Volume 3, Issue 12, December 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review on Detection of Web Based Attacks using Data Mining Techniques

Mr. Sachin S. Patil, Prof. Deepak Kapgate, Prof. P.S. Prasad
Department of CSE, GHRAET,
Nagpur (M.S), India

Abstract - In data mining technique Decision Tree is an important method of Classification, it is used for prediction and forecasting from historical data. In this survey paper we are surveying effective methods for detection of web based attacks using various data mining techniques such as classification, clustering, partitioning. 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 survey we are presenting review of various data mining techniques useful for detection of web based attacks.

Keywords – data mining, IDS, ID3, decision trees, web based attacks, minimum support, rule confidence, Classification, Clustering.

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.

Classification algorithms used in detection of web based attacks such as intrusion, anomalies, in the recent years with the development in the network and database technologies public and customers are depends on the web based applications such as e-commerce, these types of websites comes along with serious security attacks, most of the serious vulnerabilities occurs along with the HTTP requests at the web server. The HTTP requests contains the requests which contains anomalies and attack data and some data which is error free all these rough data sets are need to be preprocessed before apply at the data mining algorithm, prepossessed data are applied on the data mining algorithm and then classification rules are generated.

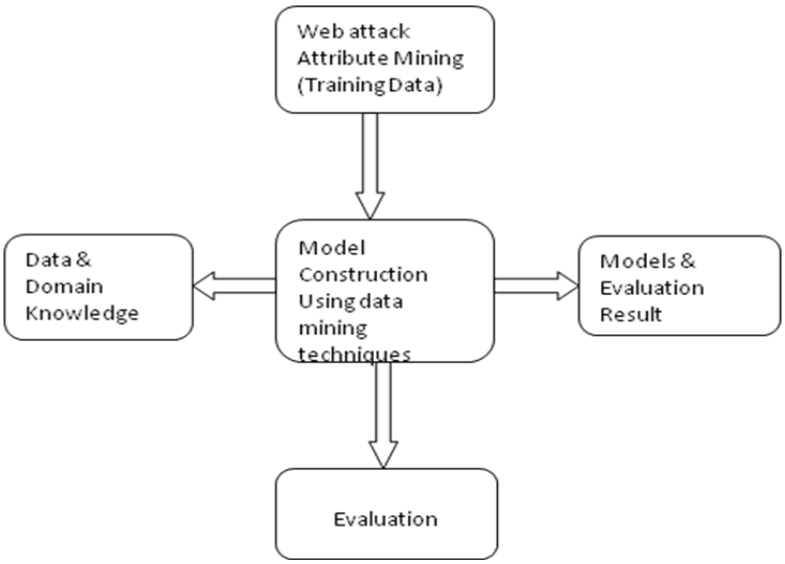

Fig I. Data mining Process for detecting web based attacks.

II. INTRUSION DETECTION SYSTEM (IDS)

To recover from the serious attacks, protect their resources many companies adopting the technique known as Intrusion Detection System (IDS), intrusion is an activity that breaks the integrity, confidentiality, and availability of the company's data. Intrusion detection system is an software application which not only detects the known and unknown attacks but also finds the source of the attacks [5]. The anomalies can arise due to various causes such as malfunctioning network devices, bad configuration in network services and operating systems, network overload, malicious denial of service attacks, ill advised applications installed by users, high level users' effort to discover network and gather information about it and its devices, and network intrusions that disrupt the normal delivery of network services. These anomalous events will disrupt the normal behavior of some measurable network data [12]. IDS system analyses the incoming network packets (web requests) looking for malicious behavior. The main functions of IDS are a) Detecting the intrusion in the network of computers and b) Providing forensic information to the organization so that they will detect the attacks in future.

Depending on the characteristics of the intrusion the IDS are classified as MIDS (Misuse IDS) and AIDS (Anomaly IDS) where MIDS are the attacks which are any know patters of attacks which are easily detected and recovered whereas AIDS are the attacks which are deviates from the profile of ordinary behavior [5]. Web based attacks are audited from the various sources such as log files, router log data, network traffic. Based on the source of data being audited IDS are classified as i) Host based IDS (HIDS) ii) Network based IDS (NIDS) iii) Protocol based IDS.(PIDS) [6] NIDS are the systems which are observing incoming network packets, traffic and detects the intrusion such as DoS attacks, port scans etc. Data used by the NIDS are includes packet header data, packet statistics and application layer data. NIDS are good at discerning attacks that involve low lever manipulation of the network and attacks that alter the network activity. Whereas Host Based Intrusion Detection System (HIDS) monitors the single computer its operating sytem, processes, memory & CPU usages. In HIDS system a sensor is installed they can monitor the activities of the operating system or any particular application; a host sensor might protecting the web server by auditing the HTTPS requests.

HIDS Vs. NIDS

Drawbacks of the HIDS are they are insulated from network activities, must be installed on every host whereas the Advantages of HIDS are It can detects the attacks that do not involve the network, can analyze what an application is doing. Drawbacks of NIDS are it must be able to keep up with the network speed, it may have problem with encrypted channels. Advantages of NIDS are it can monitor multiple hosts at a time, can correlate attacks against multiple hosts, it does not affect the host performance. and can detect attacks that are not visible from single hosts.

Types of Attacks:

Generally there are four categories of attacks:

i) **DoS (Denial of Service)**: Trying to prevent a legitimate user from accessing the service in the target machine.

ii) Probe: Scanning a target machine for information about potential vulnerabilities.

iii) R2L (Remote to Local) : When attacker is obtained non authorize access in to a network or machine.

iv) U2R (User to Root) : When target machine is already invaded, but the attacker attempts to gain access with super user privileges.

III. DATA MINING TECHNIQUES FOR DETECITON OF WEB BASED ATTACKS

In this section we present a survey of data mining techniques that have been used for detection of web based attacks:

A. Machine Learning:

Machine learning is the technique which is used to study the computer algorithms that improve automatically with experience. It finds the general rules in large amount of data sets that automatically learns user's interest. Machine learning techniques are best suitable for two techniques i.e. Classification (supervised learning) and Clustering (unsupervised learning), both of these techniques are useful for detection of web based attacks. Machine learning algorithm can be categorized in to two types of techniques i.e. supervised learning and unsupervised learning, supervised learning algorithms are trained on labeled examples i.e. input where the desired output is known whereas unsupervised learning algorithms are those that operates on the unlabelled examples i.e. input where the desired output is unknown.

B. Classification Techniques:

In classification technique, the task is to take each instance of dataset and assign it to a particular class. It is a supervised learning technique. A classification based intrusion detection system will classify all the network traffic into either malicious or normal. Classification techniques evaluate and classify the data into known classes. Each data sample is marked with a known class label. Also these techniques are used to learn a model using the training set data sample. This model is used to classify the data samples as anomalous behavior data or the normal behavior data. The classification techniques can be categorized in to following sections for detection of web based attacks:

a) Association Rule:

Association rule mining determines association rules and/or correlation relationships among large set of data items. Association rule mining is the effective technique for detection of intrusions, in this technique it searches frequently occurring item set from a large data set. In association rule the system is a set of association rules and frequent patterns that can be applied to network traffic to classify it properly so that the generated rule set is easy to understand, hence the security analyst can understand it. The mining process of decision association rule can be divided into two steps i) Frequent item set generation: It generates all set of items whose support is greater than the specified threshold called as minsupport (minimum support). ii) Association rule generation from the previously generated frequent item sets, it generated the association rules in the form of if then statements that have confidence greater than the specified threshold called as minconfidence (minimum confidence). In the network intrusion detection system where we can collect the samples from log files from server or client where each row represents an audit record and each column represents audit records. The intrusions and user activities shows frequent correlations among the network data. Consistent behavior in the network data can be captured in association rules, rules based on network data can continuously merge the rules from a new run to aggregate rule set of all previous runs. Thus with the association rule, we get the capability to capture behavior for correctly detecting intrusions.

The basic procedure for association rule mining is as follows:

- $I = \{i_1, i_2, ..., i_n\}$: a set of all the items
- Transaction T: a set of items such that $T \subseteq I$
- Transaction Database D: a set of transactions
- A transaction $T \subseteq I$ contains a set $X \subseteq I$ of some items, if $X \subseteq T$
- *An Association Rule:* is an implication of the form $X \Rightarrow Y$, where $X, Y \subseteq I$.
- A set of items is referred as an itemset. A itemset that contains k items is a k- itemset.
- The support s of an itemset X is the percentage of transactions in the transaction database D that contain X.
- The support of the rule $X \Rightarrow Y$ in the transaction database D is the support of the items set $X \cup Y$ in D.
- The confidence of the rule $X \Rightarrow Y$ in the transaction database D is the ratio of the number of transactions in D that contain $X \cup Y$ to the number of transactions that contain X in D.

Apriori algorithm:

1. The Apriori algorithm is an efficient algorithm for finding all frequent item sets.
2. The Apriori algorithm implements level-wise search using frequent item property, the Apriori algorithm can be additionally optimized.

Algorithm:

- L_k: Set of frequent itemsets of size k (with min support)
- C_k: Set of candidate itemset of size k (potentially frequent itemsets)

L_1 = {frequent items};
for (k = 1; L_k !=∅; k++) **do**
    C_{k+1} = candidates generated from L_k;
        for each transaction t in database do
            increment the count of all candidates in C_{k+1} that are contained in t
    L_{k+1} = candidates in C_{k+1} with min_support
return $\cup_k L_k$;

Ya-Li Ding,[30] proposes a novel signature searching to detect intrusion based on data mining, which is an improved Apriori algorithm. They evaluate the capability of this new approach with the data from KDD 1999 data mining competition.

Lalli & Palanisamy [29] proposes a system to discover temporal pattern of attacker behaviors, which is profiled using an algorithm EAA (Enhanced Apriori Algorithm). This is experimented with a simple interface to display the behaviors of attacks effectively.

b) Fuzzy Logic:

In contrast to standard set theory in which each element is either completely in or not in a set fuzzy set theory allows partial membership in sets. Fuzzy logic is based on fuzzy set theory. This provides a powerful mechanism for representing vague concepts. Data mining methods are used to automatically learn patterns from large quantities of data. The integration of fuzzy logic with data mining methods will help to create more abstract patterns at a higher level than at the data level. Patterns that are more abstract and less dependent on data will be helpful for instruction detection.

Sumathi M [31] proposes a system that combines anomaly, misuse and host based detection. Simple Fuzzy rules allow us to construct if-then rules that reflect common ways of describing security attacks. For host and Network based intrusion detection use fuzzy rules and machine learning along with self organizing hash maps.

ZHANG Jian[32] proposes a fuzzy default theory to transform reasoning and response engine of IDS, based on the proving of IDS as non-monotonic, and set up an intelligent IDSFDL-IDS, the experimental result shows that FDL-IDS increased the detection speed and sensitivity and decreased the cumulative cost as compared with traditional intrusion detection expert system.

German Florez [33] proposes a system to detect anomalous behavior, they generate fuzzy association rules from new audit data and compute the similarity with sets mined from "normal" datasets of fuzzy association rules that are mined from network audit data as models of "normal behavior." If the similarity values are below a threshold value, an alarm is issued. They describe an algorithm for computing fuzzy association rules based on Borgelt's prefix trees, modifications to the computation of support and confidence of fuzzy rules, a new method for computing the similarity of two fuzzy rule sets, and feature selection and optimization with genetic algorithms.

c) Genetic Algorithm:

A Genetic Algorithm (GA) is a programming technique used for detection of intrusions that mimics biological evolution as a problem solving strategy. GA uses an evolution and natural selection, recombination and mutation operators. GA uses an evolution and natural selection that uses a chromosome-like data structure and evolve the chromosomes using selection, recombination and mutation operators. Functioning of genetic algorithm starts with randomly generated population of individuals. Thru various generations these population evolved and individual's quality gets improved. In every generation, three basic operators of genetic algorithm i.e. selection, crossover means exchanging the genes between two chromosomes while mutation means random changing of a value of a randomly chosen gene of a chromosome. These individuals are representation of the problem required to be solved. Different positions of each individual can encoded as bits, characters and numbers.

Mohammad Sazzadul Hoque [7] proposes an Intrusion Detection System (IDS), by applying genetic algorithm (GA) to efficiently detect various types of network intrusions. Parameters and evolution processes for GA are discussed in details and implemented. This approach uses evolution theory to information evolution in order to filter the traffic data and thus reduce the complexity. They were using the KDD99 benchmark dataset and obtained reasonable detection rate.

d) Support Vector Machine (SVM):

SVM is a classification and regression based technique for intrusion detection. Some standard support vector machines (SVMs) which are powerful tools for data classification, classifies two-category points by assigning them to one of two disjoint half spaces in either the original input space of the problem for linear classifiers, or in a higher dimensional feature space for nonlinear classifiers. In SVM it performs classification by constructing hyper planes in a multi-dimensional space that separates two classes. SVM tries to achieve maximum separation between the classes.

Rung-Ching Chen [34] presents a combined approach for RST (Rough Set Theory) and SVM (Support Vector Machine) to detect intrusions. Firstly RST is used for preprocessing and reduce dimensions and extracts features then these extracted features are inputs to SVM and then SVM is used learn and test separately.

Snehal A. Mulay[35] proposes for solving multiclass problem using a combinational approach of SVM and Decision Tree algorithms in which it decreases training and testing time and improves the efficiency for intrusion detection.

E.Raju [36] proposes an intrusion detection approach using SVM in which they applies KDD99 data set and the experimental result showed that effectiveness of low frequent attacks such as R2L and U2R attacks compared to neural network based attacks.

e) Decision Trees:

A decision tree is the most popular classification techniques used for detection of intrusion, anomalies and web based attacks. Decision trees are tree like structure which generates the association rules. Decision Trees algorithms are used in machine learning and pattern recognition; there are various decision tree learning algorithms such as ID3, C4.5, and CART [3].

ID3 ALGORITHM

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].

The following two terms used in ID3 algorithm:

i) Entropy

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Suppose S is a set consist of s data samples and has m different class attributes. This defines m different classes: Ci(i=1,2,…,m). Suppose Si is the number of data samples of the class Ci. The quantity of information of the given sample classification is given by the above formula. where, Pi is the probability of the sample belonging to class Ci, and it's value equals to the estimate value of Si/S. Suppose A attribute has v distinct values {a1,a2,...,av}. S would be divided into v subsets {s1s2... sv} according to attribute A. Sj contains data sample of collection S under A's value being equal to aj's. Suppose Sij is the number of samples in subset Sj that belong to class Ci. If we choose A as the test attribute, the information entropy to divide the current sample set is as following.

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + \ldots + s_{mj}}{s} I(s_{1j}, s_{2j}, \ldots, s_{mj})$$

The smaller value of E(A), the better partitioning result. For a given subset Sj, its expect information is as following.

$$I(S_{1j}, S_{2j}, \cdots, S_{mj}) = -\sum_{i=1}^{m} P_{ij} \log_2 (P_{ij})$$

The larger the data sets the less is the entropy. where *i p* is the proportion/probability of S belonging to class *i*. Logarithm is base 2 because entropy is a measure of the expected encoding length measured in bits.

ii) Information Gain

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

Gain (A) calculates the difference between the two former, namely, information acquisition with the highest information gain of attribute which is chosen to determine the attributes, or the property with the minimum information entropy attribute is chosen according to the minimum entropy principle. The larger value of Gain(A), the smaller uncertainty of classification. Calculate the information gain of each attribute, and select the attribute with the highest information gain as the test property of given set S. Create a node and mark it with the property, and then recursively create branches for each property and divide the sample. In the end, it will construct a decision tree.

Algorithm:

The algorithmic steps are as follows:

- Create a root node for the tree
- If all examples are positive, Return the single-node tree Root, with label = +.
- If all examples are negative, Return the single-node tree Root, with label = -.
- If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples.
- Else
    - A = The Attribute that best classifies examples.
    - Decision Tree attribute for Root = A.
    - For each possible value, *vi*, of A,
        - Add a new tree branch below Root, corresponding to the test A = *vi*.
        - Let Examples(*vi*), be the subset of examples that have the value *vi* for A
        - If Examples(*vi*) is empty
            - Then below this new branch add a leaf node with label = most common target value in the examples
        - Else below this new branch add the subtree ID3 (Examples(*vi*), Target_Attribute, Attributes – {A})
- End
- Return Root

C. Clustering Techniques:

Clustering is the technique in which the data sets are divided into groups of similar objects. It is an unsupervised learning approach. Clustering technique is plays an important role in detection of intrusions, in which the malicious data are group into one cluster and normal data are divided in to another cluster.

There are two methods of clustering i.e. partitioning clustering and hierarchical clustering methods.

a) Partitioning Methods :

In partitioning clustering technique it partition the data sets into number of individual clusters and then evaluate them by some criteria.

K-means Clustering Algorithm:

In k-means clustering methods, each cluster is represented by the centre of the cluster.

The k-means algorithm is an algorithm to cluster ‘n’ objects based on attributes in to ‘k’ partitions where k<n.

An algorithm for partitioning ‘n’ data points in to ‘k’ disjoint subsets ‘Sj’ containing data points so as to minimize the sum of squares criterian.

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} \left| x_n - \mu_j \right|^2,$$

Where x_n is vector representing the n th data point and u_j is the geometric centroid of the data points in Sj.

The basic algorithmic steps are as follows:

1. Select k points as the initial centroids.
2. Repeat
3. Form k clusters by assigning all point to the closest centroid.
4. Recompute the centroid of each cluster.
5. Until the centroid don't change.

Yasser Yasami[19] presents a novel method for combining the k-means algorithm and ID3 algorithm for intrusion detection for that they applies the ARP data, they uses the k-means algorithm first to make clustering using Euclidian distance similarity. Then they applies the ID3 algorithm on each cluster, they applies threshold rules for decision making on test instance normally or abnormally.

M. Varaprasad Rao [37] proposes work on network intrusion data and algorithm applies on the KDD99 data set. Before applying directly on k-means, they first applying preprocessing and normalization steps and the experimental results are showed satisfactory.

Sharma, Sanjay Kumar [38] proposes a method for anomaly based intrusion detection using k-means algorithm via naïve based classification algorithm, the proposed technique is effective for intrusion detection rate when applied on KDD cup'99 data set compared to naïve based approach.

b) Hierarchical Clustering Method:

In data mining hierarchical clustering is a method of cluster analysis which makes the hierarchy of clusters. There are two types of clustering i.e. Agglomerative and Divisive, in agglomerative technique is called bottom up approach in which observation starts at its own cluster and then pairs of clusters are merge as it moves up the hierarchy. In case of divisive clustering, it is top down approach in which all observation starts in one cluster and splits are performed recursively as one move down to the hierarchy. Hierarchical clustering tool is widely used for detection of web based attacks.

There are various hierarchical clustering algorithms i.e., BIRCH, ROCK, Chameleon, DBSCAN, OPTICS, DENCLUE, CLIQUE, and PROCLUS.

Comparative chart of Different Data Mining Techniques for detection of web based attacks:

Parameters	**Support Vector Machine**	**Neural Networks**	**Clustering Techniques**	**Decision Trees**
Support for multi-classification	Binary Classifier	Naturally extended	Very Good	Excellent
Accuracy	Good	Very Good	Very Good	Excellent
Speed of Learning	Average	Average	Very Good	Very Good
Tolerance to missing values	Good	Average	Excellent	Very Good
Speed of Classification	Excellent	Excellent	Excellent	Excellent
Tolerance to redundant attributes	Very Good	Good	Very Good	Good
Tolerance to irrelevance attributes	Excellent	Average	Very Good	Very Good
Tolerance to highly dependent attributes	Very Good	Very Good	Good	Good
Dealing with desecrate, continues, binary attributes	Only Continues, Binary	Only Continues, Binary	All	All
Tolerance to noise	Good	Good	Excellent	Very Good
Over fitting problems	Good	Average	Very Good	Good
Incremental learning	Good	Very Good	Good	Good
Explanation ability/classification/ transparency of knowledge	Average	Average	Excellent	Excellent

IV) CONCLUSION AND FUTURE WORK

With the development of computer networks and the invention of e-commerce technologies, now a days we are depends on the web based applications, so naturally there is a wide chances of attacks at web applications so it is the necessity of time to detect the web based attacks. Data mining is efficient techniques for detection of web based attacks, there are various techniques are available; each technique has some advantages and disadvantages. In this paper we are presenting survey for various data mining techniques such as classification and clustering techniques. For detection of

intrusions and web based attacks of and every activities such as ID3 is an classification algorithm used for constructing a decision tree, it is a predictive model to predict a unknown classifier, it is a greedy top down approach, there are many applications are proposed based on ID3. ID3 algorithm has some drawbacks as it produced decision trees with large no of anomalies or useless data due to which the decision tree built with many branches, and the information measure gain measure tends to prefer attributes with many values. Also once the decision tree is built it may contain useless rules. To overcome all these drawbacks of ID3 algorithm, it is the need of time to make improvement in ID3 algorithm, with the increasing use of network and database technologies various intrusion, anomalies attacks are possible now a days. With the use of internet web servers are popular targets of the attacks with HTTP request at web server. Hence further research is needed to use the improved ID3 algorithm for detection of web based attacks such as SQL attacks, anomalies, intrusions.

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"

[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] Jun-Hui Liu, "Optimized ID3 algorithm based on attribute importance and convex function", IT in Medicine and Education (ITME), 2011 International Symposium on (Volume:2)

[22] Kilian Q.Weinberger, Lawrence K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification",Journal of Machine Learning Research 10 (2009) 207-244 Submitted 12/07; Revised 9/08; Published 2/09

[23] Wai-Ho Au, Member, IEEE, Keith C. C. Chan, Andrew K. C. Wong, Fellow, IEEE,and Yang Wang, Member, IEEE ,"Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data", Manuscript received Sep. 15, 2004.

[24] Pave1 Brazdil, "A Method of Processing Unknown Attribute Values by ID3", 1992 IEEE.

[25] Lai-Cheng cao, "Detecting Web Based attacks by Machine Learning", Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006 IEEE.

[26] Hongwu LUO, "An improved ID3 algorithm based on attribute importance-weighted", 2010 IEEE.

[27] Chen Jin, Luo De-lin, "An Improved ID3 Decision Tree Algorithm", Proceedings of 2009 4th International Conference on Computer Science & Education, 2009 IEEE.

[28] ZHANG Jian," Intrusion Detection System based on Fuzzy Default Logic", The IEEE International Conference on Fuzzy Systems, 2003
[29] Guangqun Zhai, Research and Improvement on ID3 Algorithm inIntrusion Detection System", 2010 Sixth International Conference on Natural Computation (ICNC 2010), 2010 IEEE.
[30] Ya-Li Ding, "A novel signature searching for Intrusion Detection System using data mining", Machine Learning and Cybernetics, 2009 International Conference IEEE
[31] Sumathi M, "An Efficient Intrusion detection system for Network behaviors using Fuzzy logic based Rules", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8,August 2013
[32] Zhang Jian"Intrusion Detection System based on Fuzzy Default Logic" The IEEE International Conference on Fuzzy Systems, 2003 IEEE.
[33] German Florez," An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection"
[34] Rung-Ching Chen,"Using Rough set and support vector machine for network intrusion detection", International Journal of Network Security & Its Applications (IJNSA),Vol 1, No 1, April 2009
[35] Snehal A. Mulay,"Intrusion Detection System using Support Vector Machine and Decision Tree", International Journal of Computer Application (0975-8887), Voulume-3-No-3, June-2010.
[36] E.Raju, "Network Intrusion Detection using Support Vector Machines" International Journal of Computer Science and Management Research, Vol 2 Issue 1 January 2013
[37] M. Varaprasad Rao, "Algorithm for Clustering with Intrusion Detection using Modified and Hashed k-means Algorithm", Advances in computer science and Engg & Appl, AISC 167
[38] 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

