Detection of Web Based Network Attacks Using ID3 Algorithm

Ms. Pooja Kokate[1], Prof Ajay Phulre[2]
[1, 2] Dept of CSE
[1, 2] SBITM Betul (M.P.), India

Abstract- *In data mining technique Decision Tree is an important method of Classification, it is used for prediction and forecasting from historical data. ID3 is one of the popular decision tree algorithm, ID3 algorithm has been used broadly for its simple idea, effectiveness and efficiency. In this paper we proposed and implemented effective methods for detection of web based attacks using improved ID3 algorithm. With the advent of network and the e-commerce technologies we are depends on e-transactions, so the internet works using browsers, attacker has to attacks on the websites and the web servers, at the web server it maintains the log file which consists of web application queries, for applying data mining algorithms we requires only that the web application queries be slightly pre-processed before application. In this paper we are presenting a effective concept on ID3 algorithm for detection of web based attacks, for this we have taken the data set from 'SmarSniff' tool, the experimental results shows that the improved algorithm is effective in decrease the data amount and reduce the impact of data with poor quality and thus improves the efficiency and effectiveness of ID3 algorithm.*

Keywords– data mining, IDS, ID3, decision trees, web based attacks, Response time.

I. INTRODUCTION

Data mining is the process of mining useful, meaningful information from large volume of data; the data may be inconsistent, noisy, fuzzy, random, and incomplete. In data mining technique, there are various tools and techniques used which are based on statistical methods, these methods are very effective for mining meaningful information, the techniques such as correlation analysis, evolution analysis, classification analysis and evolution analysis [1][2]. In classification analysis decision tree technique is used, in which classification rules are generated and we get the useful information from that. Classification analysis is commonly used in detecting web based attacks such as intrusion detection, detecting anomaly. Decision tree is a flow-chart like tree structure that consists of nodes that form a rooted tree in which each non leaf node indicates a test on an attribute, each branch represents outcome of the test, and each leaf node holds a class label. The topmost node in a tree represents a root node.

Decision tree algorithms adopt a greedy approach in which a decision trees are constructed in top down recursive divide and conquer manner. ID3 (Iterative Dichotomizer 3) algorithm is a decision tree algorithm used for classification it was developed by J. Ross Quinlan (1983). ID3 algorithm to make the decision tree because the ID3 algorithm has a clear concept using Shannon's information theory, and can be simply implemented, it is essentially a attribute based learning algorithm that constructs a decision tree based on a training set of data and a entropy measure to build the leaves of the tree[4]. Two terms used in ID3 algorithm i.e, Entropy & Information Gain, entropy is used to calculate the homogeneity of a learning data set and information gain is used for calculating the expected reduction in entropy. After calculating the entropy and information gain of each attribute, we have to decide which attribute has highest information gain the generate the decision tree according to the information gain, the attribute which has highest information gain is considered as a root node and others are parent node and this process continues iteratively and thus generates a decision tree.

II. LITERATURE SURVEY

Feng Yang, Hemin Jin, Huirnin Qi,“ Study on the Application of Data Mining for Customer Groups Based on the Modified ID3 Algorithm in the E-commerce” [1] propose a study uses the Taylors formula to transform the algorithm to reduce the amount of data calculation and the generation time of decision tree & improve the efficiency of decision tree classifier. The properties selected by this formula can overcome the shortcomings that the ID3 algorithm is easy to be favor of choosing more attribute values as the test attribute, but also greatly reduce the time of algorithm for generating decision trees thus reduce the computational cost speed up the construction of decision trees and improving the efficiency of decision tree classifier.

Giovanni Vigna, “A Stateful Intrusion Detection System for World-Wide Web Servers” [8] proposes a

WebSTAT, a STAT based intrusion detection system. The WebSTAT system has been evaluated in terms of its ability to detect attacks and the performance impact of the detection processes on deployed web servers. Their goal is to perform early detection of malicious activity and possibly prevent more serious damage to the protected site.

K. Hanumantha Rao, "Implementation of Anomaly Detection Technique Using Machine Learning Algorithms" [6] proposes a anomaly detection system based on the combinational approach of K-means and ID3 algorithm for classifying the two clusters classifying the normal and anomalies activities. The K-means clustering method first partitions the training instances into two clusters using Euclidean distance similarity. On each cluster, representing a density region of normal or anomaly instances, we build an ID3 decision tree.

WU Sen," Improved Classification Algorithm by Minsup and Minconf Based on ID3" [14] proposes a improved classification algorithm based on minsup and minconf concept based on ID3 to decrease the data amount and reduce the impact of data with poor quality and we apply the log file at the web server to detect the web based attacks. This improved algorithm introduces two new concepts, 'support of test attribute set to class' and 'rule confidence', which are used to improve the decision tree construction process by both pre pruning and post pruning and ultimately to increase the efficiency and effectiveness of classification.

III. IMPLEMENTATION DETAILS

In the implemented work, we have collected the data using 'SmartSniff' tool, SmartSniff is a network monitoring utility that allows to capture TCP/IP packets that pass through your network adapter, and view the captured data as sequence of conversations between clients and servers. We have collected entries such as protocol, local address, remote address, local port, packet size etc. so for applying the ID3 algorithm on them, preprocessed the entries in the SmartSniff tool separately as no. of rows, columns, size etc and generate the lists according to each separate record. We have considered three parameters for detecting source IP address, we consider Source IP address, Size, Weight and different possibilities of Source IP address as Malicious, Suspicious and Clean. We designed one database that contains the malicious IP addresses, then compare the each list with designed database if the record contains the IP address as same as that of database, then considered it as a malicious IP address. If size of the data packet is less than 100 kb then it considered as a clean IP address, at last we have to count the weight on the source IP address if the weight is more then we consider it as a suspicious IP address, for calculating the weight of each IP address we have designed another database. Thus after performing the above process, properties files are generated containing i) attributes :- IP address, size, weight ii) Categories :- Yes, No we are classifying the IP addresses as clean, malicious and suspicious, size as low, normal, high and weight as weak and strong.

Based on the above preprocessed data sets, property files are generated for designing the ID3 classifier. We designed two property files for ID3 algorithm and for improved ID3 algorithm to generate the decision trees separately. In the simple ID3 the property file takes all the entries whose source IP address size is less than 100 KB and more than 100 KB, but in the improved ID3 property file contains only such entries whose size is below 100 KB, so the size of improved ID3 property file is less compared to ID3 property file. In the improved ID3 algorithm user may set the threshold values as per their choices on size and weight based on the threshold values the improved ID3 tree generated in less in size compared to the ID3 classification tree, and thus separates the source IP address as clean, malicious and suspicious.

Algorithm 1

(For Preprocessing of SmartSniff data set & generating property file)

Assumptions:

IIL	→	IDS info list.
N	→	No. of IIL list size
IDTO	→	Intrusion Data Transaction Object
W	→	Weight
S	→	Size
Src_IP	→	Source IP
SPsizeT	→	Specified Packet size Threshold
Swt	→	Specified weight Threshold

Begin

Step 1: Read the input log file and split complete file in to the no. of records, attributes, size.
Step 2: Read each record indentifying the vital IDS parameters from each line to store them in to list (Src_Ip, size & weight parameters)
Step 3: Repeat for I ← 1 to N
IDTO ← IIL
W ← IDTO
Src_IP ← IDTO
Size ← IDTO

```
// check Src_IP is malicious in to malicious database.
        if (chekIPAddressMalicious) then
                write malicious associated parameters into property file
        else if (size<=100 && Size! = 0)
        then
                write clean with associated parameters in to property file.
        else if  ( is IPAddressPresent ) then increase the weight of IPAddress
        else
                insert record of  IP address into a database table (tbipweights)
        if  (W > SWT) then
if  (size > SPsizeT) then
        write malicious with associated parameter into a property file.
Else
        Write suspicious with weight strong associated parameter into a property file
End if
Else
        If ( !size < SPsizeT) then
        Write suspicious with weak weight parameter into a property file
End if
Else
        Write suspicious with high weak parameter into a property file.
End else
End else
End else
End else
End else
End else
End for.
End
```

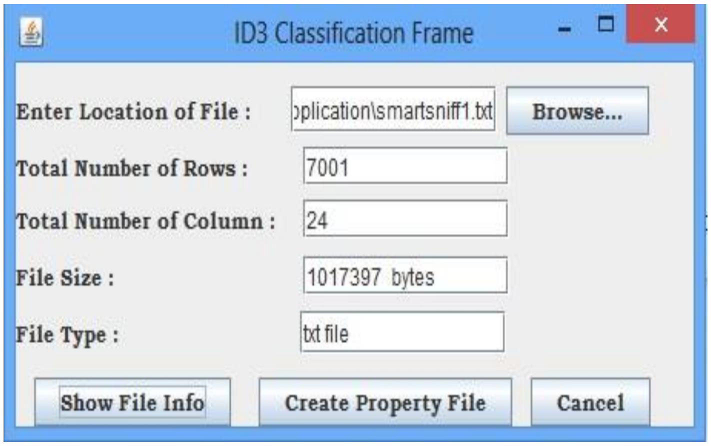
Algorithm 2 (For generating ID3 classifier tree)

Output of Algorithm 1 is given to input to the Algorithm 2.

Step 1: Read property file created by Algorithm 1.
Step 2: Read attributes, categories from the property info file.
Step 3: Read the data start from example.
Step 4: Set the size and weight parameters threshold as per users choice
Step 4: Read each line and calculate entropy & gain of each category value
Step 5: Generate the decision tree based on high gain value and entropy.

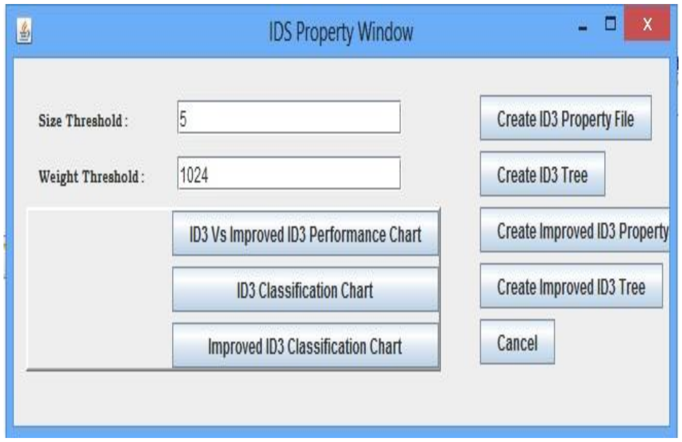
IV. RESULT AND DISCUSSION

The working environment for proposed ID3 algorithm for detection of web based attacks is implemented using Java (Netbeans) and Ms-Sql server. As shown in fig.1, user provides the inputs as ‘SmartSniff’ log file and generates the preprocessed information.



Enter Location of File : pplication\smartsniff1.txt
Total Number of Rows : 7001
Total Number of Column : 24
File Size : 1017397 bytes
File Type : txt file

Fig. 1.ID3 Classification Frame.



Size Threshold : 5
Weight Threshold : 1024

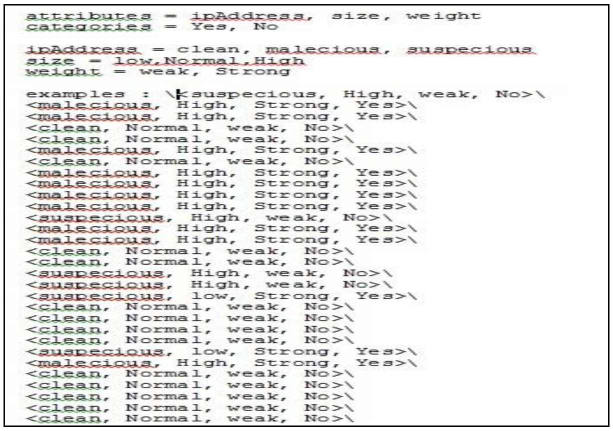
Fig.2. IDS Property window.

As shown in fig. 2. Will generates the property file information based on the preprocessed information and generated ID3 tree, for generating improved ID3 tree user has to set the threshold values on size and weight parameters and based on this information generates a property file and generates a improved ID3 tree.

Results Calculated:

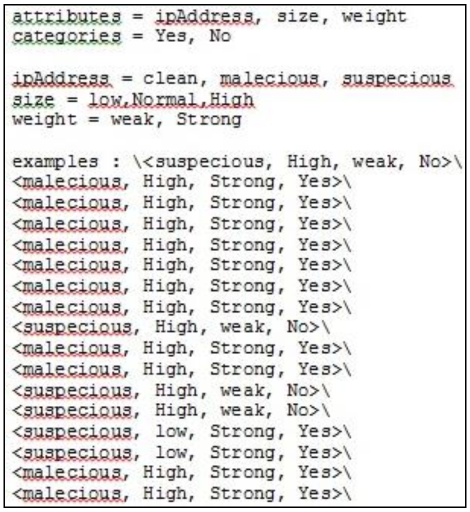
Fig. 3. Shows the generated ID3 property file as attributes as ipAddress, Size & Weight, categories as yes, no, and classifies different parameters of IP address as clean, malicious and suspicious & size as low, normal & high, weight as weak & strong, the property file generated by ID3

algorithm is larger in size compared to property file generated by improved ID3 (Fig.4)



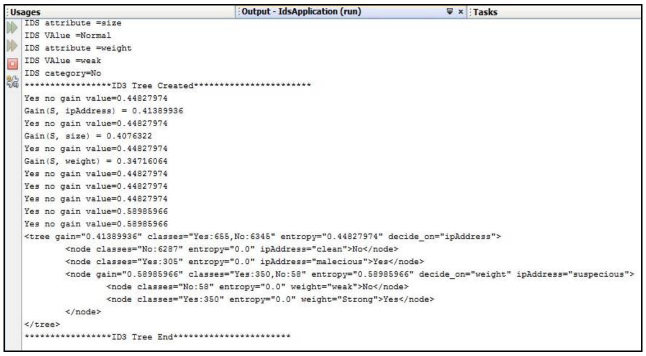
```
attributes = ipAddress, size, weight
categories = Yes, No

ipAddress = clean, malecious, suspecious
size = low,Normal,High
weight = weak, Strong

examples : \<suspecious, High, weak, No>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<clean, Normal, weak, No>\
<clean, Normal, weak, No>\
<malecious, High, Strong, Yes>\
<clean, Normal, weak, No>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<suspecious, High, weak, No>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<clean, Normal, weak, No>\
<clean, Normal, weak, No>\
<suspecious, High, weak, No>\
<suspecious, High, weak, No>\
<suspecious, low, Strong, Yes>\
<clean, Normal, weak, No>\
<clean, Normal, weak, No>\
<clean, Normal, weak, No>\
<clean, Normal, weak, No>\
<suspecious, low, Strong, Yes>\
<malecious, High, Strong, Yes>\
<clean, Normal, weak, No>\
<clean, Normal, weak, No>\
<clean, Normal, weak, No>\
<clean, Normal, weak, No>\
<clean, Normal, weak, No>\
```

Fig. 3. Property file generated by ID3 Algorithm.



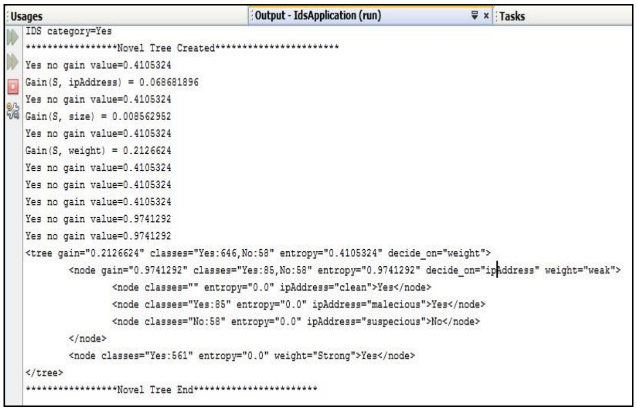
```
attributes = ipAddress, size, weight
categories = Yes, No

ipAddress = clean, malecious, suspecious
size = low,Normal,High
weight = weak, Strong

examples : \<suspecious, High, weak, No>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<suspecious, High, weak, No>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
<suspecious, High, weak, No>\
<suspecious, High, weak, No>\
<suspecious, low, Strong, Yes>\
<suspecious, low, Strong, Yes>\
<malecious, High, Strong, Yes>\
<malecious, High, Strong, Yes>\
```

Fig. 4. Property file generated by Improved ID3 Algo.



```
IDS attribute =size
IDS VAlue =Normal
IDS attribute =weight
IDS VAlue =weak
IDS category=No
*****************ID3 Tree Created***********************
Yes no gain value=0.44827974
Gain(S, ipAddress) = 0.41389936
Yes no gain value=0.44827974
Gain(S, size) = 0.4076322
Yes no gain value=0.44827974
Gain(S, weight) = 0.34716064
Yes no gain value=0.44827974
Yes no gain value=0.44827974
Yes no gain value=0.44827974
Yes no gain value=0.58985966
Yes no gain value=0.58985966
<tree gain="0.41389936" classes="Yes:655,No:6345" entropy="0.44827974" decide_on="ipAddress">
        <node classes="No:6287" entropy="0.0" ipAddress="clean">No</node>
        <node classes="Yes:305" entropy="0.0" ipAddress="malecious">Yes</node>
        <node gain="0.58985966" classes="Yes:350,No:58" entropy="0.58985966" decide_on="weight" ipAddress="suspecious">
                <node classes="No:58" entropy="0.0" weight="weak">No</node>
                <node classes="Yes:350" entropy="0.0" weight="Strong">Yes</node>
        </node>
</tree>
*****************ID3 Tree End***********************
```

Fig. 5. ID3 tree generated.



```
IDS category=Yes
*****************Novel Tree Created***********************
Yes no gain value=0.4105324
Gain(S, ipAddress) = 0.068681896
Yes no gain value=0.4105324
Gain(S, size) = 0.008562952
Yes no gain value=0.4105324
Gain(S, weight) = 0.2126624
Yes no gain value=0.4105324
Yes no gain value=0.4105324
Yes no gain value=0.4105324
Yes no gain value=0.9741292
Yes no gain value=0.9741292
<tree gain="0.2126624" classes="Yes:646,No:58" entropy="0.4105324" decide_on="weight">
        <node gain="0.9741292" classes="Yes:85,No:58" entropy="0.9741292" decide_on="ipAddress" weight="weak">
                <node classes="" entropy="0.0" ipAddress="clean">Yes</node>
                <node classes="Yes:85" entropy="0.0" ipAddress="malecious">Yes</node>
                <node classes="No:58" entropy="0.0" ipAddress="suspecious">No</node>
        </node>
        <node classes="Yes:561" entropy="0.0" weight="Strong">Yes</node>
</tree>
*****************Novel Tree End***********************
```

Fig. 6. Improved ID3 tree.

Fig. 5 & Fig. 6. Shows the tree generated by ID3 & improved ID3, it is observed that the tree generation time by improved ID3 is less.

Comparative Analysis:

A. Comparison of Separated IP addresses

As shown in fig. 7.a, & fig. 7.b the graphs shows the comparative of separation of malicious, suspicious, and clean IP addresses from total IP addresses between traditional ID3 & Improved ID3 algorithm.

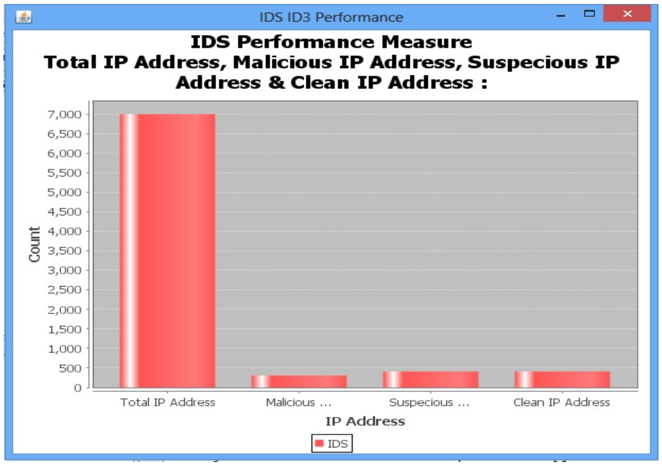


Fig. 7.a, Separation of IP address by ID3

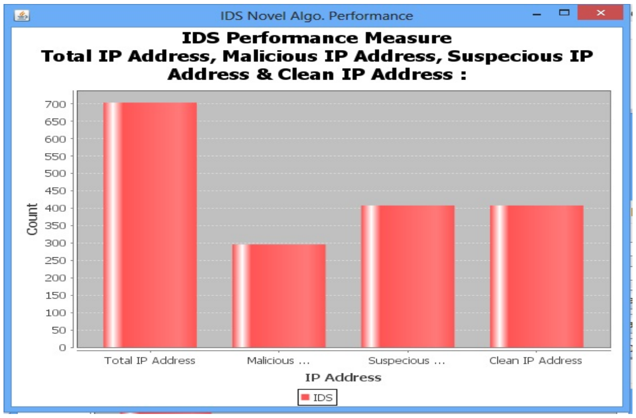


Fig. 7.b, separation of IP addresses by improved ID3.

B. Comparision of processing Time Observed by User

The graph 8. shows the comparative analysis for response time of traditional ID3 algorithm & improved ID3 algorithm, it is observed that the total processing time required for construction of improved ID3 tree less.

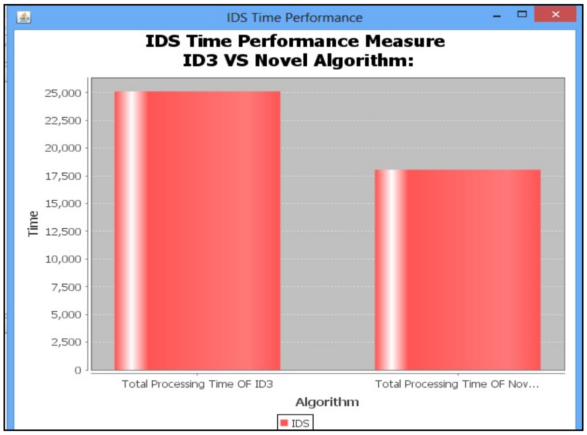


Fig.8. Total processing time.

V. CONCLUSION

In this implemented work, we have applied ‘SmartSniff’ tool dataset, here implemented effective concept is called improved ID3 algorithm which drastically minimizes the response time observed by user which leads to the improvement in response time, also the separation of total IP addresses in to malicious, suspicious and clean IP addresses are also compared between traditional and improved ID3. Finally experiments show that proposed algorithm improves the performance and efficiency.

REFERENCES

[1] Feng Yang, Hemin Jin, Huirnin Qi,*“ Study on the Application of Data Mining for Customer Groups Based on the Modified ID3 Algorithm in the E-commerce”,* 2012 International Conference on Computer Science and Information Processing (CSIP), 2012 IEEE

[2] IU Qin, *“Data Mining Method Based on Computer Forensics-based ID3 Algorithm”,* 2010 IEEE.

[3] Joong-Hee Leet, *” Effective Value of Decision Tree with KDD 99 Intrusion Detection Datasets for Intrusion Detection System”* Feb. 17-20, 2008 ICACT 2008.

[4] Mrutyunjaya Panda, Manas Ranjan Patra *“A Comparative Study of Data Mining Algorithms for Network Intrusion Detecton”* First International Conference on Emerging Trends in Engineering and Technology 2008, IEEE.

[5] Victor H. Garc´ıa, Ra´ul Monroy, and Maricela Quintana, *“Web Attack Detection Using ID3”,*

[6] K. Hanumantha Rao, *“Implementation of Anomaly Detection Technique Using Machine Learning Algorithms”*, International Journal of Computer Science and Telecommunications [Volume 2, Issue 3, June 2011]

[7] Mohammad Sazzadul Hoque*,” An Implementation of Intrusion Detection System Using Genetic Algorithm”,* International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012

[8] Giovanni Vigna, “A Stateful Intrusion Detection System for World-Wide Web Servers” Computer Security Applications Conference, 2003. Proceedings. 19th Annual 8-12 Dec. 2003, IEEE.

[9] Jaroslaw Skaruz, “Intrusion Detection in Web Applications: Evolutionary Approach”, Proceedings of the International Multiconference on Computer Science and Information Technology pp. 117–123, 2009, IEEE.

[10] Fei Li, “Data Mining-Based Credit Card Evaluation Ffor Users of Credit Card”, Proceedings of the Third Intemational Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004, IEEE.

[11] Yuesheng Tan, “Applications of ID3 Algorithms in Computer Crime Forensics”, 2011 IEEE.

[12] Yasser Yasami, “A Novel Unsupervised Classification Approach for Network Anomaly Detection by K-Means Clustering and ID3 Decision Tree Learning Methods”

[13] Joong-Hee Leet, Effective Value of Decision Tree with KDD 99 Intrusion Detection Datasets for Intrusion Detection System”, ICACT 2008

[14] WU Sen,” Improved Classifi'cation Algorithm by Minsup and Minconf Based on ID3”, 2006 IEEE.

[15] Ms. Smita Nirkhi, Dr. R.V. Dharaskar, Dr. V.M. Thakre, “Data Mining: A Prospective Approach for Digital Forensics”, IJDKP, Vol 2, No. 6, November 2012.

[16] Yiwen Zhang Lili Ding, Yun Wang, “Reaserch and Design of ID3 Algorithm Rule- B ased Anti- Spam Email filtering”, IEEE, 2011.

[17] Luis Filipe da Cruz Nassif, Eduardo Raul Hruskka, “ Document Clustering for Forensic Analysis: An Approach

for Improving Computer Inspeciton", ITIFS, Vol 8, No. 1, January 2013, IEEE.

[18] Sebastian Kurowski, Sandra Frings, "Computational Documentaion of IT Incidents as Support for Forensics Operations", 2011, sixth international Conference on IT Security Incident Management and IT Forensics., 2011.

[19] Yasser Yasami", An unsupervised network anamoly detection approach by k-means clustering and ID3 algorithm"2008 IEEE.

[20] L.Sathish Kumar, Mrs.A.Padmapriya ,"ID3 Algorithm Performance of Diagnosis For Common Disease", International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 5, May 2012 .

[21] Sharma, Sanjay Kumar," An improved network intrusion detection technique based on k-means clustering via Naïve bayes classification", Advances in Engineering, Science and Management (ICAESM), 2012, IEEE