ID 3 Algorithm for Intrusion Detection

Anant R. More **Vikas N. Nandgaonkar** **Dr. Manoj Nagmode** **Pramod P. Patil**

***Abstract*- ID3 algorithm is a standard, popular, and simple method for data classification and decision tree creation. Since privacy-preserving data mining should be taken into consideration, several secure multi-party computation protocols have been presented based on this technique. Many times the system get vulnerable to new attacks. This paper is the implementation of proposed model described in [1].We have implemented a model to embed primitive intelligence in the network intrusion detection systems using C# . This model is based on Quinlain's ID3 algorithm of decision tree construction and inductive learning. This model detects unknown attacks with the help of optimized decision tree from available set of data also follow predefined rules for accurate decision making for system Administrator. The communication overhead has been kept reasonably low to make the whole protocol efficient and practical.**

***Keywords* : Network Security, Cryptography, network intrusion detection**

I. INTRODUCTION

Nowadays, many data mining systems are dealing with distributed database among two or more parties, while each party wants to keep its own information private. This is the case in applications in various environments such as medical and insurance. One popular technique in data mining to classify the information is ID3(Iterative Dichotomizer 3) algorithm by which a decision tree is produced from existing data. There is one main formula for Entropy , which can be used in ID3 algorithm to select the attribute with the best information-gain value at each step of this process. Although, according to the surveys for splitting criteria, such as [4], the results of using Entropy and other protocols are very similar, almost all existing protocols use Entropy to compute information-gain to find the best split at each node. Entropy normally tries to create balanced tree. Thus, in distributed computation of the decision tree, where communication cost is the most important issue, we can use the one with better performance, regardless of the negligible difference in their final decision tree. Also, some applicants prefer to test their database with different types of existing techniques and select the best one depending on the final result and their needs for specific problems. Therefore, different protocols, using various techniques are needed to be proposed in this field of study. In this paper, we introduce a secure solution for the ID3 algorithm in which Entropy is used to compute information gains for the remaining attributes in the current node of the decision tree. We present a multi-party protocol to securely compute each expression of the formula obtained and three secure multi-party sub-protocols for addition , duplication and square division . First one generates private output shares for involved parties such that their multiplication becomes equal to the addition of the input shares. Second one generates private output shares such that their addition becomes equal to the multiplication of the input shares. Third one, square division, computes a sub-formula of the Information Gain formula.

II. BEHAVIOR OF RELATED ANALYSIS SYSTEM

Learning model is the classification model according to the subjective will and the objective ability in the process of students learning [1]. Constructing learning model is not a simple classification, but must classify and synthetically evaluate learner's learning ability, learning mode and motivation firstly. In order for the correct analysis and evaluation of learner's learning model, we have devised a network learning behavior intelligent analysis system to collect data and mine data, then realize the classification and evaluation finally, as shown in figure 1. The system composed of Antibody database, Intrusion Detection module, and data packet capture module. The data collection module is mostly used for the collection and quantification of learning ability, behavior, strategies and tendencies, which impact learner's learning.

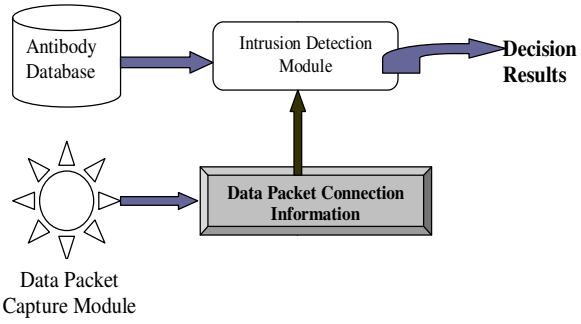


Fig.1 Behavior of Intrusion detection system

Statistical analysis done by Intrusion Detection module embedded process the collected data by data mining technology, and explores the personality characteristics. The results of statistical analysis can be the basis of classification and evaluation module, and provide the foundation for the

establishment of the learner characteristics results model. Classification and evaluation module is further mining classification of the history data of network learning behavior, and analyze and evaluate the relationship between personal learning behavior and learning effect. In terms of feature selection, several researches have proposed identifying important intrusion features through wrapper and filter approaches. Wrapper method exploits a machine learning algorithm to evaluate the goodness of features or feature set. It provides better performance of selecting suitable features .

III. THE IMPLEMENTED MODEL IN C

Most of the contemporary Intrusion Detection Systems are composed of four parts. A packet capturing mechanism, a classifier, a database of known attack patterns and an optional user interface. The packet capturing mechanism captures network traffic from an identified network segment and pass it on to the classifier. Figure 2 shows our implemented system having one module of classifier. The classifier

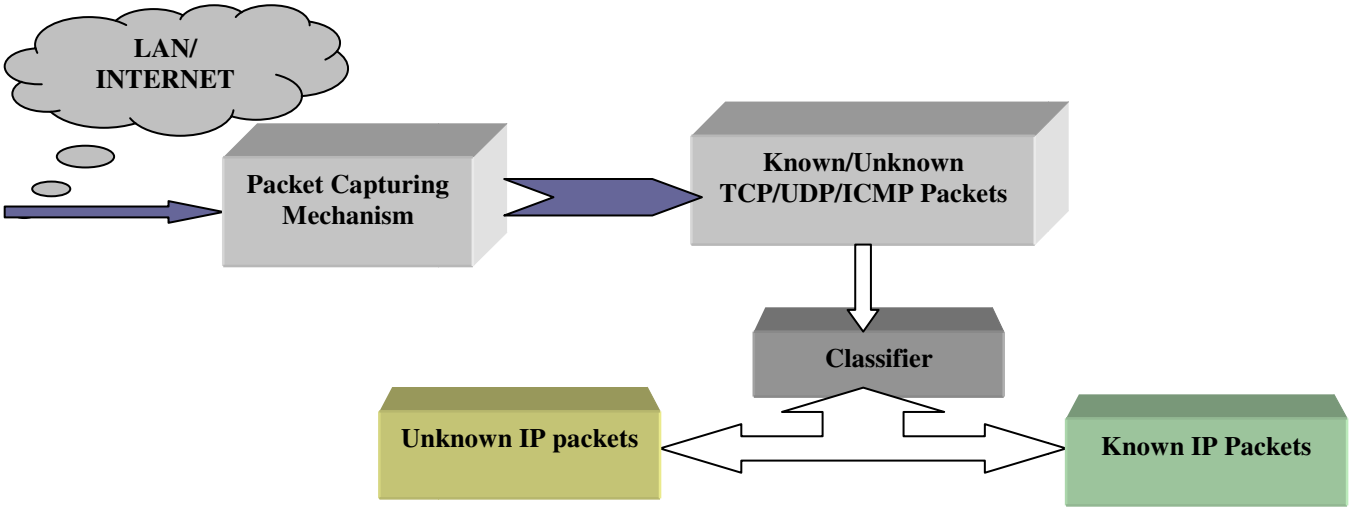


Fig. 2 Contemporary IDS with Classifier

classifies the incoming traffic as "innocent" or Suspicious" based on the results of comparison with attack Patterns already there in the database. The result of this comparison is either logged for off line analysis or is displayed by the user interface. Location of a classifier in a conventional IDS can be seen in figure2 . The classifier performs the job of classification of income network traffic into either innocent or suspicious traffic. If the packet arrived is classified as innocent no logging is done. If on the other hand, the packet is classified as suspicious, then the event has to be logged and/or the user is to be informed. This information to the user or system administrator can either be a mere beep or the system (NIDS) can be designed to send an e-mail to system administrator (through a suitably programmed gateway) about the occurrence of a particular event The classifier of most rule based NIDSs work on the principle of patter matching. If the packet or a sequence thereof exhibits a particular behavior already in the database of the classifier it will be classified according to decision attribute given in the database. If the system exploiter is using a new technique or if that particular behavior is not already programmed, the packet will be classified as "Innocent". For a rule base based system it is termed as firing or not firing or rule. No system has so far been developed to take inference from the available data and subsequently add a new rule to classify an unknown sequence of packets. Figure 3.gives data flow model for implemented system where Analyzer user interface has input from packet capture mechanism , it take decision whether packets from known or legal address and has valid or non destructive packets . Two decisions taken "SUCCESS" and "FAILURE" .

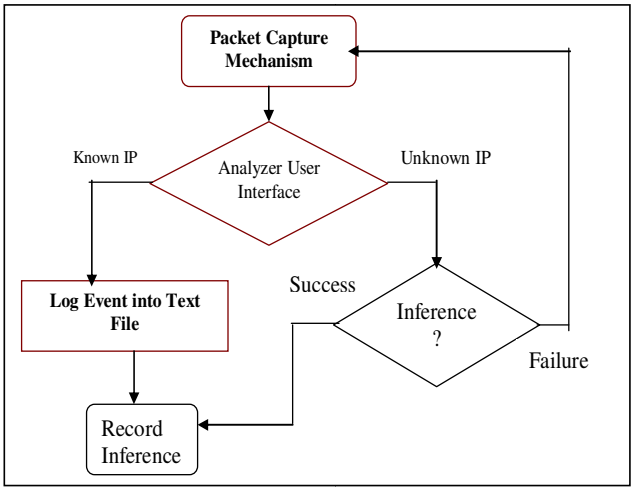


Fig. 3 Analyzer user Interface to Contemporary ID3

The Quinlain's ID3 algorithm makes extensive "use of decision trees. A decision tree is a tree in which each non leaf node is labeled with an attribute or a question of some sort, and in which the branches at that node correspond to the possible values of the attribute, or answers to the question. For example, if the attribute was shape, then there would be branches below that node for the possible values of shape, say square, round and triangular. Leaf nodes are labeled with a class. Decision trees are used for classifying instances one starts at the root of the tree and taking appropriate branches according to the attribute or question asked about at each branch node, one eventually comes to a leaf node. The label on that leaf node is the class for that instance. As I said, each node of a decision tree is linked to a set of possible solutions. Each parent node, that is each node that is not a leaf (and thus has children) is associated with a test, which splits the set of possible answers into subsets representing every possibility of the test's outcomes. In a application such as this one, a special form; of decision trees is used that is known as the I1dentification Tree," or simply an ID tree. An ID tree is nothing but a decision tree in which all possible divisions are created by training the tree against a list of known data. The purpose of an ID tree is to take a set of sample data, classify the data and construct a series of test to classify an unknown object based on like properties.

IV. ID3 ALGORITHM

ID3 algorithm is a typical decision tree algorithm. It analyzes known types of objects according to a fixed set of attribute or characteristic, and produces a decision tree, and then the decision tree put all the objects in the correct classification [2]. It uses the concept of mutual information when choosing important characteristics, forms the decision tree by using the subset of training examples, and excerpts mutual information as the discriminance. First, find out factors that have best sense, and divide the data into several subsets, and then each subset can be divided by the factors that have best sense, till all subset contains the same type of data, thereby result in a decision tree. ID3 algorithm has clear theory, simple technique and strong learning ability; it is suitable for processing mass resources distribution issues. But ID3 algorithm has its drawbacks: the calculation of mutual information depends on the characteristics that have much eigenvalue ,there is a hypothesis if we set the mutual information as a feature selection method. That is the proportion between positive examples and negative examples in training example subsets should be the same with the proportion in the real problems. But it cannot be guaranteed the same under normal circumstances, and there is deviation when calculate the mutual information in training set. _ ID3 algorithm is sensitive to the noise (the errors in training sets). _ ID3 decision tree will be changed along with the increasing of training sets. And it is inconvenient to the growing of training examples.

A. Role of Entropy in ID 3

Traditional ID3 algorithm chooses attributes, and often tend to choosing the attributes that get more values, because the weighted sum method makes the classification of examples set tend to the metadata group that discarding small data group, but the attribute has more properties is not always optimal one. The attributes in the learning model building process include the knowledge level of originally subject in learning ability database, the multiple factors of learning mode in learning mode database, and the learning motivation classification in learning motivation database. The final decision tree classification results are not certainly consistent with the actual situation according to the traditional ID3 classification because there are many types of attributes based on Entropy.

B Introduce The User Interest α

In the decision tree established by increasing user interest, the information entropy corresponded to the root node is the largest. Along with the construction of the decision tree, information entropy gradually decreased until the entropy of leaf nodes turn to zero (i.e. all objects of a node in the same category). Therefore, it is hoped that each choice of testing attribute can reduce the entropy at the presto speed, and then make every branch of the decision tree as short as possible and eventually build a smaller tree. It is the purpose of ID3. The traditional ID3 algorithm does not take into account the influence of the relationship between attributes on the attributes choosing, and results in the choice of redundant attributes that have little meaning or no significance to the real classification. Algorithm demands the maximal relationship between the selected attribute and the genus (i.e. the information gain in ID3 algorithm), and the minimal relationship with the used attributes in the same branch (interactive information) [3]. This will avoid the choice of redundancy attributes, and accelerate the pace of entropy reducing and thus build a better tree. In order to distinguish attributes importance, we introduce the user interest when calculating the information entropy to distinguish the dependence of attributes. The user interest $\alpha(0\leq\alpha\leq1)$ to be known as the user interest to uncertain knowledge, and it is determined by the decision-makers according to the prior knowledge or area knowledge. It is a vague concept, usually referred to certain prior knowledge, including area knowledge and expert advice. And in the study of decision tree, it is referred to the factors that influence the generation

and selection of the decision tree rules except the examples set used for the formation and modification of the decision tree in its training process. Suppose that a training examples set is X, the purpose is to divide the training examples into n classes, recorded as C=(X1, X2, …Xn). On the assumption that the number of i^{th} training examples is | Xi | = Ci, the probability that an example belongs to this training examples is P (Xi). If we choose the attribute A to test, with a set of properties *a*1,*a*2,*a*3,...*ai*, the number of examples that belonged to the i^{th} category when $A = a\,j$ is Cij

P(Xi : A = aj) = Cij / | X | (1)

The value of $P(Xi{:}A{=}aj)$ is the probability that the test attribute A belongs to the i^{th} category. *Yj*is the examples set when A= *aj*, then the degree of uncertainty to the decision tree classification is the entropy of the training examples set to attributes A:

$H(Yj)= -\sum P(Xi|A{=}aj)\log 2P(Xi|A{=}aj)$ (2)

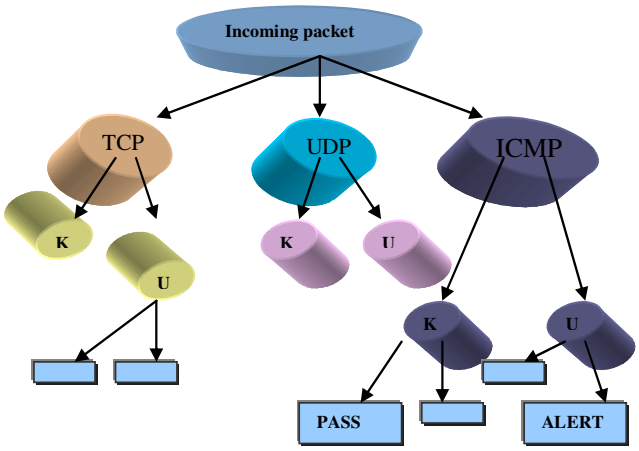
We increase the user interest αwhen calculating the taxonomic information entropy of each leaf node $X\,j$ when $A = aj$extended from attribute A, and then strengthen the label of important attribute, and reduce the label of non-important attribute. The formula as follows:

$H(X|A)=\Sigma[P(A{=}aj)+\alpha]H(Xj)$ (3)

The information provided by attributes A for classification (the information gain of attribute A) is:

$I(X{:}A)=H(X)-H(X|A)$ (4)

V. Formation of Optimized Decision Tree



U:Unknown , K : Known

Fig . 4. Optimized Decision tree

Calculate entropy (disorder) for each of the six attributes (protocol, sender's IP address, Sender's port number, data contents or packet payload and data size). Using formulae mentioned in section 4.1 Sender's IP address can be selected for first test at the root -node, A sample test can be "'is sender's IP address from a known set of IP addresses". An other test can be whether this packet is from a machine which is "'internal" as per the network hierarchy. The attribute "protocol" is selected next because it minimizes the entropy in the known IP addresses subset. Similar calculations on the other subsets and sub-subsets will generate following optimized decision tree which is capable of identification of all the samples in the training set.

Estimating to classification accurate rate of decision tree is very important, this we can estimate a given classification accurate rate to unknown correct data indicia will do[4]. Holdout and k-fold cross-validation is two kinds of technology which estimate accurate rate of classification. In holdout, given data are separated by two sets randomly: training set and testing set. Usually, two thirds data are assigned to the collection training, by which the classification is induced and its accurate rate is estimated through testing set. Random sample, transformation of holdout, is k repetition of holdout, whose global accurate rate is the average value of all iteration accurate rate. In k-order cross-validation, the primary data is divided into k disjoint subsets ,and each of them is same in size. Training and testing are repeatedly operated k times, in i^{th} iteration, is used to testing set, and the other is used to training set. That is, the classification in the first iteration is trained on and tested on; the classification in the second iteration is trained on and tested on; the process continues until the end. The accuracy estimation is the value of right classification in k iteration dividing the sum of primary data. In stratified cross-validation, all the subsets are stratified, which make the distribution of sample in each subset be same with that in primary data. The other methods for estimating classification accuracy include bootstrapping and leave-one-out. In the previous, we select identical and recovery sample; while the latter is special case of k-order cross-validation, and s is the sum of primary sample. Recently, the most common method is 10- stratified cross-validation because of the relatively lower bias and variance. Generally speaking, this kind of method is suggested to use.

A. IF-ELSE Rules for formation of Decision Tree

Rule 1

If packet Destination Address= Unknown
If Protocol.Type="TCP"
If Packet.Destination_Port(Known)=23
Alert="YES"
If Packet.Destination_Port(Unknown)!=23
Alert="YES"

Rule 2 If Protocol.Type="UDP"
If Packet.Destination_Port(Known)=23
Alert="YES"
If Packet.Destination_Port(Unknown)!=23
Alert="YES"

Rule 3 If packet Destination Address= known

If Protocol.Type="TCP"
If Packet.Destination_Port(Known)=23
Alert="NO"
If Packet.Destination_Port(Unknown)!=23
Alert="YES"

Rule 4

If Protocol.Type="UDP"
If Packet.Destination_Port(Known)=23
Alert="NO"
If Packet.Destination_Port(Unknown)!=23
Alert="YES"

VI. PERFORMANCE ANALYSIS

In Figure 4. we show how decision tree is generated using above mentioned rules also examples abstracted from sample bank by ID3 algorithm, and get the following result showing all information into Analyzer User Interface build in C#.NET.
Figure 5 shows source IP with port information, destination IP with port number which it accessed, packet type whether packet is TCP, UDP or ICMP with known and unknown attributes. At the end we calculate how many packets are of TCP type or UDP type or ICMP type so that it can be beneficial for the Administrator to take proper decision to protects the network.

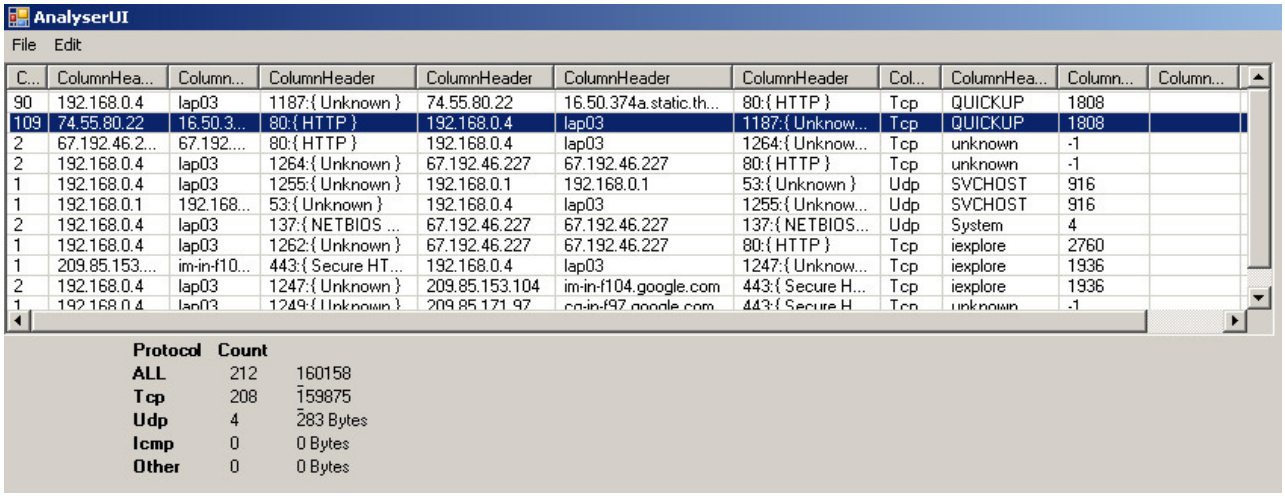


Fig. 5 Result of the Implemented model with Summary of Protocol and Count

CONCLUSION

In this paper, based on ID3 algorithm, by introducing the concept of entropy and IF-ELSE rules we have optimized decision tree . Above system gives better performance on heavy networks . Using this algorithm, we can abstract high reliability rules by removing unknown packets from incoming packets and generate decision tree with good structure and high reliability. The experiments indicate that, the accuracy of decision tree is improved, from which we can abstract good rules. In this test, the accuracy rate of rules is not very high, which is relevant to small data bank. The larger scale the data bank is, the more the useful data digging from the bank, the higher the accuracy rate of rules is, and the efficiency and performance of the algorithm will get better and the superiority of algorithm will be obvious.

REFERENCES

[1] Saeed Akhtar, A Proposed Model To Use ID3 Algorithm In The Classifier of A Network Intrusion Detection System
[2] U.M. Fayyad, K.B.Iran. Mlti-interval Discretization of Continuous-valued Attributes for Classification Learning. Proceedings of the 13th International Joint Conference on Artificial Intelligence. Morgan Kaufmann: In R. Bajcsy(Ed.), 1993: 1022-1027
[3] D. Michie. Personal Models of Rationality. Journal of Statistical Planning and Inference, 1990, 23(25): 381-399
[4] Steven E Smaha. Haystack: An Intrusion Detection system. In Fourth Aerospace Computer Security Applications Conference, pages 37-44, Tracor Applied Science Inc., Austin, Texasdecember 1988.
[5] Teresa F Lunt. Intruders in Computer Conference on Auditing and Technology,1993.Detecting systems. Computer Technology, 1993.

AUTHOR'S PROFILE

	**Prof. More Anant** Asst. Prof. & HOD NMIET, Pune , BE, ME Electronics Engg., LMISTE Anant_anu@yahoo.com
	**Prof. Nandgaonkar V. N.** Asst. Prof & HOD. NMIET Pune , B.E., M.E. Computer Engg., LMISTE, vikas.nandgaonkar@gmail.com
	**Author's Name : Prof. Patil Pramod,** Lecturer, NMIET, Pune pramodpatil88@gmail.com
	Dr. Manoj Nagmode Prof. MIT COE, Pune , LMISTE