International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 5 1294– 1298

Decision Tree Based Intrusion Detection System with Sampling

Pankaj R. Ambartani
School of Information Technology, VIT University
Vellore, India
pankajambartani@gmail.com

Bhushan B. Shinde
School of Information Technology, VIT University
Vellore, India
bhushan.shine@yahoo.com

Prof. Usha Devi G
School of Information Technology, VIT University
Vellore, India
ushadevi.g@vit.ac.in

***Abstract*—** The level of sophistication of the cyber threats has increased the network defenders are bound to use every resource possible in their arsenal to protect the networks. The most important objective for any security management dealing with a huge high speed network is to find any unexpected variation in the traffic pattern as a result of different attacks. Intrusion Detection System (IDS) has become the integral part of today's Information security infrastructures. Intrusion detection system is used to detect unauthorized access and manipulation of the computer system. This paper uses the decision tree algorithm with stratified weighted sampling to improve the accuracy rate of intrusion detection. We proposed learning with preprocessing step i.e. to generate the samples using the stratified sampling technique to surcharge the shortcomings of the ID3 algorithm. KDD Cup99 dataset is used for the purpose of experimentation and further evaluation.

Keywords- *Intrusion Detection; Decision Tree; KDD cup 99 Dataeet; Pre-processing; Startified Weighted Sampling*

*****

I. INTRODUCTION

As the standards and the number of attacks on the computer networks has increased, therefore it has become extremely difficult not only to avoid but also to detect the computer threats APT (Advanced Persistent Threats) has proved itself as a matter of concern not only for the organizations but has alarmed the nation security as well as per the report from Rachwald-2010). Intrusion Detection is often used as another wall to protect computer systems. Intrusion Detection may be explained as "the process to recognize anybody trying to use any computer system which they are not authorized to, also the legitimate access holders exploiting their resources [1].

The intrusion detection can be categorized as- Misuse detection and Anomaly detection. Misuse detection is the capability of a system to identify a known sequence of activities, often referred to as the signatures identified as threats.

The anomaly detection works on the traffic deviation from the expected established pattern of traffic in a network. By the decision tree analysis the logic of decision tree can be implemented in the intrusion detection system. A decision tree is a method of predicting from the figures of statistics which are converted into a tree like structures to form a statement of the pattern. Decision tree can prove itself substancial in the retrieval of data for the purpose of making decisions. The decision tree starts with a root node which splits recursively as per the possible conditions and its decision. Decision tree is effective as in induction for the unseen instances only if those have some relation with the origin of the notion. Sampling refers to drawing a sample or selecting a subset of elements from a population. The usual goal in sampling is to produce a representative sample,

would be a "mirror image" of the population from which it was selected.

II. LITERATURE REVIEW

Intrusion detection begins in 1980's and after that numbers of practical aspects have been made to form intrusion detection systems [2-4]. Panda and Ptra [5] resolute a method to discover signature of definite attacks by naive Bayers. For analyzing the output they used KDD dataset. Meng Jiaiang [6], for clustering and resolving the information employed the K-mean Algorithm. The clustering refers to partition of data into subset of comparable objects. Every subset, refers a cluster, constitute of objects that are comparable done atween themselves and incomparable to objects of remaining subsets.

Jiong Zhang [7] uses the momentum forest algorithm in anomaly based NIDS for the Intrusion Detection. Cuixio Zhang, shanshan Sun, Guobling Zhang [8] uses heterogeneous method for intrusion Detection which patterns the heterogeneous elements combining the anomaly and misuse detection that the anomaly detection system is formed using not constantly observed clustering process and the algorithm is a refined version of K-means Algorithm. This refined algorithm acquires the determined points of K-means enhanced trilateral triangle theorem. Gary Stein [9] put the practical use of genetic algorithm and the decision tree algorithm to detection of intrusion and method for the characteristic reduction.

III. DECISION TREE

Decision tree methodology is a easy, predictive and efficient method which make use of divide and conquer policy in a top down approach making use of the greedy algorithm. The process starts from the root node and precedes onwards each non leaf node. First of all select an attribute to verify the sample set. After that divide the in-process sample set and makes many subsets of samples as per the verification, every sum-sample containing a new leaf node. Continue the same process till we get an appropriate requisite. The ID3 algorithm is suitable for very few records in a sample set and it can't

manage the lost values. When the sample set's size is increased, the decision tree is not adaptable to changes. The best thing about the decision tree is that it provides a level of abstraction preventing the user to avoid knowing much about the off screen procedures. The ID3 algorithm uses the information reduced to observe any change in attribute thus aiding the contribution in the construction of the tree. The ID3 decision tree, quantification of the information is done and the procedure is known as entropy. Entropy is referred to scale the ambiguity in a data set. If all the events of a set belong to the same category, then the entropy is considered to be zero signifying unambiguity.

The following steps are repeated continuously [10]:-
1) Calculation of the information gained for every Attribute.
2) The attribute with the largest amount of gained Information is choosen for the split.
3) If chosen attribute is distinct, all the feasible variables are attatched to it, otherwise if ongoing then a cut point with the maximum amount of information is selected.
4) Thereafter split, check if the node is leaf or not; if not then they are the roots of the sub-tree.
5) Step 1 to 4 is recursively repeated.

Let P1, P2. … Pn be the Probabilities

Where $\sum_{i=1}^{n} P$i=1,

Entropy E is calculated as

$$E(P1, P2 \ldots Pn) = \sum_{i=1}^{n} (Pi . \log (1/Pi)) \quad \ldots\ldots\ldots\ldots (1)$$

$$\text{Information Gain } (A, G) = I(A) - \Sigma p(Ai)\, I(Ai) \quad \ldots\ldots\ldots (2)$$

Where A= Selection of attribute
G= Gain of information
The proposed system utilizes the gained information to make the decision tree.

A. *Proposed Decision Tree Algorithm*

```
PROC Create_tree (I, ATBT)
{
        Create (I);
        IF (the risk factor of all the elements of the data sets
        in the sample is     similar)
        THEN (return L as leaf)
        ELSE
        {
        FOR (every attribute in ATBT)
        {
        IF (the attribute is never utilized in categorizing the
        attributes)
          THEN (gather the Information of the attribute of the
        node)
        }
        IF (The attribute with the maximum information gain
        (>0) is labeled as ABT) THEN
        {
        Label the node as the node needs splitting
        Next step as per the value of ABT
        Divide R into Rk, Split the node R;
        }
        ELSE
        {
        Read node equivalent to leaf;
        }
           FOR (every branch Rk) Create_Tree(R, ATBT);
        }
}
```

Where I=Information,
ATBT=ATTRIBUTE

IV. SAMPLING

Sampling is the process in which we select a Representative Sample which act as the mirror of the whole set of data after the analysis of a part. The main reason behind using sampling is to make the optimized use of the resources available and to obtain an estimation which is as much as possible nearer to the actual values. If the elements of the subset are of the homogeneous nature then the sampled set is more likely to be small and vice-versa. Sampling is universally accepted as the way in which subset so drawn from the whole set quiet efficiently.

A. *Sampling Techniques*

Simple Random Techniques: All the elements have the same probability of making into the sample.

Weighted Sampling: In this type of sampling the probability of each element differs from each other or in other words non-uniform n its chances of inclusion depends upon the predefined parameters[11]. Total available data into areas and the values of the parameters inside strata are expected to be similar.

B. *Stratified Weighted Sampling*

```
Star_Sampling(S, P, Q)
{
F_Estimate =0
For i=1……P
{
If P.W>X
{
Current_Estimation==0;
For j=1……Q
{
Current_Estimate +=S(
Uniform_Range ((i-1)/S, i/S))
}
F_Estimate +=Current_Estimate/Q;
}
}
Return F_Estimate;
}
```

Where X=Weight at any instant of time

V. KDD CUP 1999 DATASET

The KDD dataset is the version of dataset containing only the network information (Tcpdump information). The term *knowledge discovery in Database* or KDD can be defined as

the procedure to find knowingness of information and to stress on the top hierarchical application of any specific data mining methods that aim to mine knowledge from information taking large databases into consideration. The KDD is composed of single connection vectors having 41 characteristics of each with a mark on all of them as normal or attack [12], along precisely a single type of attack. The experimental data reflect from KDD cup Dataset [13] is as follows:

TABLE I. KDD cup 41 feature

#	Feature Name	#	Feature Name
1	Duration	22	is_guest_login
2	protocol_type	23	Count
3	Service	24	srv_count
4	Flag	25	serror_rate
5	src_bytes	26	srv_serror_rate
6	dst_bytes	27	rerror_rate
7	Land	28	srv_rerroe_rate
8	wrong_fragment	29	same_srv_rate
9	Urgent	30	diff_srv_rate
10	Hot	31	Srv_diff_host_Rate
11	num_failed_logins	32	dst_host_count
12	logged_in	33	dst_host_srv_count
13	num_compromided	34	dst_host_same_srv_rate
14	root_shell	35	dst_host_diff_srv_rate
15	su_attempted	36	dst_host_same_src_port_rate
16	num_root	37	dst_host_srv_diff_host_rate
17	num_file_creations	38	dst_host_serror_rate
18	num_shell	39	dst_host_srv_serror_rate
19	num_access_files	40	dst_host_rerror_rate
20	num_outbound_cmds	41	dst_host_srv_rerror_rate
21	is_hot_login		

- Denial of Service-Attacks which are framed such as to block the user having access to certain amount of services.
- Probe-A Probe is exploring the target network with the motive to uncover some information about the network which may be misused by other type of attacks.
- Remote-to-Local-In this type of an attack the hacker has no access to the internal network, thus it try to gain control over the network from outside most probably internet.
- User-to-Root-In this case the hacker has a valid access into the network. The attack is aimed at increasing the penetration inside the network so as to gain access to the actions which were not authorized earlier.

The *content* characteristics proposed by the knowledge harvested from the domain are used to determine the payload of the TCP packet, for e.g. number of log-in attempts which has failed.

Inside the connection, the *same host* characteristic analyzes the acknowledged connections having similar destination host to estimate the figures relating the protocol behavior. The *similar same service* features examine the connections having the similar services as the present connection since the spent two seconds.

VI. PROPOSED APPROACH

A. System Architecture

The suggested architecture for the system is presented in the figure 1. Preprocessing is done to transform the values into its numeric form.

The data derived from KDD CUP"99 could be a sequence of system calls which is textual in nature having forty-one characteristics.

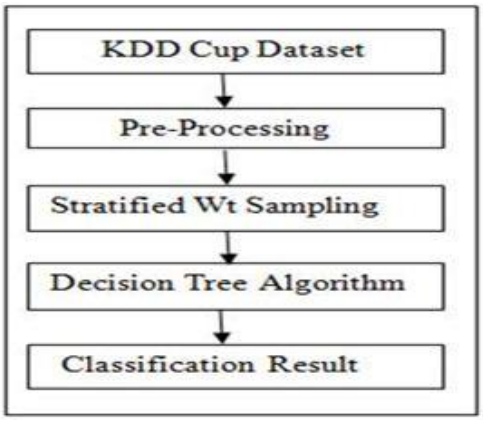

Figure1. Flowchart of Proposed System

B. Analysis of Pre-Processing

Every Intrusion detection pattern has its own pros and cons which inspired us for making a new framework altogether. It's always a better option to join the components usable.

The utmost prior objective of the proposed system is to make a system that avoids the hoax calls and also adds to the betterment of the system. Issues contributing are as follows:

- The ways to furnish a desirable and proportionate data for computation for Intrusion Detection.
- The way to differentiate between the hoax attacks hence improve the detection chances.
- Finding pattern in the patterns of breach and representing them in proper data types so that the manager or executor could lay down proper policies for future reference.

In *Connection Record*, the "Connection" represents stream of packets of data of a particular service protocol, say: Transferring an image via HTTPS protocol.

C. Preprocessing Step to KDD Data

The whole system suggested here is composed of the following components [14] as Preprocessing of data, Fusion Decision stage and Data Callback stage.

In the first stage i.e. preprocessing, data compression is done as much as possible without hampering the amount of information content. Derivation of Attribute-where the derivation of a new attribute id done from a given set of attributes. Deduction of the sample-set: Merge 2 samples if their identity goes beyond a predefined value.

Fusion Decision Phase for identifying the hoax alarm from the actual one in order to elevate the performance of the system. In Data Callback stage the clusters of non-defined data sets into the set of data meant for testing. This methodology ensures consistency of the system.

VII. SIMULATION PROCEDURE

The input to the proposed system where a portion consisting 10% of KDD Cup 99 dataset for testing and training and separate subsets are kept aside for them. Out of which the training subset of data is categorized further into 5 sets to evaluate the response in case of the 4 types of attack sparing one for normal behavior. The data which is to be tested is separated for DoS, U2R, R2L, Probe and normal data. Here we have classified the nature of the user operation as normal and abnormal (attack), where the abnormal behavior is actually the set of the predefined actions already present in the attacks. The main objective of the decision tree is to identify difference between the behaviors. Training set of the data consists of the normal data along with 4 other

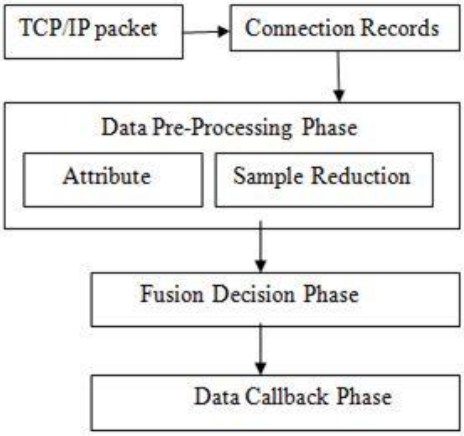

Figure2. Framework of Integrated Decision System

categories of attacks which is fed into the system to recognize the apt qualities. Thereafter by implementing the decision tree, system produces determinate and indeterminate rules, of which tree is created from the definite ones. The testing set of the data is the fed to our system which based on the earlier defined rules decide whether the input is normal or abnormal. The result which is obtained is used thereafter to calculate the encompassing accuracy which depends further based on its capability to categorize the instance correctly.

TABLE II. Instance in Dataset in their Respective classes

Dataset / Attack	10K	20K	30K	40K
Normal	300	650	1000	1400
DoS	260	580	770	1000
U2R	140	220	450	600
R2L	180	330	580	700
Probe	120	320	200	300
Total instance	1000	2000	3000	4000

VIII. RESULT AND ANALYSIS

Our proposed model is expected to deliver the Accuracy Rate of the system by 6% i.e. 94% and reduction in the Error Rate to less than 3%. After implementing the 41 features of the KDD Dataset on the decision tree algorithm The detection rate of correct categorized entity and misclassified entity id given by the following formula which is used to calculate the performance of the system:-

Accuracy Rate =

$$\frac{Total\ number\ of\ correct\ classified\ entity}{Total\ number\ of\ entity} * 100$$

Error Rate =

$$\frac{Total\ number\ of\ miscategorized\ entity}{Total\ number\ of\ entity} * 100$$

Table III shows the accuracy rate of many set of data. Proposed approach has nearly 94% accuracy.

TABLE III. Comparison Accuracy Rate if ID3 and Proposed System

Algorithm	Dataset			
	10K	20K	30K	40K
ID3	88.78%	88.23%	87.86%	86.45%
Proposed	94.74%	94.54%	94.02%	93.85%

Table IV shows the Error rate of proposed approach respectively with many set of Data. Proposed approach has less error rate nearly 3%.

TABLE IV. Comparison Errror Rate of ID3 and Proposed System

Algorithm	Dataset			
	10K	20K	30K	40K
ID3	8.78%	8.24%	9.64%	9.81%
Proposed	2.81%	3.26%	3.18%	3.92%

Figure 3 and Figure 4 shows comparison graphical analysis of Accuracy rate and Error rate of existing Algorithm and Proposed approach.

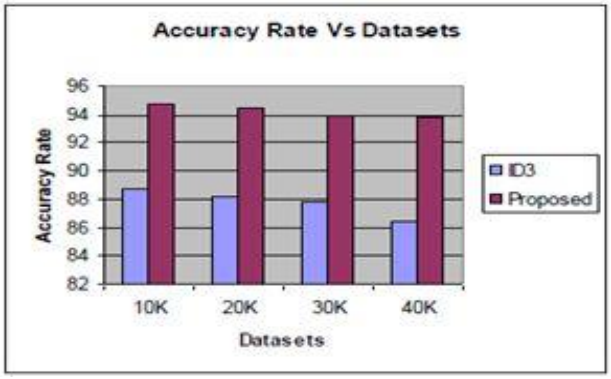

Figure3. Graphical Analysis of Accuracy Rate for Praposed System

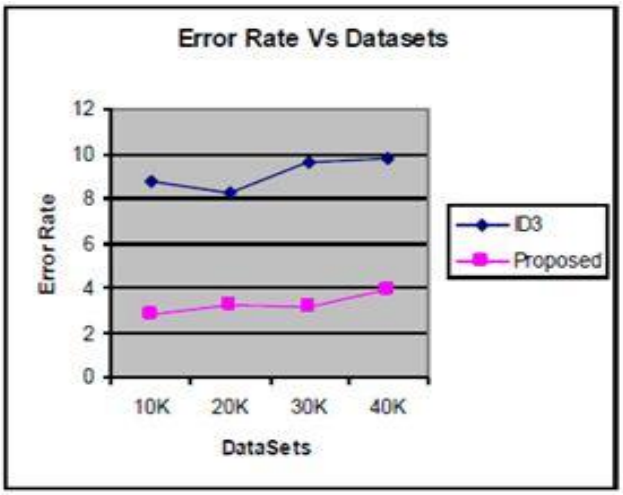

Figure4. Graphical Analysis of Error Rate for Praposed System for Different Dataset

IX. CONCLUSION

After the implementation of the Stratified Sampling and Decision Tree on an Intrusion Detection System, we have been able to classify the breach as attack or normal. Where the decision tree has been implemented to improve the precision of the system, the preprocessing done on the data obtained from KDD in three stages ensured that the performance of the system remains consistent with significant reduction in the rates of the mistakes occurred. Implementation of Decision tree algorithm on the stratified samples proved worthy for coping up from the problems observed in ID3. Thus, the proposed methodology can be implied on the varied sized datasets with a higher level of accuracy than the existing algorithms decreasing the use of the memory and CPU. Hence the new system fulfills the criteria of being reliable for ducting intrusion in the network.

REFERENCES

[1] http://www.webopedia.com/intrusion detection.

[2] Shari Rubin, Somesh Jha and Barton Millers, "Protomatching Network Traffic for High Throughput Network Intrusion Detection" In Proceedings of the Proceedings of the 13th ACM conference on Computer and Communication Security,pages 47-58,ACM,2006.

[3] Marco Cove, Divide balzarotti, Viktoria Felmetsger and Giovanni Vigna Swaddler,"An approach for the Anomaly-Based Detection" Symposium on Rcent advances in Intrusion Detection (RAID), pages 63-86.Springer, 2007.

[4] Pavel Kachurka,Vladimir Golovka , "Neutral Approach into Real-Time Network Intrusion and Recognision" The 6th IEEE International Conference on Intelligence Data Acquisition and Advance Computing System: Technology and Application,15-17 September 2011,pp.393-397,IEEE 2011.

[5] M.Panda and M.R Ptra,"Network Intrusion Detection using naïve Bayes" International Journal of Computer Science and Network Security (IJCSNS), Volume-7,No.12,December 2007, PP.258-263.

[6] Meng Jianliang, Shang Haikun, "The application on intrusion detection based on K-means cluster Algorithm", International Forum on Information Technology and Application, 2009.

[7] Jiong Zhang and Muhammas Zulkernine, "Anomaly based Network Intrusion detection with unsupervised outlier detection," School of Computing Queen's University Kingston, Ontario, Canada.IEEE International Conference ICC 2006, Volume-9, PP.2388-2393, 11-15 June 2006

[8] Cuixiao Zhang, Gouging Zhang, Shanshan Sen, "A mixed unsupervised clustering based Intrusion detection model" third International Conference on Generic and Evolutionary Computing, 2009.

[9] Gary Stein,Bing Chen,"Decision Tree Classifier for network intrusion detection with GA based feature selection," University of Central Florida,ACM-SE 43,Proceedings of 43rd annual Southeast regional Conference,Volume-2,2005.ACM,New York,USA

[10] Mohammadreza Ektefa, Sara Memar, Fatimah sidi , Lilly Suriani Affendey, " Intrusion Detection using Data Mining Technique" IEEE 2010

[11] Saar-Tsechansky, M. and F. Provost, "Active Sampling for Class Probability Estimation and Ranking", Machine Learning 54:2 2004.

[12] R. Shanmugavadivu, Dr. N. Nagrajan, "Network Intrusion Detection System Using fuzzy Logic" Indian Journal of computer Science and Engineering.

[13] http://kdd.ics.uci.edu/database/kddcup99/kddcup99.ml.

Wang Ling, Xiao Haijun, "An Integrated Decision System for Intrusion Detection " 2009 International conference on Multimedia Information Networking and Security.