

Douligeris, C., & Mitrokotsa, A. (2004). DDoS attacks and defence mechanism:classification and state-of-the-art. The international Journal of Computer and Telicommunications Networking, 44(5), 643 – 666.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

Ghosh, A. K., Schwartzbard, A., & Schatz, M. (1999). Learning program behavior profiles for intrusion detection. Paper presented at the 1st USENIX Workshop on Intrusion Detection and Network Monitoring.

Kumar, M., Hanumanthappa, M., & Kumar, T. V. S. (2011). Intrusion Detection System-False Positive Alert Reduction Technique. Paper presented at the Proc. of Int. Conf. on Advances in Computer Engineering.

Lappas, T., & Pelechrinis, K. (2007). Data mining techniques for (network) intrusion detection systems. Department of Computer Science and Engineering UC Riverside, Riverside CA, 92521.

Lee, W., & Stolfo, S. J. (2000). Data mining approaches for intrusion detection: Defense Technical Information Center.

Naveen, N., Srinivasan, D. R., & Natarajan, D. S. (2011). A Unified approach for real time intrusion detection using intelligence data Mining techniques. Paper presented at the IJCA Special Issue on Network Security and Cryptography.

Rasheed, M. M., Ghazali, O., Norwawi, N. M., & Kadhum, M. M. (2011). A traffic signature-based algorithm for detecting scanning internet worms. International Journal of Communication Networks and Information Security (IJCNIS), 1(3).

Summers, R. C. (1997). Secure computing: threats and safeguards. McGraw-Hill, Inc.

**Muhammad Tayyab:** Working as Network Administrator at City University of Science & Information Technology (CUSIT), Peshawar, Pakistan. MS(CS) Scholar. Areas of interest are ad-hoc networks & operating systems.
e-mail: hafiztayyab@yahoo.com