Real Time and Offline Network Intrusion Detection using Improved Decision Tree Algorithm

G. Sunil Kumar
Lecturer in M.C.A
Maris Stella College, Vijayawada, India

ABSTRACT

Data mining has been used extensively and broadly by several network organizations. Classification based algorithms provide a significant advantage in order to detect attacks in the training data. Network applications usage is being increased every day as the internet usage is exponentially increasing. In the same way, Network attacks detection is gradually decreased as data source is increasing. There is a need to develop some robust decision tree in order to produce effective decision rules from the attacked data. In this paper improved, decision tree is implemented in order to detect network attacks like TCP SYN , Ping of Death, ARP Spoof attacks. This improved tree is also tested on famous network intrusion dataset Kddcup 99 dataset. Experimental result shows this improved decision tree classifier gives effective decision rules compare to existing decision tree techniques like ID3 and C45 algorithms. Finally, this robust decision tree evaluates less false positive and true negative alarm rates compare to existing algorithms.

General Terms

Intrusion detection, Denial of service, Attack.

Keywords

DDOS, TCP, UPD, C45

1. INTRODUCTION

Networking bring us not only benefits, such as more computing power, resources and provides better performance for a given price, but also have some challenges and risks, in the field of a system as well as information security. In the last two decades, significant research has been taken place in the network security and several approaches have been developed for building secure networks.

Network Packet Analysis and Filtering plays a prominent role in many security-related techniques, such as access control mechanism, intrusion detection and Effective firewall. Through packet filtering different types of network features are analyzed within the communication network. The key features in the packet-filtering technique is flexibility and efficiency. Flexibility means the filter can be easily customized for different packet patterns while the efficiency means the ability of a filter to capture interesting packets. In the last few years, a dramatic increase in the types of attacks, network intrusion detection plays a vital role for information assurance. Firewalls Has to do some protection, they do not provide maximum protection and still need to depend on intrusion detection system. The major purpose of intrusion detection is to protect computer systems prepare for and deal with attacks. Ids's collect data from different sources within computer systems and networks. In most systems, this information is then compared to predefined rules or patterns of Misuse to identify attacks and vulnerabilities. However, there are different methods of intrusion detection apart from the use of data mining algorithms. These techniques, along with behavioral data forensics, generates attack rules of normal user behavior and will notify the security analyst if a deviation from that normal behavior occurs. In most cases intrusion detection systems, however, both host and network intrusion detection systems combine to find network attacks detection as well as prevention mechanisms from both inside and outside endpoints. Even though we have some predefined rules and signatures still, the intrusion detection system itself has a risk associated to it because of the partial absence of human intervention in some case scenarios.

The process of capturing data packets that are crossing a LAN network is called as packet capture. These captured data packets can be analyzed using various software tools and thus detect the root cause of various network attacks, identify security flaws, and ensure data communications and network usage with outbound policy. There are various kinds of software tools that are used to capture packets, sniffer is one such tool that helps us capture packets off the wire. The job of a sniffer program is to detect bottlenecks within a network and other related problems by monitoring, capturing and analyzing traffic in a specific network. A network manager uses this information provided by sniffer programs to mange the flow of network traffic efficiently and securely. Sniffer programs can also be used to log network traffic passing over a digital network. Sniffer captures packets from data streams that travel over a network and helps in decoding and analyzing its content according with any specifications.

Data mining techniques have not been adopted in the Information technology security. This is due to the lack of information obtaining the necessary network data and a lack of domain knowledge about the applied approaches. As security experts works on network traffic, the barriers for collecting the data have been gradually decreased over time. New tools exist that support the functionality required to process networking data for the purposes of data mining, such as commercial tools,free open source software packages, and easy to use scripting languages. In academia, many research papers have been published on the subject of data mining for intrusion detection; from building decision trees with honeypot data to classifying threats in real-time , these documents have provided significant, accurate results regarding data mining for intrusion detection.

A decision tree can be expressed as a recursive partition of the instance data space. The decision tree consists of nodes that forms a rooted tree, meaning it is a directed tree with a topmost node called “root” that has no incoming edges. All remaining nodes have only one incoming edge. All other nodes are called leaves (also known as terminal or decision nodes).A node with outgoing edges is called an internal or test node. Each leaf is assigned to one class (for ex: attack or normal) representing the most appropriate target value. Figure 1 describes a decision tree that reasons whether or not a potential attack will respond to a direct network features. Leaf may hold the probability of the target class attribute having a certain value. Internal nodes are denoted as circles, whereas leaves are denoted as triangles .Instances or Rules are classified by navigating tree from top to down leaf node, according to the outcome of test node along the path.

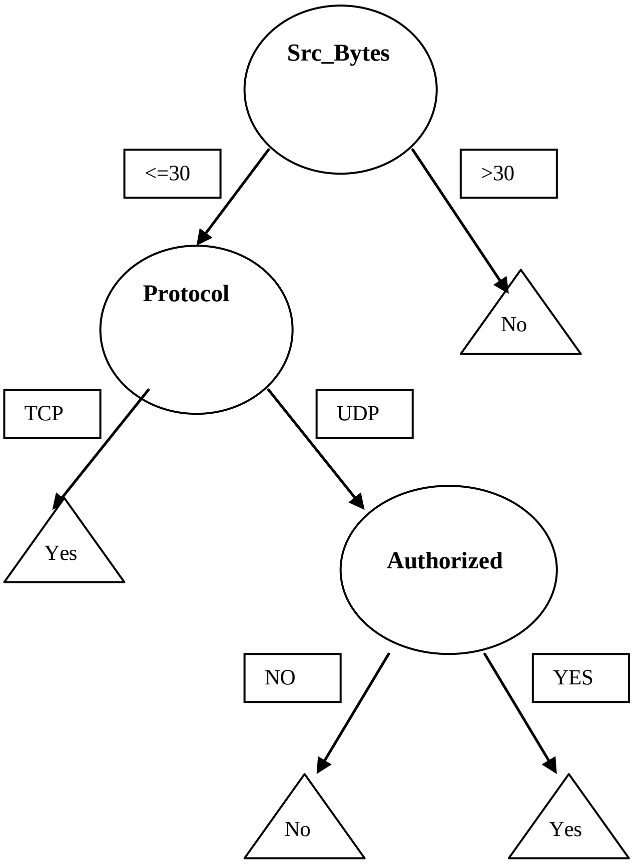


Figure 1: Decision Tree representing response to network data.

2. RELATED WORK

A fundamental issue in decision tree inductive learning is the attribute selection measure at each non-terminal node of the tree. The information gain measure is the most popular one for addressing this issue. However, A natural bias in information gain is that it favors to select attributes with many values and harms its performance. Because of this, they will have a very high information gain and very poor predictor relative to the training examples over test examples. One way to overcome its bias is to penalize these attributes with many values by incorporating a term called split information and becomes the gain ratio measure.

However, a practical issue that arises in using gain ratio to select attributes is that the split information can be zero or very small. To Overcome this practical issue Dianhong Wang [1], applied heuristic method, which calculates the gain of each attribute and then applies the gain ratio measure only those attributes with above the average information gain.

The average gain AverageGain(S,A) of an attribute A, relative to a set of training instances S, can be defined by Equation 1.

$$\text{AverageGain(S,A)}= \frac{\text{Gain(S,A)}}{|A|} \qquad \text{---------(1)}$$

where |*A*| is the number of values of the attribute *A* and *Gain*(*S*,*A*) is the *information gain* that the attribute *A* partitions the training instances *S*.This approach doesn't handle missing and numeric attributes. So, applying our improved measure to the domains that involve missing and numeric attribute values.

Except for the information gain measure and its improved versions, Lopez de Mantaras[2] presented a distance-based attribute selection measure. His experimental study proves that the distance based measure is not biased toward attributes with large numbers of values, and avoids the practical issues towards the gain ratio measure.

Mingers[3] provides an experimental study of the relative accuracy of different attribute selection measures in the decision tree in order to overcome the bias in the tuples.

Nageswara Rao, Dr. D. Rajya Lakshmi, Prof T. Venkateswara Rao et at[4] proposed robust statistical preprocessor in order to improve the accuracy. But the limitation in that paper is existing c45 does not handle when the dataset is large.

C45 ALGORITHM

The C4.5 Algorithm is the extension of ID3 algorithm. It used a mechanism of learning from large datasets. The attribute selection of the algorithm is based on an assumption :the complexity of decision tree and the amount of information is represented by given attribute are closely related .C4.5 expands the classify range to digital attribute s. That metric standard of two-class entropy ,the most of the algorithm is based on the information entropy which is contained by produced nodal points of decision tree is least [9].The so called entropy is representative of degree of disorder of objects in the systematology. It is easy to understand that the smaller entropy the smaller disorder .In the other word the more sequential in the record collection, the more consistent .This is the target we seek ,too .Suppose the set S is a training sample ,the formula of entropy as follows:

C4.5 builds decision trees from a set of training data, using the concept of information entropy. The training data is a set S.s1,s2,s3 .... represents samples in the dataset S.Each si = xl,x2,... is a sample vector where xl,x2,... represents features or attributes of the sample. The training data associated with a vector C = cl,c2,... where cl,c2,...cn represents the class to which each sample belongs to dataset. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets in one class or the other. Its criterion is the normalized information gain that results from choosing an attribute for splitting the data. The attribute with the highest information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sublists. This algorithm has a few base cases[5]. All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree same to choose that class. · None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.

Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value[10].

3. PROPOSED MODEL

Proposed Research work introduces a new framework for offline analysis as shown in Figure 2. For network intrusion detection. In this framework KDD99cup [6] dataset is given to Preprocessing stage which includes robust statistical techniques for both outlier detection as well as effective feature selection.

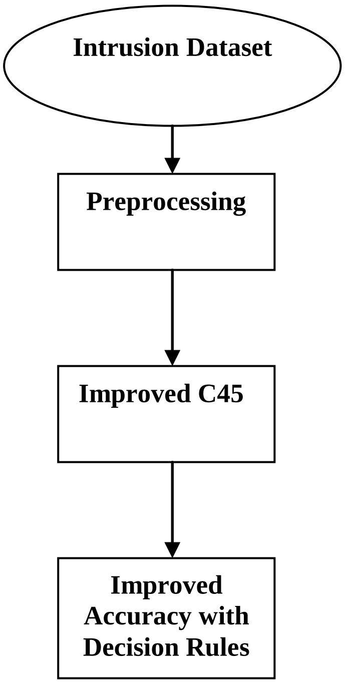


Figure 2: Proposed Model Architecture

3.1 IMPROVED C45 MODEL

Let A1,A2,A3…An be the attributes in the network intrusion dataset,S1k,S2K……Snk be the K samples of each attributes A1,A2,….An. Let V1,V2……Vk be the different values in each attribute that satisfies the target attribute class.Srcbytes and Destbytes are used as threshold features.n1,n2 denotes number of subsamples which belongs to anomaly class and normal class.

STEP1: Transform nominal and string attributes to numerical types.

STEP2: For each numerical attribute in the attribute list L.

Calculate the polynomial linear Normalization by using eq 1)

Where a,b,c are the estimated parameters obtained by using Parabola curve fitting as follows

$$F(x) = a(x - \sum_{i=1}^{n} x_i / n)^2 / \sigma_X^2 + b(x - \sum_{i=1}^{n} x_i / n) / \sigma_X + c$$

Let a parabola $y = a + bx + cx^2$ ...(1) which is fitted to a given data ($x1, y1$) , ($x2, y2$) , ($x3, y3$) ,……, (xn, yn) .

Let y_λ be the theoretical value for $x1$ then $e_1 = y_1 - y_\lambda$

$$\Rightarrow e_1 = y_1 - (a + bx_1 + cx_1^{\ 2})$$

$$\Rightarrow e_1 = y_1 - (a + bx_1 + cx_1^{\ 2})$$

$$\Rightarrow e_1 = y_1 - a - bx_1 - cx_1^{\ 2}$$

$$\Rightarrow e_1^{\ 2} = (y_1 - a - bx_1 - cx_1^{\ 2})^2$$

$$\Rightarrow S = \sum_{i=1}^{n} e_i^2$$

$$\Rightarrow S = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^{\ 2})^2$$

According to the principle of least squares method, the value of S should be minimum, therefore

$$\partial S / \partial a = 0$$

$$\partial S / \partial b = 0 \qquad \text{----- 2)}$$

$$\partial S / \partial c = 0$$

Solving equation (2) and dropping suffix, we have

$$\sum_{i=1}^{n} y = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^{\ 2}$$

$$\sum_{i=1}^{n} xy = n\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\ 2} + c\sum_{i=1}^{n} x_i^{\ 3}$$

$$\sum_{i=1}^{n} x^2 y = n\sum_{i=1}^{n} x_i^{\ 2} + b\sum_{i=1}^{n} x_i^{\ 3} + c\sum_{i=1}^{n} x_i^{\ 4}$$

Above equations are known as normal equations. On solving equations we get the estimated values a,b and c.

STEP3: create a node N;

STEP4: if tuples in D are all of the same class, C then

STEP5: return N as a leaf node labeled with the class C;

STEP6: if attribute list is empty then

STEP7: return N as a leaf node labeled with the majority class in D; // majority voting

STEP8: apply Attribute selection to each attribute(L, attribute list) to find the "best" splitting criterion;

Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information is selected. Given a collection S of c outcomes The expected information needed to classify a tuple in D is given by

In the kdd99 dataset we have two class labels ie normal and anomaly.Hence

Modified Information or entropy is given as

ModInfo(D)= $-S_i \sum_{i=1}^{m} l\,\text{og}\sqrt{S_i}$,m different classes

ModInfo(D)= $-S_i \sum_{i=1}^{2} l\,\mathrm{og} \sqrt{S_i}$

$$= -S_1 \log \sqrt{S_1} + S_2 \log \sqrt{S_2}$$

Where S_1 indicates set of samples which belongs to target class 'anamoly', S_2 indicates set of samples which belongs to target class 'normal'.

Information or Entropy to each attribute is calculated using

$$Info_A(D) = \sum_{i=1}^{v} |D_i| / |D| \times ModInfo(D_i)$$

The term Di /D acts as the weight of the jth partition. ModInfo(D) is the expected information required to classify a tuple from D based on the partitioning by A.

Information gain is defined as the difference between the original information requirement) and the new requirement .That is,

$$Gain(A) = Mod\inf o(D) - \inf o_A(D)$$

Finding Best Split:

In order to decide which attribute is best split measure ,correlation coefficient is used as a threshold as

$$r = \sum XY - \bar{X}\bar{Y} \Big/ \sqrt{SD_x} . \sqrt{SD_y}$$

```
Let A= MaxGain{AttributeList}
If(r>0 and A>r })
{
A is positively alerted and the node is selected.
}
Elseif(r<0 and A>r)
{
A is negatively alerted and the node is discarded.
}
Elseif(r=0 and A>r)
{
A is unalerted and next highest MaxGain is selected.
}
Else
 A is discarded
```

Depending on the alert type severity, the decision on the root node and the child nodes are selected.

STEP9: Recurs on the sub lists obtained by splitting on a best, and add those nodes as children of node.

Stopping Criteria

Tree growing phase continues until a stopping condition is triggered. The following conditions are common terminating conditions:

1. All tuples in the training data belong to a single value . 2. The number of subcases in the terminal node is less than the minimum number of subcases for parent nodes.

4. REALTIME NETWORK ATTACKS

The purpose of this experiment is to present a smart software tool for packet sniffing and for intrusion detection purposes. The study involves 3 basic modules.

1. Packet Capturing
2. Packet Interpretation and Storing
3. Intrusion detection using Improved DT.

1. Packet Capturing:

Packet capture is the act of capturing data packets crossing a computer network. As soon as the system is connected to the internet, If a packet sniffer is installed in the computer, the packets sniffer is capable of capturing all the packets that passes over a network. A packet sniffer can view a wide variety of information that is being transmitted over the network as well as the network it is linked.

2. Packet Interpretation and Storing:

Once data is captured, it can be analyzed right away or stored and analyzed later. Packets are made to stored in a database (access/oracle/mysql) while they are capturing for future interpretation. Robust detection techniques are used in order to detect ping of death and denial of service type of attacks.

3. Intrusion detection:

Intrusion detection (ID) is a type of security management system for computers and networks. An ID system gathers and analyzes information from various areas within a computer or a network to identify possible attacks. This IDS is mainly designed to identify two types of attacks.

Ping of Death Attack:

Ping of death is a kind of denial of service attack caused by an attacker deliberately sending an IP packet larger than the specified 65,536 bytes allowed by the IP protocol. Many operating systems didn't know what to do when they received an overflowed packet, so they crashed, or rebooted. Attacks took advantage of this defect by fragmenting packets that when received would total more than the allowed number of bytes and would effectively create a buffer overload on the operating system at the receiving end, crashing the system[7].

Tcpsyn Flood Attack:

Transport control protocol is used by many application layer protocols like the HyperText Transfer Protocol (HTTP) .TCP is connection oriented and it maintains information about buffers, windows, and other resources to count segments and track lost segments. When a normal Transport control protocol connection starts, a destination host receives a SYN start packet from a source machine and sends back a SYN ACK packet. The destination host must then catch an ACK packet of the SYN ACK packet before the connection get established. This is referred to as the "TCP three-way handshake".

5. EXPERIMENTAL RESULTS

All experiments were performed with the configurations Intel(R) Core(TM)2 CPU 2.13GHz, 2 GB RAM, and the operation system platform is Microsoft Windows XP Professional (SP2) and Windows -7. Kddcup 99 dataset is used for network intrusion detection. For implementation and designing the tool Weka[8] and java packet capture packages are used.

5.1 The following figure 3 capture packets using the adaptor in either windows xp or windows-7 environment.Winpcap tool has to be installed before capturing the packets in lan.

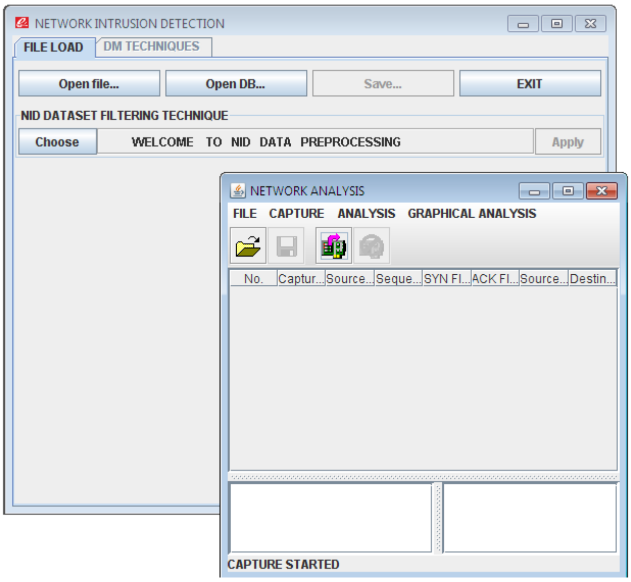
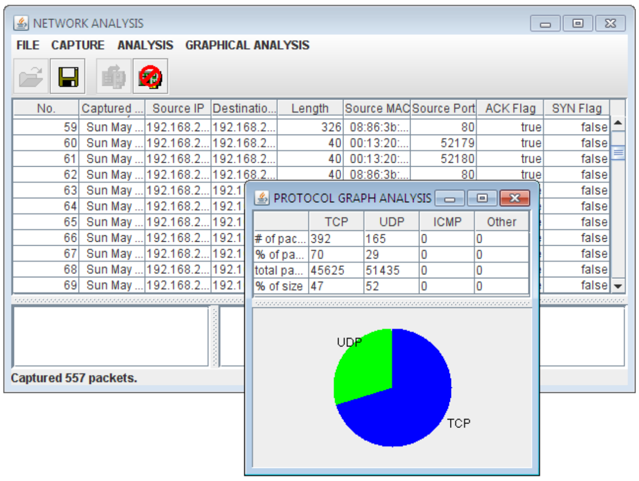


Figure 3: Interlinking both real time packet analyzer to data mining.

5.2) Figure 4 demonstrates real time packets capture as well as graphical pie chart representation of the protocols.



No.	Captured ...	Source IP	Destinatio...	Length	Source MAC	Source Port	ACK Flag	SYN Flag
59	Sun May ...	192.168.2...	192.168.2...	326	08:86:3b:...	80	true	false
60	Sun May ...	192.168.2...	192.168.2...	40	00:13:20:...	52179	true	false
61	Sun May ...	192.168.2...	192.168.2...	40	00:13:20:...	52180	true	false
62	Sun May ...	192.168.2...	192.168.2...	40	08:86:3b:...	80	true	false
63	Sun May ...	192.168.2...	192.1					false
64	Sun May ...	192.168.2...	192.1					false
65	Sun May ...	192.168.2...	192.1					false
66	Sun May ...	192.168.2...	192.1					false
67	Sun May ...	192.168.2...	192.1					false
68	Sun May ...	192.168.2...	192.1					false
69	Sun May ...	192.168.2...	192.1					false

	TCP	UDP	ICMP	Other
# of pac...	392	165	0	0
% of pa...	70	29	0	0
total pa...	45625	51435	0	0
% of size	47	52	0	0

Captured 557 packets.

Figure 4 : Packet analysis

5.3) Figure 5 denotes cumulative curve representation of protocols and IP versions dynamically.

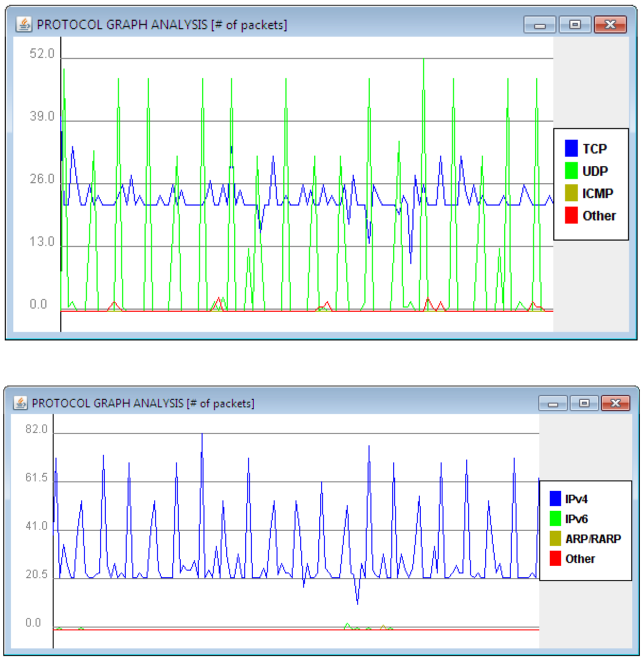


Figure:5 Cumulative Curve

5.4) Following results gives the improved C45 performance on 10% KDD dataset with 5291 instances :

Table I: Improved C45 performance on 10% KDD dataset

Property	Existing C45	Improved C45
Tree Size	126	120
Number of leaves	95	92
Correctly Classified Instances	5083 (96.06%)	5129 (96.93%)
InCorrectly Classified Instances	208 (3.93%)	162 (3.06%)

5.5 Threshold curve for area under ROC

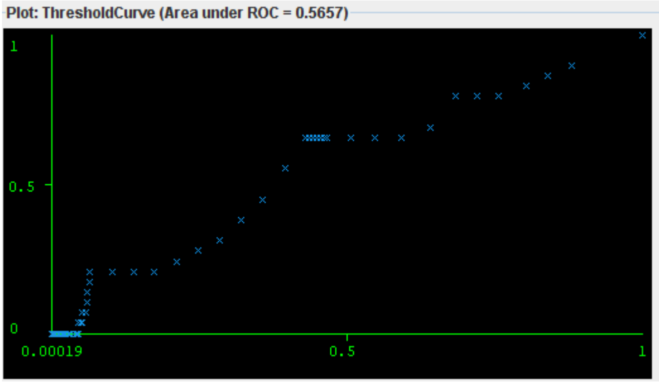
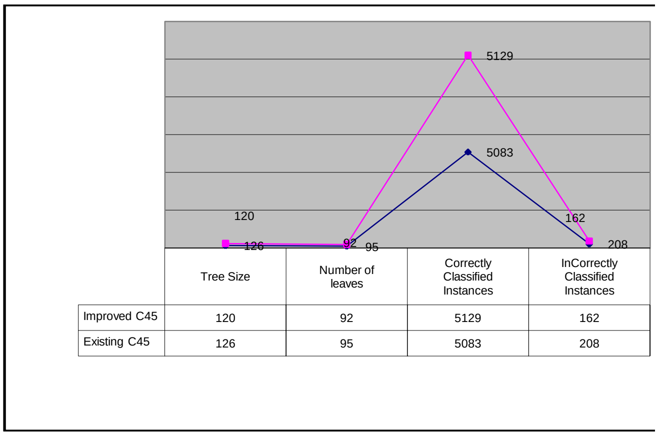


Figure 6: Threshold curve for Decision tree

5.6 Accuracy Comparision between Existing C45 and Improved C45 algorithms:



	Tree Size	Number of leaves	Correctly Classified Instances	InCorrectly Classified Instances
Improved C45	120	92	5129	162
Existing C45	126	95	5083	208

Figure 7: Accuracy Comparision

6. CONCLUSION AND FUTURE WORK

Experimental results shows proposed decision tree gives better attack classified results compare to existing C45 technique. Proposed Algorithm gives 96.9 percent of accuracy for detecting attacks with less false positive and true negative rates. In this research work real time network traffic data are analyzed for identifying attacks like dos, spoof type of attacks. In future proposed algorithm is applied on the live network captured data for comparasion between kdd99 dataset with the captured attacked data.

7. REFERENCES

[1] Dianhong Wang, Liangxiao Jiang “An Improved Attribute Selection Measure for Decision Tree Induction”,IEEE Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007).

[2] R. L. de Mantaras. A distance-based attribute selection measure for decision tree induction. Machine Learning, 6:81–92, 1991.

[3] J. Mingers. An empirical comparison of selection measures for decision-tree induction. Machine Learning, 3:319–342, 1989.

[4] Nageswararao,Dr.D.RajyaLakshmi,Prof T.Venkateswara Rao,” Robust Statistical Outlier based Feature Selection Technique for Network Intrusion Detection ,(IJSCE 2012).

[5] http://en.wikipedia.org/wiki/C4.5_algorithm .

[6] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani “A Detailed Analysis of the KDD CUP 99 Data Set”, IEEE 2009.

[7] http://www.webopedia.com/TERM/P/ping_of_death.html

[8] Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall,” Weka: Practical Machine Learning Tools and Techniques with Java Implementations”

[9] J. R. Quinlan, "C4.5: programs for machine learning", Morgan Kaufmann Publishers, 1993.

[10] Hybrid Neural Network and C4.5 for Misuse Detection Zhi-Song Pan, Song-Can Chen, Gen-Bao Hu, Dao-Qiang Zhang, Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi’an, 2-5 November 2003.