MADAM ID FOR INTRUSION DETECTION USING DATA MINING

Prabhjeet Kaur*

Amit Kumar Sharma*

Sudesh Kumar Prajapat*

ABSTRACT

Data Mining for IDS is the technique which can be used mainly to identify unknown attacks and to reduce false alarm rates in anomaly detection technique. Various Research Projects using Data Mining techniques for Intrusion Detection are proposed one of which is MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection) used to detect both Misuse detection (used to identify known attacks) and Anomaly detection (used to predict unknown behavior of attacks). It uses data mining technique on different data sets captured by continuous auditing of data on network.

This paper focus on MADAM ID which includes types of intrusion it detect like DDOS attack, various types of alarm ratesit generated, C4.5 algorithm which is used to classify the data as normal and abnormal and how it is better than ID3 algorithm, types of result it generated with example, total cost it includes, drawback of MADAM ID, future scope of data mining in intrusion detection.

We use Wireshark tool for auditing packets on network and WEKA tool for pre-processing on the given data set, classify them by J48 tree which is an implementation of C4.5 algorithm and detect various alarm rates.

***Keywords**: MADAM ID, C4.5 algorithm, J48, DOS attack, cost estimation*

*Department of Computer Science, Central University of Rajasthan, Kishangarh, Ajmer, Rajasthan, India.

1. INTRODUCTION

In current scenario network security is one the main problem faced by all users. Various threats or intrusions are emerging day to day that can harm the security of the system and confidentiality of information. Intrusion detection system (IDS) is the way to detect the threats and attacks that can violate the security of the system and to preserve data integrity and system availability from various attacks [1]. Today, data mining techniques are used to provide better solution for good IDS.It requires efficient and accurate analysis of large amount of system and network audit data.MADAM IDused data mining technique which is used both for misuse and anomaly detection. In this data mining techniques is being used to evaluate, examine and classifying large amount of network data for Intrusion Detection [2, 3].Many different data mining techniques such as clustering, classification and association rules are providing us several algorithms for Intrusion detection [1]. In this paper we discuss C4.5 algorithm which are depending on classification rules of data mining and analyse data implementing this algorithm by Wekaas an analysing tool. We use decision tree approach for this.This analysis helps in detecting various types of attacks like DDOS attack and helps us to prevent our system from these attacks.

2. OVERVIEW OF CLASSIFICATION TECHNIQUE

Classification helps to put instances into predefined classes. Classification is a data mining technique based on machine learning which is used to predict group membership for data instances. For example, we may wish to use classification to predict whether the patient in a hospital is “sick” or “healthy”

This can be done by use of various types of classifiers [4]. This type of classification task is commonly referred to as supervised learning as class label for each training tuple is given. In this there is a specified set of classes, and example objects are labeled with the appropriate class. The main goal is to generalize from the training objects that will enable novel objects to be identified as belonging to one of the classes. The success of classification learning is heavily dependent on the quality of the data provided for training. If data is not adequate then it leads to misclassification which results in poor results.

Classification consists of two steps:

- Training: It is the supervised learning of a training set of data to build a model.
- Testing: It includes classifying the data according to that model.

3. ANALYIS OF DATA SET

We analyze the data set given in Fig 1 taken from networktraffic by Wireshark tool andthen analyze it by Weka tool.

No.	Time	Source	Destination	Protocol	Length	Info
1	0	62.41.58.10	192.168.1.11	HTTP	1474	Continuation or non-HTTP traffic
2	0.027795	62.41.58.10	192.168.1.11	HTTP	1474	Continuation or non-HTTP traffic
3	0.027928	192.168.1.11	62.41.58.10	TCP	54	49755 > http Ack=2841
4	0.071856	62.41.58.10	192.168.1.11	HTTP	1474	Continuation or non-HTTP traffic
5	0.09516	62.41.58.10	192.168.1.11	HTTP	1474	Continuation or non-HTTP traffic
6	0.095296	192.168.1.11	62.41.58.10	TCP	54	49756 > http Ack=2841
7	0.101742	62.41.58.10	192.168.1.11	HTTP	1474	Continuation or non-HTTP traffic
8	0.121737	192.168.1.11	62.41.58.10	TCP	54	49750 > http Ack=1
9	0.131743	192.168.1.11	62.41.58.10	TCP	54	49753 > http Ack=1
10	0.153958	62.41.58.10	192.168.1.11	HTTP	1474	Continuation or non-HTTP traffic
11	0.174762	62.41.58.10	192.168.1.11	HTTP	1474	Continuation or non-HTTP traffic
12	0.174912	192.168.1.11	62.41.58.10	TCP	54	49758 > http Ack=2841
13	0.189314	62.41.58.10	192.168.1.11	HTTP	1474	Continuation or non-HTTP traffic
14	0.189464	192.168.1.11	62.41.58.10	TCP	54	49756 > http Ack=5681
15	0.202527	62.41.58.10	192.168.1.11	HTTP	1474	Continuation or non-HTTP traffic
16	0.224482	62.41.58.10	192.168.1.11	HTTP	1474	Continuation or non-HTTP traffic

Fig 1: Showing dataset named as id1

a. PREPROCESSING

It helps to import the data from various sources in the form of csv and other types of file. Various preprocessing filters used are discretization, attribute selection, transforming and combining attributes, etc.

Fig 2 shows the preprocessing on info attribute which are of 7 different types.

Out of which 6 of them are unique and label name “Continuation or non-HTTP traffic” are 10 in number.

Selected attribute

Name: Info Type: Nominal
Missing: 0 (0%) Distinct: 7 Unique: 6 (38%)

No.	Label	Count	Weight
1	Continuation or non-HTTP traffic	10	10.0
2	49755 > http Ack=2841	1	1.0
3	49756 > http Ack=2841	1	1.0
4	49750 > http Ack=1	1	1.0
5	49753 > http Ack=1	1	1.0
6	49758 > http Ack=2841	1	1.0
7	49756 > http Ack=5681	1	1.0

Fig 2

Fig 3 shows the graph generated by preprocessing on “Info” attribute as shown in Fig 2. Here blue line indicates the weight of info attribute named as”Continuation or non-HTTP traffic”. Similarly,weight of other info attribute is shown.

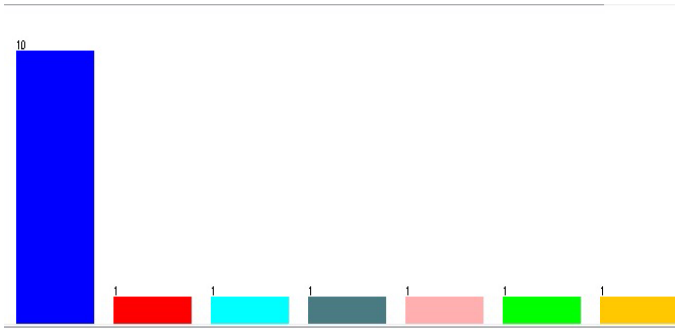

Fig 3

b. CLASSIFICATION

In this we apply the decision tree classifier that is J48 which

is the implementation of C4.5 algorithm.It calculate the number of instances and attributes of given data set and after that it evaluate the given training data set by using use “training set option”.

```
Relation:     id1
Instances:    16
Attributes:   7
              No.
              Time
              Source
              Destination
              Protocol
              Length
              Info
Test mode:    evaluate on training data
```

Fig 4

c. C4.5 ALGORITHTM

This algorithm can be used to generate decision tree that can be used to classify data instances in different classes which helps in further analysis detecting valid results [5].

This algorithm made number of improvements on ID3 algorithm. These are:

- It can handle both continuous and discrete attributes.For continuous attributes, it creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Missing attribute values are simply not used in gain and entropy calculations.

- It can handle attributes with different costs.

Fig 5 shows the decision tree of data set given in Fig 1.

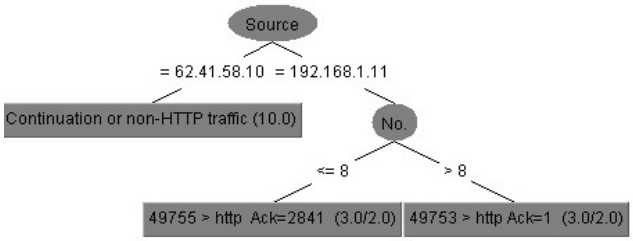

Fig 5

By J48 classifier the data set given in Fig 1 can be divided into various classes and its accuracy can be calculated by various parameters. These are

- True positive (TP): This alarm corresponds to the numberof detected attacks and it is in fact attack[1].
- True negative (TN): This alarm corresponds to the numberof detected normal instances and it is actually normal [1].
- False positive (FP):This alarm corresponds to the number of detected attacks but it is in factnormal [1].
- False negative (FN): This alarm corresponds to the numberof detected normal instances but it is actuallyattack, in other words these attacks are the targetof intrusion detection systems [1].
- Precision (A) =number of correctly classified instances of class A / number of instances classified as belonging to class A [6].
- Recall (A) = number of correctly classified instances of class A/ number of instances in class A [6].
- The true positive rate (TPR) TPR = TP / (TP + FN)
- The true negative rate (TNR) = TN / (TN + FP)
- False positive rate(FPR)= FP / (TN + FP)
- The false negative rate (FNR)= FN / (TP + FN)

Table 1 gives the detailed accuracy by class of data set given in Fig 1.

TP Rate	FP Rate	Precision	Recall	Class
1	0	1	1	Continuation or non-HTTP traffic
1	0.133	0.333	1	49755 > http Ack=2841
0	0	0	0	49756 > http Ack=2841
0	0	0	0	49750 > http Ack=1
1	0.133	0 .333	1	49753 > http Ack=1
0	0	0	0	49758> http Ack=2841
0	0	0	0	49756 > http Ack=5681
Weighted Avg.	0.75	0.017	0.667	0.75

Table 1

4. MADAM ID FEATURES AND LIMITATION

A. DETECTION OF DDOS ATTACK

The mission of this attack to send unusual traffic towards the victim to make it busy [7, 8].In Fig 1 we say that the destination host with 192.168.1.11 IP address have http service request in the past few seconds. So, there is the possibility of denial of service attack with that service.

According to this type of analysis we classify the data set as normal or abnormal and detect various types of intrusions.

B. COST EVALUATION

It can be evaluated by various cost factors. These are

- Damage cost: The maximum amount of damage occurs when intrusion detection system is not available.

- Response cost: Is the cost to take an action when intrusion is found.
- Consequential cost: Cost due to connection.

For MADAM ID various cost levels features are

- At the beginning of an event
- At the middle to end of an event
- At multiple events in a time window[9]

C. LIMITATION OF MADAM ID

- It is applied only at connection level.
- Within connection classification of contents are very challenging [9].

II. DRAWBACK OF INTRUSION DETECTION SYSTEM BASED ON DATA MINING.

- It is very consuming due to massive efforts in analyzing the data and categorizes them according to useful and raw data that also results in high cost and manpower.
- The behavior of the attack changes frequently and thus the data used for auditing also which requires continuous monitoring of data and filtering efforts to identify them as normal suspicious [10].

5. CONCLUSION AND FUTURE SCOPE

This paper described MADAM ID which used classification technique of data mining that improve theperformance of Intrusion Detection Systems (IDS).Weka tool helps to get the proper understanding of this technique by analyzing on different aspects of given data set using C4.5 algorithm. Experimental result shows how to identify DDOS attack most common type of intrusion violating security.Future work will include payingmore attention to provide better methods ofproper auditing the data which is most difficult task and classify them efficiently. To work on projects which are basis of detection of unknown behavior of attack which greatly affect the security of various systems.

REFERENCES

[1] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi and Lilly Suriani Affendey,"Intrusion Detection Using Data Mining Techniques",978-1-4244-5651-2/10/$26.00 ,2010 IEEE
C. Xiang, M. Y. Chong and H. L. Zhu" Design of Multiple-Level Tree Classifiersfor Intrusion Detection System", Proceedings of the 2004 IEEEConference on Cybernetics and Intelligent SystemsSingapore, 1-3 December, 2004

[2] C. Xiang and S. M. Lim,"Design Of Multiple-Level Hybrid Classifier ForIntrusion Detection System",0-7803-9518-2/05/$20.00,2005 IEEE

[3] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, 2ndEd.

[4] J.R.Quinlan. C4.5: Programs for machine learning. Morgan Kaufman Publishers, 1993.

[5] P Srinivasulu1, D Nagaraju, P Ramesh Kumar, and K Nageswara Rao," Classifying the Network Intrusion Attacksusing Data Mining Classification Methods and their Performance Comparison"IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.6, June 2009

[6] Kanwal Garg and Rshma Chawla,"Detection Of DDOS Attack Using Data Mining",(IJCBR)ISSN (Online) : 2229-6166, Volume 2 Issue 1, 2011

[7] Yoohwan Kim, Wing Cheong Lau, Mooi Choo Chuah And Jonathan H. Chao"Packetscore: Statistical-Based Overload Control Against Distributed Denial-Of-Service Attacks" IEEE INFOCOM 2004 ,The 23rd Annual Joint Conference of the IEEE Computer and Communications Societies, Hong Kong, China, March 7-11, 2004. IEEE, 2004.

[8] AleksandarLazarević, Jaideep Srivastava and VipinKumar, "Data Mining For Intrusion Detection", ArmyHigh Performance Computing Research Center Department of Computer Science University of Minnesota, Tutorial on the Pacific-Asia Conference on Knowledge Discovery in Databases 2003

[9] Anoop Singhal and Sushil Jajodia, "Data warehousing and data mining techniques for intrusion detection systems", Distrib Parallel Databases (2006) 20:149–166, Springer Science+Business Media, LLC 2006.