# The Role of Decision Tree Technique for Automating Intrusion Detection System

**Neha Jain, Shikha Sharma**
Department of Software Engineering, Suresh Gyan Vihar University, Jaipur
Department of Computer Science & Information Technology, ECA, Ajmer

**Abstract**-Security of computers and the networks that connect them is increasingly becoming of great significance. Intrusion detection is a mechanism of providing security to computer networks. Although there are some existing mechanisms for Intrusion detection, there is need to improve the performance. Data mining techniques, such as decision tree analysis, offers a semi-automated approach to detect threats. In this paper, decision tree technique is applied on a small set of network data. Then build a decision tree model, and incorporates the model's logic into snort signatures or firewall rules.

**Keywords:** Denial of Service, Data mining, IDS, Network security, decision tree

## I. INTRODUCTION

One of the main challenges in the security management of large-scale high-speed networks is the detection of suspicious anomalies in network traffic patterns due to Distributed Denial of Service (DDoS) attacks or worm propagation [1][2] A secure network must provide the following:

1. Data confidentiality: Data that are being transferred through the network should be accessible only to those that have been properly authorized.
2. Data integrity: Data should maintain their integrity from the moment they are transmitted to the moment they are actually received. No corruption or data loss is accepted either from random events or malicious activity.

3. Data availability: The network should be resilient to Denial of Service attacks.

An intrusion detection system (IDS) inspects all inbound and outbound network activity and identifies suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system. When a potential attack is detected the IDS logs the information and sends an alert to the console. IDS try to find data packets that contain any known intrusion-related signatures or anomalies related to Internet protocols.

## II IDS TAXONOMY

The goal of IDS is to detect malicious traffic. In order to accomplish this, the IDS monitor all incoming and outgoing traffic. There are several approaches on the implementation of IDS. Among those, two are the most popular [11]:

***A. Anomaly detection*:** This technique is based on the detection of traffic anomalies. The deviation of the monitored traffic from the normal profile is measured. Various different implementations of this technique have been proposed, based on the metrics used for measuring traffic profile deviation.

***B. Misuse/Signature detection:*** This technique looks for patterns and signatures of already known attacks in the network traffic [12]. A constantly updated database is usually used to store the signatures of known attacks. The way this technique deals with intrusion detection resembles the way that anti-virus software operates. [3][4][5][6][7].

## III. DRAWBACKS OF IDS

Intrusion Detection Systems (IDS) have become a standard component in security infrastructures as they allow network administrators to detect policy violations.

Current IDS have a number of significant drawbacks:

1. *Current IDS* are usually tuned to detect known service level network attacks. This leaves them vulnerable to original and novel malicious attacks.

**2. *Data overload:*** Another aspect which does not relate directly to misuse detection but is extremely important is how much data an analyst can efficiently analyze. That amount of data he needs to look at seems to be growing rapidly. Depending on the intrusion detection tools employed by a company and its size there is the possibility for logs to reach millions of records per day.

**3. *False positives:*** A common complaint is the amount of false positives a IDS will generate. A false positive occurs when normal attack is mistakenly classified as malicious and treated accordingly.

**4. *False negatives:*** This is the case where an IDS does not generate an alert when an intrusion is actually taking place. (Classification of malicious traffic as normal)

Data mining can help improve intrusion detection by addressing each and every one of the above mentioned problems.

## IV. DATA MINING: DEFINITION

Data mining is, at its core, pattern finding. Data miners are experts at using specialized software to find regularities (and irregularities) in large data sets. Here are a few specific things that data mining might contribute to an intrusion detection project:

- Remove normal activity from alarm data to allow analysts to focus on real attacks

- Identify false alarm generators and "bad" sensor signatures
- Find anomalous activity that uncovers a real attack
- Identify long, ongoing patterns (different IP address, same activity)

To accomplish these tasks, data miners use one or more of the following techniques [8][9][10]:

- *Data summarization* with statistics, including finding outliers
- *Visualization*: presenting a graphical summary of the data
- *Clustering* of the data into natural categories
- *Association rule discovery*: defining normal activity and enabling the discovery of anomalies
- *Classification*: predicting the category to which a particular record belongs

## V. DECISION TREE: A CLASSIFICATION TECHNIQUE

A decision tree is defined as "a predictive modeling technique from the fields of machine learning and statistics that builds a simple tree-like structure to model the underlying pattern [of data]".

Decision trees are one example of a classification algorithm. Classification is a data mining technique that assigns objects to one of several predefined categories. From an intrusion detection perspective, classification algorithms can characterize network data as malicious, benign, scanning, or any other category of interest using information like source/destination ports, IP addresses, and the number of bytes sent during a connection.

## VI. WHY USE DECISION TREE

Decision trees is a viable tool in the intrusion detection toolkit, the technique needs to satisfy a minimum set of requirements. The technique needs to be beneficial to the intrusion analysis mission and produce real results for an organization.

In addition, decision trees must be unique among existing tools. If other tools exist with the same functionality provided by decision trees, then decision trees may be redundant and unnecessary.

## VII. USING DECISION TREE FOR INTRUSION DETECTION

Two prerequisites for the analysis are data collection (i.e. identifying and collecting data of interest) and tool acquisition and selection (i.e. identifying and deploying data mining tools). The gathered data requires a pre-processing phase to move it into the form necessary for decision tree algorithms. After the data is processed, decision trees can be trained using the processed data and tools. Running and analyzing the result of this data is an important next step to understand the resulting model and its rule sets. The final step is using the results of the analysis to run the decision rules in real-time.

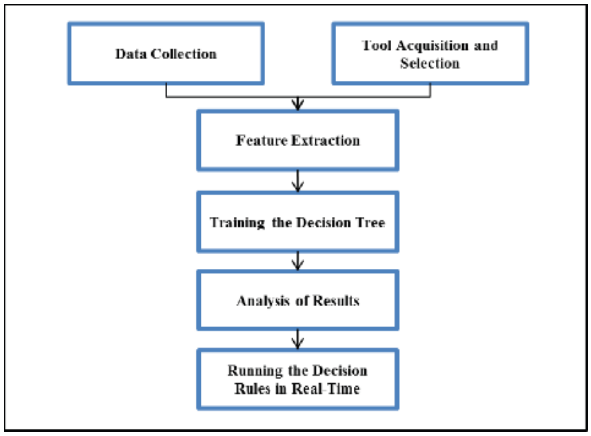


**Fig 1:** Process to implement decision tree for Intrusion Detection

## VIII. IMPLEMENTATION

1. Implementing decision trees can require some network data. Download the following files from http://www.openpacket.org/
   example.com-3.pcap
   example.com-4.pcap
   example.com-5.pcap
   zeus-sample-3.pcap (botnet traffic)
2. Implementing decision trees can require various tools:

- *Feature extraction tools*: used during the data pre-processing phase. For this we use tcptrace tool to perform feature extraction from pcap files.
  Download and install tcptrace from http://www.tcptrace.org/
- *Data mining analysis tool*: Weka is used for this purpose. Of all the open-source tools, Weka has been described as "perhaps the best-known open-source machine learning and data mining environment"
  Download and install Weka from http://www.cs.waikato.ac.nz/ml/weka/

3. Now the Feature Extraction tool is used to collect and structure the features from a dataset in a format that can be used for training the decision tree.

Steps:

i) For each pcap file, run the command:
tcptrace --csv -l filename1.pcap > filename1.csv
(where filename is the name of the pcap file)

ii) From each csv file, remove rows 1-8 (the row before conn #)

iii) From each csv file, delete the following columns EXCEPT

- port_a
- port_b
- total_packets_a2b
- total_packets_b2a
- unique_bytes_sent_a2b
- unique_bytes_sent_b2a

iv) Add new column called "class" to each spreadsheet. Fill in each cell of the new column with either "normal" or "malicious".

v) Copy and paste all cells from the spreadsheets into a single csv file called analysis_of_traffic.csv

4. Run Weka

i) From the Weka GUI Chooser, click on the Explorer button

ii) From the Weka Explorer GUI, click on Open File.

iii) Using the explorer, open the traffic_analysis.csv file

iv) Click on the Classify tab at the top of the Weka Explorer GUI
Click on the "choose" button to select a classifier
From the menu, expand the trees icon
Click on the J48 Tree classifier

v) On the Classify GUI, click on the Start button to start the classifier

5. Weka Output

This decision tree in the output listed above states that connections with port_a <= 1049 and port_b <= 445 are malicious, otherwise the connection is normal.
For reference, tcptrace defines port_a as the port of the machine initiating the connection and port_b as the port of the machine receiving the connection.

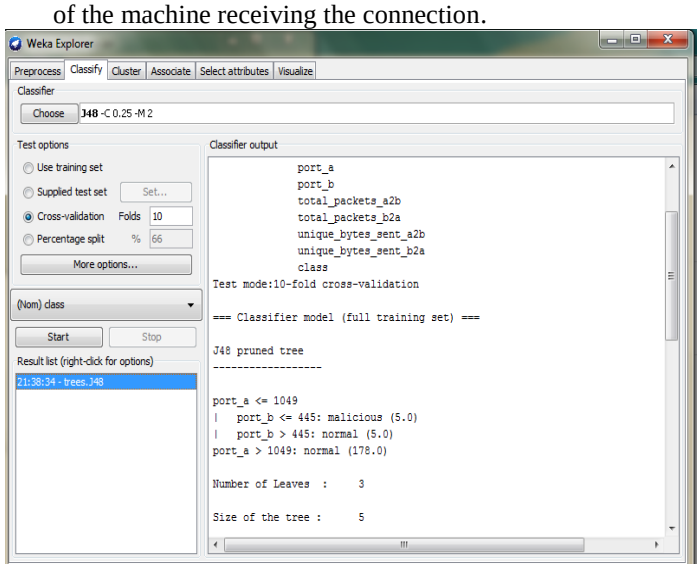


**Fig 2:** Weka output

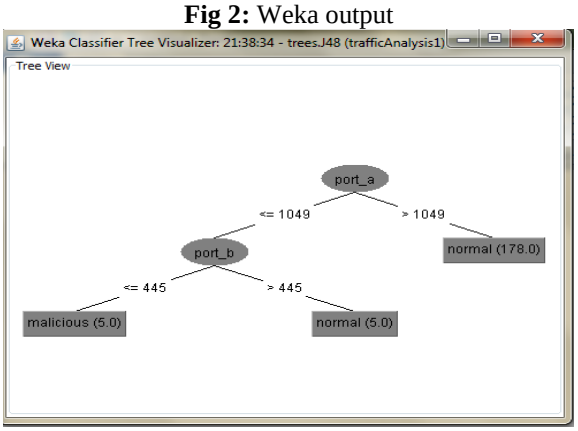


**Fig 3:** Tree view

## IX. CONCLUSIONS

This paper has presented a decision tree technique that categorize new piece of information into a number of predefined categories. Decision tree uses a pre-classified dataset to learn to categorize data based on existing trends and patterns. After the tree is created, the logic from the decision tree can be incorporated into number of intrusion detection technologies including firewalls and IDS signatures.

With the increasing incidents of cyber attacks, building an effective intrusion detection models with good accuracy and real-time performance are essential. Data mining is relatively new approach for intrusion detection. More data mining techniques should be investigated and their efficiency should be evaluated as intrusion detection models.

## REFERENCES

[1] Christos Douligeris, Aikaterini Mitrokotsa, "DDoS attacks and defense mechanisms: classification and state-of-the-art" ,Computer Networks: The International Journal of Computer and Telecommunications Networking,Vol. 44, Issue 5 , pp: 643 - 666, 2004.

[2] Z. Chen, L. Gao, K. Kwiat, Modeling the spread of active worms,Twenty- Second Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), Vol. 3, pp. 1890 1900, 2003.

[3] Mithcell Rowton, Introduction to Network SecurityIntrusion Detection,December 2005.

[4] Biswanath Mukherjee, L.Todd Heberlein, Karl N.Levitt, "Network Intrusion Detection",IEEE, June 1994.

[5] Presentation on Intrusion Detection Systems, Arian Mavriqi.

[6] Intrusion Detection Methodologies Demystified, Enterasys Networks TM.

[7] Protocol Analysis VS Pattern matching in Network and Host IDS, 3rd Generation Intrusion Detection Technology from Network ICE

[8] Han, J. and Kamber, M. (2000). Data Mining: Concepts and
Techniques, Morgan Kaufmann Publisher.

[9] Mannila, H., Smyth, P., and Hand, D. J. (2001). Principles of Data Mining. MIT Press. Mannila, H., Toivonen, H., and Verkamo, A. I. (1997)

[10] Berry, M. J. A. and Lino_,G. (1997). Data Mining Techniques. John Wiley and Sons, Inc

[11] Mithcell Rowton, Introduction to Network Security Intrusion
Detection, December 2005.

[12] Mounji, A. (1997). Languages and Tools for Rule-Based Distributed Intrusion Detection. PhD thesis, Faculties Universitaires Notre-Dame dela Paix Namur (Belgium).