INTERNATIONAL JOURNAL OF RESEARCH IN TECHNOLOGY AND MANAGEMENT (IJRTM)
ISSN 2454-6240
www.ijrtm.com

ADVANCING NETWORK SECURITY IN ICFRE DATA CENTRE

Pashupati Bhatt
Sr. Lecturer, ITM Dehradun, India,
ppbhatt12@gmail.com

Chandradeep Bhatt
Lecturer, KIT, Dehradun, India
bhattchandradeep@gmail.com

Km. Neha
Lecturer, ITM, Dehradun, India,
nehagodiyal11@gmail.com

ABSTRACT

Network security is of immense priority today for organizations large and small. It is also important for independent internet users to understand the basics of securing their data over the internet and prevent hacks into their information systems. The process of securing your network should be simple to install and effective in implementation. It should protect the network from threats which are known and unknown. Unauthorized access to the resources is a major issue for government organizations. It is important to ensure that no outsider is able to corrupt or impede the network's services. The unauthorized access to resources should be managed as such that it does not thwart employees from being able to avail resources for productive work. Thus the network security topology should be such that it helps prevent attack from an outsider as well as allow users to have adequate access to resources. Data classification is a two-step process. The first step is learning and the second step is classifying the provided data into meaningful sets or classes. Two types of approaches can be applied for the first step of classification: Supervised learning i.e. the use of classifiers or unsupervised learning i.e. cluster analysis. Unsupervised learning is required when the class labels of data used for training is not known. As the class labels in our training dataset are known, we shall use the supervised learning approach.

Key words: Network Security, Data Mining, Classification, Weka Tool

I. INTRODUCTION

Indian Council for Forest Research and Education is the apex of nation-wide research institutes under the Government of India. It is the hub of forest managers and dedicated researchers working together on issues like climate change, conserving the bio diversity, desertification and sustainable management. The council has nine regional centres and four research centres in different parts of the country. The regional research Institutes are located at Dehradun, Coimbatore, Bangalore, Jabalpur, Jorhat, Jodhpur, Shimla, Ranchi and Hyderabad and the centres are at Allahabad, Chhindwara, Aizawl and Agartala. All these centres are connected to ICFRE's datacenter for the following purposes:

1. Network Support
2. Maintaining e-governance through its facilities.
3. Providing a video conferencing centre.
4. Design and development of ICFRE website and other websites regarding national seminars and conferences.
5. Maintaining hardware and database backup.

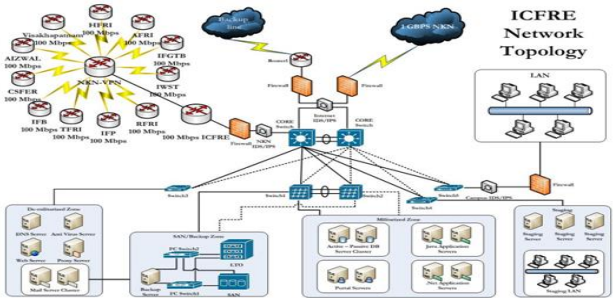
1.1 Devices Currently In Use

i. **Network Intrusion Prevention System (NIPS)**
IPS is a network security device that examines network traffic and keeps a check on malicious activity in the network. Its main functions are to identify intrusion, log information about intrusion, report intrusion and attempt to block /stop intrusion.
   a. IBM Proventia GX 5008 for NKN Internet
   b. IBM Proventia GX 5108 for NKN VPN
   c. IBM Proventia GX 5108 for ICFRE/FRI LAN

ii. **Firewall**
Firewall filters traffic which is inbound or outbound. Firewall can filter packets based on rules. They also filter traffic of packets at network layer. It can also see the legitimacy of sessions. A firewall also has the facility to evaluate content of packets at application layer.
   a. CISCO Adaptive Security Appliance 5550 for NKN Internet
   b. CISCO Adaptive Security Appliance 5550 for Internet Backup
   c. CISCO Adaptive Security Appliance 5550 for NKN VPN
   d. CISCO Adaptive Security Appliance 5550 for Campus

iii. **Vulnerability Assessment (VA) Scanner** IBM ES750
iv. **Anti Virus Server** Symantec Endpoint Protection Manager
v. **Anti Spam** IBM Proventia MS3004
vi. **Proxy Server** Blade Server 460 C.

1.2 ICFRE Network Topology



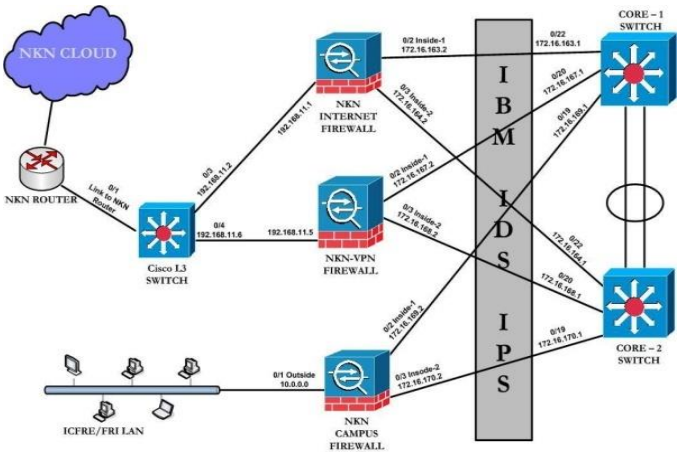


Figure 1. ICFRE Network Topology

1.3 Weaknesses in ICFRE Network:

1. No provision for controlling internally generated spams and FRI campus generated spams.
2. Most of the switches are not manageable inside Campus so proper tracking of illegal activity is not possible.
3. AAA services for Authorization, Authentication and Accounting not present.

II. ANALYSIS TOOL AND DATASET

2.1. Data Mining

Data mining can be defined as extraction of useful information from a large database with help of techniques such as statistics, machine learning and visualization to discover. This is extremely useful as data is previously incomprehensible and after extraction of comprehendible data can lead to major predictions and discoveries. It is a powerful new technology with great potential to help user focus on the most important information in their data warehouses. Data mining tools help in predicting trends, behaviors and patterns which are useful for businesses as decisions have a basis.

Data mining steps in the knowledge discovery process are as follows:

1. Cleaning of data and removing noise by making data consistent.
2. Integration of data with multiple resources.
3. Selection of relevant data for analysis
4. Transformation of data for visualizations.
5. Mining by using intelligent methods for predictions.
6. Identification of trends and patterns.
7. Presentation of knowledge in an easily comprehendible form.

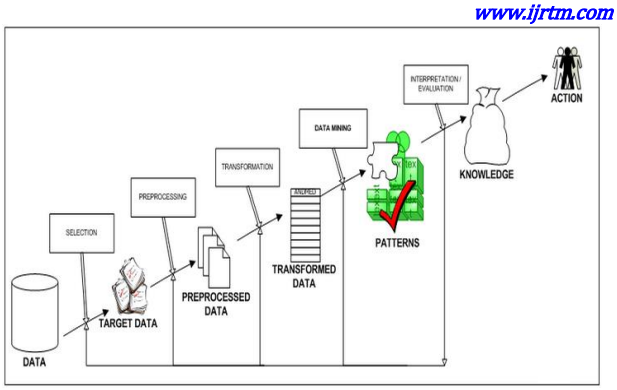


Figure 2. Processes required for Data-Mining

2.1 Weka Tool:

WEKA- Waikato Environment for Knowledge analysis. It was developed by computer science department of University of Waikato New Zealand. It has

- 49 data preprocessing tools
- 76 classification and regression algorithms
- 8 clustering algorithms
- 3 algorithms for finding association rules
- 15 attribute/ subset evaluators + 10 search algorithms for feature selection

The data used in WEKA can be imported in various file formats which are: ARFF, CSV, C4.5 and Binary. The data can also be read from a URL or from an SQL Database.

The explorer interface of Weka can be used for the following-

- Preprocessing the data.
- Building classifiers.
- Clustering data.
- Finding associations.
- Attribute selection.
- Data visualization.

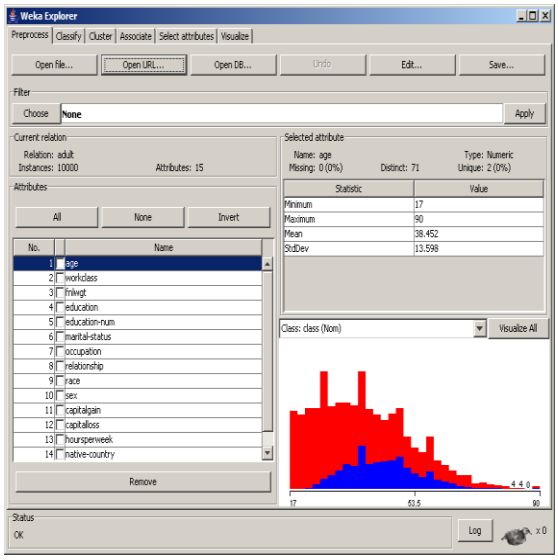


Figure 3. Weka Explorer Interface

2.2 NSL KDD Intrusion Detection System(IDS) Dataset:

The NSL KDD dataset is a refined version of KDD'99 dataset which was based on the DARPA'98 dataset. It is easily available for free online, for researchers who want to work with intrusion detection. It is the publicly available.
The training data was in raw form and had the size of about four Gb in compressed binary TCP dump data. It was network traffic collected over a period of seven weeks. This raw data was then processed into five million connection records. A connection is a sequence of TCP packets which start and end at some certain periods well defined. In this stipulated time data flows to and from a source IP address to a target IP address under some well-defined protocol. Each connection record consists of about 100 bytes and is labeled as an attack or normal traffic.
Attacks fall into four main categories:
DOS: denial-of-service, e.g. syn flood.
R2L: unauthorized access from a remote machine, e.g. guessing password.
U2R: unauthorized access to local super user (root) privileges, e.g., various 'buffer overflow' attacks.
Probing: surveillance and other probing, e.g., port scanning.
The attacks and their types in NSL KDD dataset are:
Table. 1 Attack Types

Attack Type >	DOS	Probe	U2R	R2L
Attack Name V	Back	Nmap	Perl	Warezclient
	Smurf	Portsweep	Rootkit	Warezmaster
	Teardrop	Ipsweep	Loadmodule	Ftp_Write
	Land	Satan	Buffer_Overflow	Guess_Passwd
	Neptune			Imap
	Pod			Spy

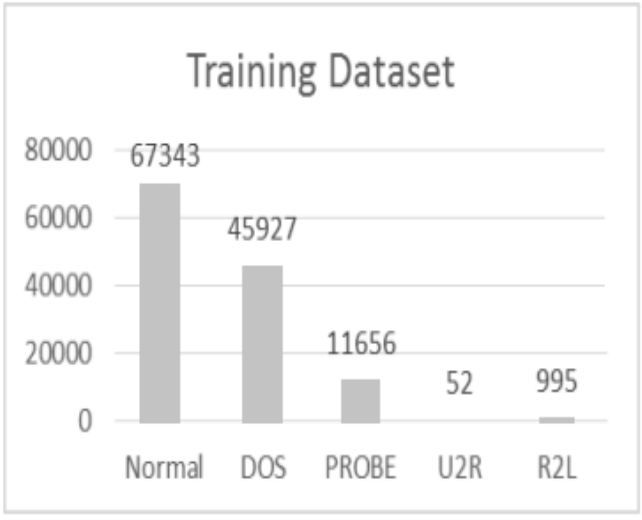


Figure 4. Classes in the NSL KDD IDS Dataset

III. CLASSIFIER

Classification is an important data mining task which is performed to classify a given identified pattern into one of the known classes. This is useful for predictive analysis on unclassified tuples obtained in the future.

3.1 Ada Boost

The AdaBoost algorithm was proposed by Yoav Freund and Robert Schapire and is an important ensemble method. It has a very accurate prediction and is simple to code. It also has several successful applications implemented.
Let X denote the instance space and Y the set of class labels. Assume Y={−1,+1}. Given a weak or base learning algorithm and a training set {(x1,y1), (x2,y2),...,(xm,ym)} where xi ∈X and yi ∈ Y (i=1,...,m), the AdaBoost algorithm works as follows:

- First, it assigns equal weights to all the training examples (xi,yi) (I ∈ {1,...,m}).
- When t rounds of learning have been completed it assigns weights as Dt.
- These weights are nothing but current error E(Ft-1 (xi)) on that sample.
- Using the training set and weight generated that is Dt, the algorithm generates a weak or base learner by calling the base learning algorithm.

The dataset has the following number of attacks and incidents associated with them:

Table. 2 Attack Types and Instances

Class	No. Of Instances	Class	No. Of Instances
Normal	67343	Ftp_Write	8
Neptune	41214	Multihop	7
Warezclient	890	Rootkit	10
Ipsweep	3599	Buffer_Overflow	30
Portsweep	2931	Imap	11
Teardrop	892	Warezmaster	20
Nmap	1493	Phf	4
Satan	3633	Land	18
Smurf	2646	Loadmodule	9
Pod	201	Spy	2
Back	956	Perl	3
Guess_Passwd	53		

3.2 Bagging

In ensemble methods several machine learning algorithms are used to obtain result and get a better predictive performance. The predictions obtained are better than the algorithms if run individually.

Thus an ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. The trained ensemble, therefore, represents a single hypothesis. This hypothesis, however, is not necessarily contained within the hypothesis space of the models from which it is built. Thus, ensembles can be shown to have more flexibility in the functions they can represent. This flexibility can, in theory, enable them to over-fit the training data more than a single model would, but in practice, some ensemble techniques (especially bagging) tend to reduce problems related to over-fitting of the training data.

3.3 Bayes Net

Bayes Net is used for graphing probabilistic relations which are values for random variables. Given a finite set X= (X1,….Xn) of discreet variables which are random in nature, each variable Xi can take up values from a finite set as Val(Xi). A bayesian network is represented by a directed acyclic graph which has the values of a discreet set of values known as joint probability distribution. The links in the graph show the influence of one variable over the other. If there happens to be a link which is directed between Xi and Xj which belong to X (X1… Xn) such that it is from Xi to Xj then Xi will be the parent of Xj. Fig 5. is an example of a Bayesian network, which is hypothetically about the medical domain with 5 variables. The corresponding CPDs are in **Error! Reference source not found.**

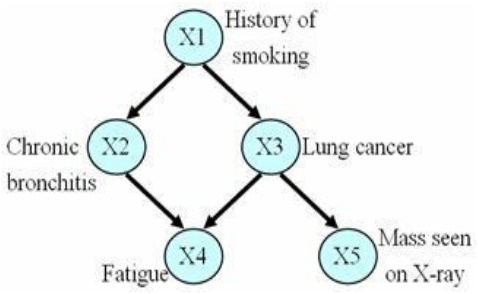


Figure 5. A simple example of Bayesian network

3.4 J48

J48 is an open source Java implementation of the C4.5 algorithm in the weka data mining tool. It builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy.

At each node of the tree, J48 chooses one attribute from the given the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain that results from choosing an attribute for splitting the data.

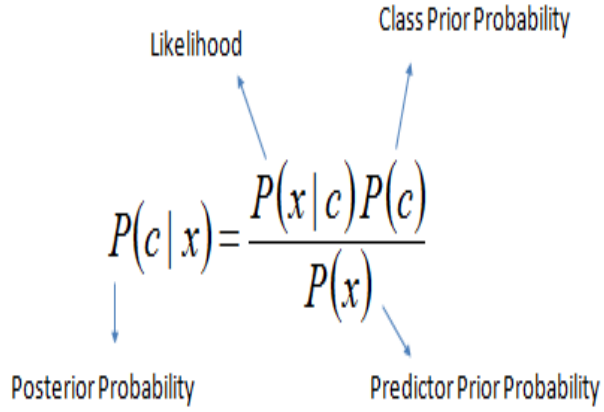
Information gain tells us which attribute in a given set of training feature vector is most relevant for discriminating between classes to be learned. A decision is made for the ordering of attributes in the nodes of a decision tree. Entropy is used for calculation of information gain, higher the entropy, more the information gain.

Information Gain= Entropy (Parent) - [Average Entropy (Children)]

The attribute with the highest normalized information gain is chosen to make the decision. For each attribute, the gain is calculated and the highest gain is used in the decision node.

3.5 Naive Bayes

It is also known as idiot's Bayes, simple Bayes, and independence Bayes. This method is easy to implement, not requiring complicated iterations. It is for the same reason it can be used for large amount of data. Being easy to understand it can also be understood by an unskilled user in classifier technology. Probabilistic approaches to classification typically involve modeling the conditional probability distribution P(C|D), where C ranges over classes and D over descriptions, in some language, of objects to be classified. Bayesian approach divides/splits the posterior distribution into a prior distribution P(C) and a likelihood P(D|C).



$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

3.6 One R

1R: learns a 1-level decision tree thus In other words, generates a set of rules that all test on one particular attribute. Basic version (assuming nominal attributes):

- One branch for each of the attribute's values
- Each branch assigns most frequent class
- Error rate: proportion of instances that don't belong to the majority class of their corresponding branch
- Choose attribute with lowest error rate

3.7 Random Forest

Random forest is a powerful approach to data exploration, data analysis and predictive modeling.

- It can handle thousands of input variables without variable deletion.

- It can run on large databases producing a highly accurate classifier while learning fast.
- It computes complexities between pairs of cases that can be used in clustering, locating outliers or by scaling give interesting views of data.

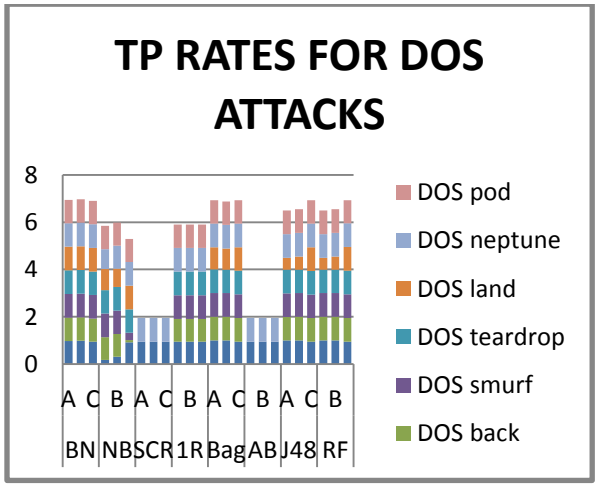
It is an ensemble method which uses recursive partitioning to generate many trees and then aggregate the results. Using a bagging technique, each tree is independently constructed using a bootstrap sample of the data. Trees go very deep and help learning patterns having high irregularity. They also tend to over fit the set used for training. The training algorithm used in random forest is Bootstrap Aggregating. Given a training set $X = x_1, \ldots, x_n$ with responses $Y = y_1, \ldots, y_n$, bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples. After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x'.

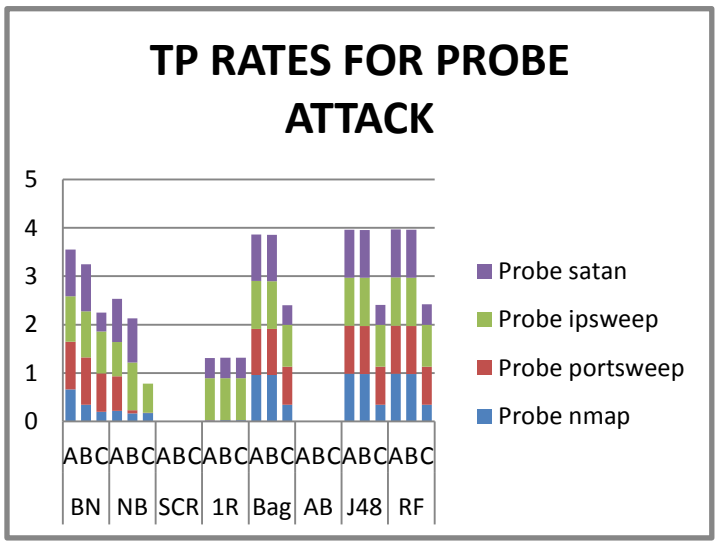
3.8 Single Conjunctive Rule

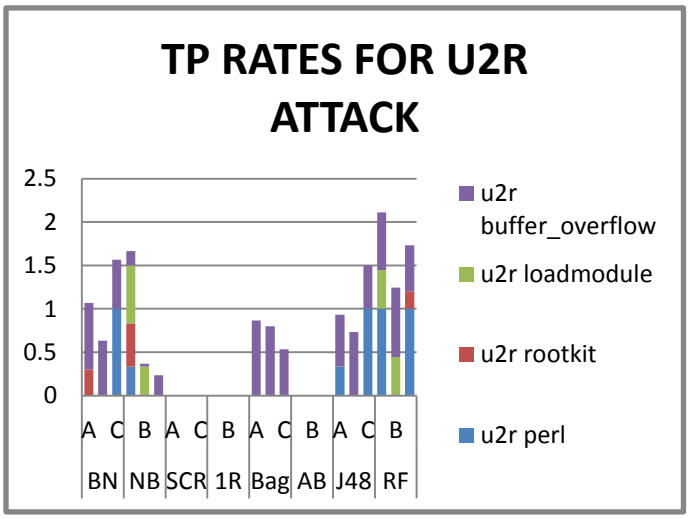
Class conjunctive rule implements a single conjunctive rule learner that can predict for numeric and nominal class labels. A rule consists of antecedents "AND"ed together and the consequent (class value) for the classification/regression. In this case, the consequent is the distribution of the available classes (or mean for a numeric value) in the dataset. If the test instance is not covered by this rule, then it's predicted using the default class distributions/value of the data not covered by the rule in the training data. This learner selects an antecedent by computing the Information Gain of each antecedent and prunes the generated rule using Reduced Error Pruning (REP) or simple pre-pruning based on the number of antecedents.

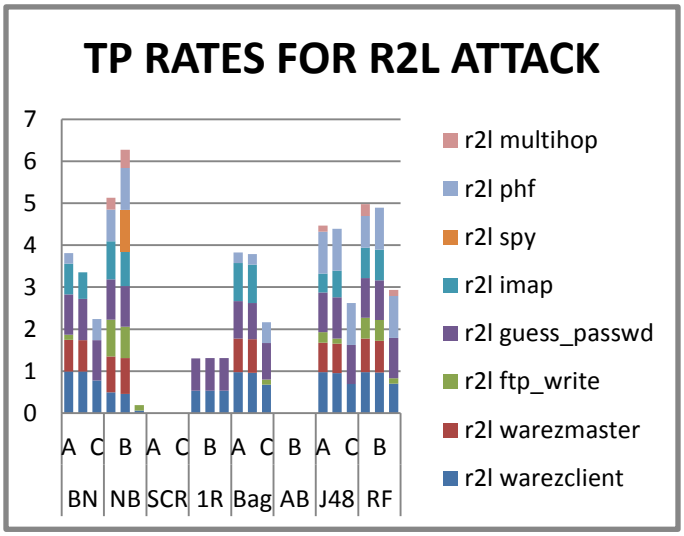
IV. COMPARATIVE EVALUATION OF CLASSIFIER

4.1 Comparative evaluation of True Positive Rates of Classifiers









4.1 Comparative evaluation of Correctly Classified Instances by Classifiers:

Classifier	Correctly classified instances			Incorrectly Classified Instances		
	in A	in B	in C	in A	in B	in C
AB	104807	104807	104807	21166	21166	21166
Bag	125427	125383	117884	546	590	8089
BN	122672	123319	117303	3301	2654	8670
J48	125667	125646	117880	306	327	8093
NB	60048	73246	107977	65925	52727	17996
OneR	114528	114527	114527	11445	11446	11446
RF	125778	125744	117871	195	229	8102
SCR	104807	104807	104807	21166	21166	21166

4.2 Comparative evaluation of Error Rates of Classifiers:

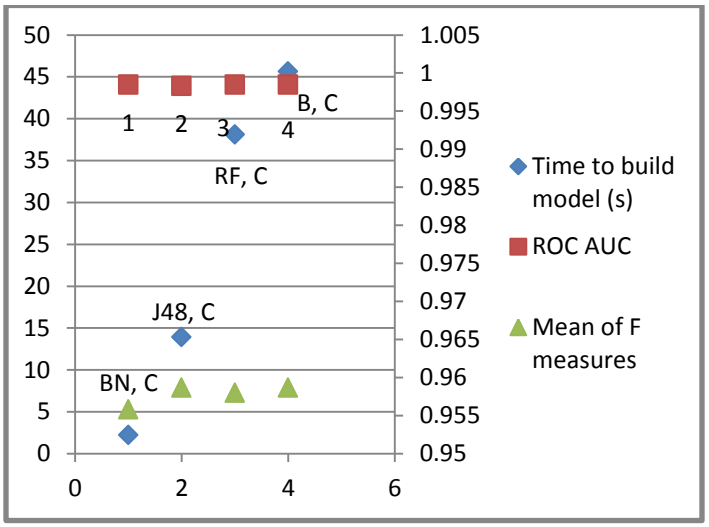
Kappa Statistic- It measures the agreement of prediction with the true class– 1.0 signifies complete agreement.

Classifier	Kappa statistic		
	in A	in B	in C
AdaBoost	0.6985	0.6985	0.6985
Bagging	0.9928	0.9922	0.8931
BayesNet	0.9572	0.9654	0.8857
J48	0.996	0.9957	0.8931
Naive Bayes	0.3846	0.4774	0.7597
One R	0.8478	0.8478	0.8478
Random Forest	0.9974	0.997	0.8931
Single conjunctive rule learner	0.6985	0.6985	0.6985

From the line chart above we can infer that BayesNet, Bagging, J48 and Random Forest have the best kappa statistics and thus their predictions are more agreeable. Also in case of Exp C, kappa statistics values show a slight dip but as Exp C reduces our time considerably we still consider them for further analysis.

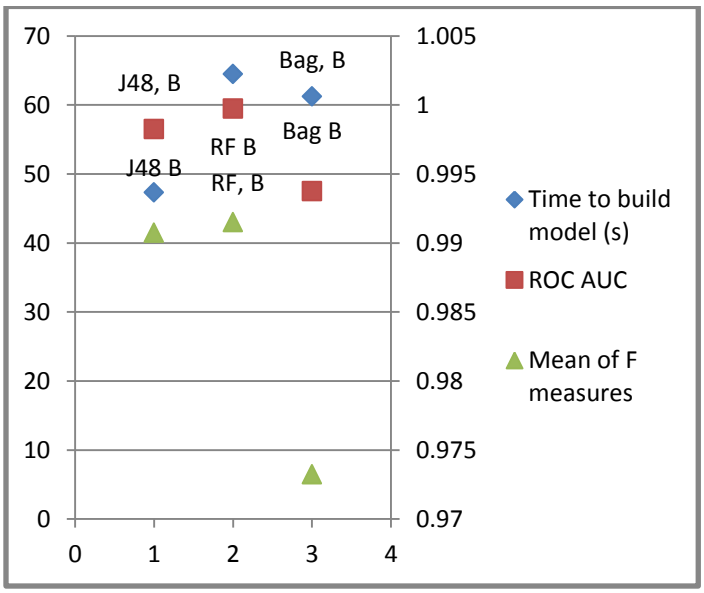
V. RESULTS

I. Results for DOS attacks



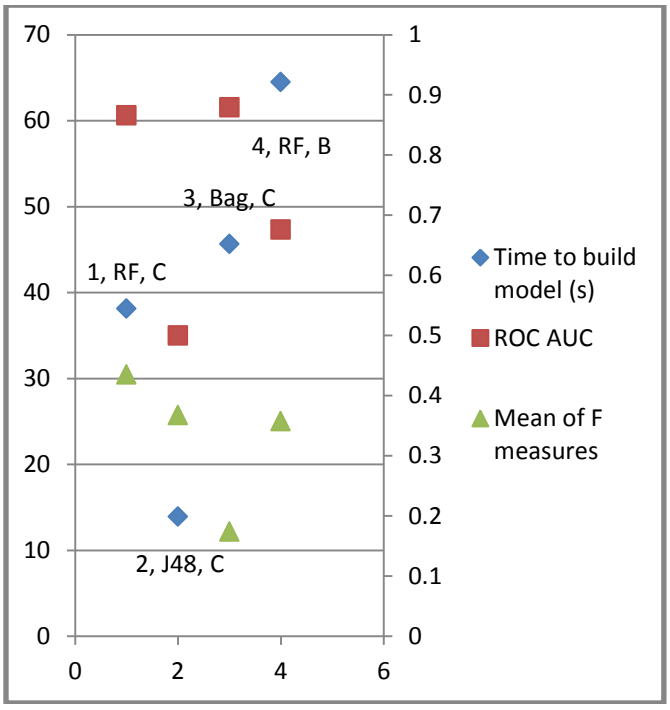
Rank	Classifier	Exp	Mean of F measures	Time to build model (s)	ROC AUC
1	BN	C	0.955833	2.22	0.9985
2	J48	C	0.958667	13.92	0.998333
3	RF	C	0.958	38.13	0.9985
4	Bag	C	0.958667	45.63	0.9985

II. Results for Probe attacks



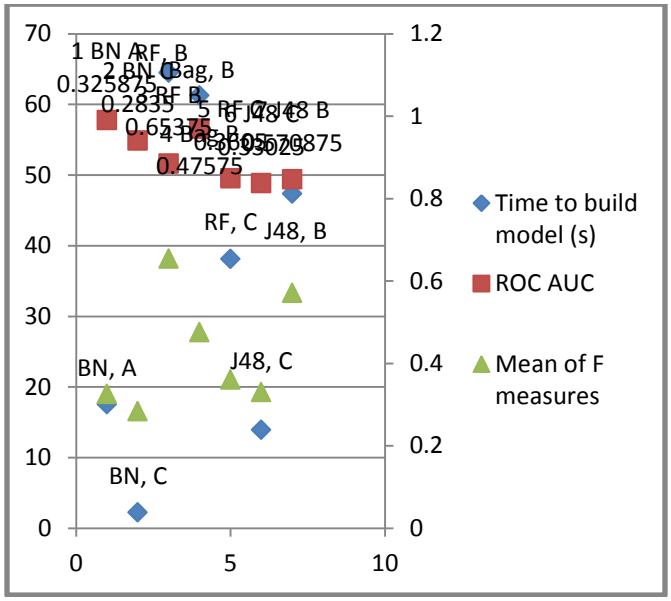
Rank	Classifier	Exp	Mean of F measures	Time to build model (s)	ROC AUC
1	J48	B	0.99075	47.35	0.99825
2	RF	B	0.9915	64.49	0.99975
3	Bag	B	0.97325	61.26	0.99375

III. Results for U2R attacks



ank	Classifier	Exp	Mean of F measures	Time to build model (s)	ROC AUC
1	RF	C	0.43475	38.13	0.866
2	J48	C	0.36725	13.92	0.5
3	Bagging	C	0.174	45.63	0.87925
4	RF	B	0.35725	64.49	0.676

IV. Results for R2L attacks



Rank	Classifier	Exp	Mean of F measures	Time to build model (s)	ROC AUC
1	BN	A	0.325875	17.55	0.9905
2	BN	C	0.2835	2.22	0.940875
3	RF	B	0.65375	64.49	0.884625
4	Bag	B	0.47575	61.26	0.96925
5	RF	C	0.3605	38.13	0.848875
6	J48	C	0.33025	13.92	0.8375
7	J48	B	0.570875	47.35	0.846625

VI. CONCLUSION

In the beginning we saw that network security is of immense importance for organizations like ICFRE which face cyber threats 24x7 all the yearlong. In order to tackle the problem of cyber-attacks the defense imperatives involve a wide variety of network security devices at crucial junctions of data exchange. Cyber Security is a constantly evolving phenomenon which has to be kept up to not only match but also anticipate the ever growing threats and vulnerabilities.

It is necessary for the management to keep itself abreast with the latest devices and applications available in the environment. The search carried out in terms of devices and technologies available vis-à-vis our requirements to counter the weaknesses of the current network security topology was successful in determining the latest approaches and advancements which may help further strengthen our network security measures.

The analysis of NSL KDD IDS dataset (downloaded from http://nsl.cs.unb.ca/NSL-KDD) using data mining classifiers was an effort to search the suitable classifier for classification of large amounts of data. The objective was to devise a method for finding out the best amongst the plethora of classifiers, available on any data mining tool, through certain measures of their performance that it can be used as a template on any set of data and also can be further improved upon.

REFERENCES

[1] Data Mining Concepts and Techniques, Jiawei Han, Micheline Kamber and Jian Pei, Third Edition, Morgan Kaufmann.
[2] On The KDD'99 Dataset: Statistical Analysis for Feature Selection, Journal of Data Mining and Knowledge Discovery, ISSN: 2229-6662 & ISSN: 2229-6670, Volume 3, Issue 3, 2012.
[3] Lazarevic A., Kumar V. and Srivastava J. (2005) Managing cyber threats: issues, approaches, and challenges.
[4] KDD'99 dataset, University of California, Irvine (1999) http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.
[5] NSL KDD IDS Dataset http://nsl.cs.unb.ca/NSL-KDD.
[6] Benchmarking Attribute Selection Techniques for Data Mining by Mark A. Hall and Geoffrey Holmes, Department of Computer Science, University of Waikato Hamilton, New Zealand.
[7] K. Kira and L. Rendell. A practical approach to feature selection. In preceedings of the Ninth International Conference on Machine Learning.
[8] Carl Endorf, Eugene Schultz, Jim Mellander, Intrusion Detection &Prevention, McGraw-Hill, 2004.
[9] http://www.saedsayad.com/naive_bayesian.htm.
[10] ICFRE Website- http://www.icfre.org/.
[11] Comparison of firewall and intrusion Detection System, rchana D wankhade, International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 5 (1) , 2014, 674-678.
[12] Performance Analysis of Data Mining Approaches in Intrusion Detection, P Amudha, H Abdul Rauf, 978-1-61284-764-1/11, IEEE 2011.
[13] 'A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection', International Journal of Engineering Research & Technology (IJERT)Vol. 2 Issue 12, December – 2013.
[14] 'Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features', Adetunmbi A.Olusola., Adeola S.Oladele. and Daramola O.Abosede. Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I WCECS 2010, October 20-22, 2010, San Francisco, USA.