Efficient Attack Detection using Information Gain, C4.5 and Decision Tree Algorithm

Chandrakant Rajpoot[1], Nirupma Tiwari[2]

[1] M.Tech Scholar Department of CSE
[1] SRCEM, Banmore, India.
[2]. Assistant Professor Department of CSE
[2]SRCEM, Banmore, India.

Abstract— **Network is increasing rapidly so, security is a major problem in networks. Internet attacks are growing, and there have been several attack approaches, consequently. Attack detection systems are using various data mining techniques to detect intrusions. In information security, attack detection is necessary because they attempt to compromise the privacy, reliability or availability of a resource. One of the primary challenges for intrusion detection is the problem of misjudgement, mis-detection and lack of real time response to the attack. In this paper, experiment results are calculated at the kddcup99 data set. Feature selection of the data set is executed using Information Gain (IG) and clear dissimilarity between normal and attack data is observed by using C4.5 decision tree algorithm.**

Keywords*—Information Gain; Intrusion Detection System; KDDCUP1999; Security; confidentiality; etc.*

I. INTRODUCTION

Data is very vital to a group. Groups typically wish to reserve the privacy of their data. With the widespread use of the internet, it has become a crucial task to maintain the secrecy and integrity of organization's vital data. A network intrusion attack can be any usage of a network that compromises its strength of the security of information that is stored on computers connected to it. A large number of activity comes under this definition, including an attempt to de-stabilize the network as a whole, gain illegal access to files or rights, simply non license use of software. Additional security actions can stop all such attacks. The aim of attack detection is to develop a system which would routinely scan network activity and detect such intrusion attacks. When an attack is identified, the system in-charge could be informed and thus take corrective action. Conventional methods for network safety include security mechanisms like user validation, cryptography and attack prevention systems like firewalls, Intrusion Detection System (IDS) address problems that are not solved by these techniques. IDS is capable of recognizing these attacks which firewalls are not able to prevent. Also, newer attacks are being developed that are able to penetrate through firewalls, so new approaches are required to defend against these new kinds of attacks.

An IDS is software and/or hardware designed to identify unwanted attempts at log on, manipulating, and/or inactivating of computer system, mainly over a network, such as the internet. One of the main challenging part is to maintain the security of large-scale high-speed networks (LSHSN) is the detection of intrusions in network traffic [1]. A vulnerable network must provide the following:

• Data privacy: Data that is being transferred through the network should be accessible only to those that have been properly authorized.

• Data integrity: Data should preserve their integrity from the moment they are transmitted to the moment they are truly received. No fraud or data loss is recognized either from random events or malicious activity.

• Data accessibility: The network should be robust to Denial of Service attacks.

Attack detection system (ADS) can be classified into two broad categories [2]: Misuse Detection and anomaly Detection.

Misuse/Signature Detection: - The system learns patterns of already known attacks. These well-read patterns search through the received data to find intrusions of the previously known types. This method is not talented in detecting new attacks that do not follow pre-defined patterns.

Anomaly Detection: - Here patterns are learned from normal data. The unseen data are checked and searched to find deviations from these learned patterns. These deviations are 'anomalies' or possible intrusions. This method is not capable of identifying the type of attack.

To understand the key ideas behind the above two approaches of IDS let us take an example considers a security guard present at an entrance who is responsible for allowing only valid persons to pass through the gate. One approach that the guard may follow would be to maintain a database of photographs of well-known culprits who should not be allowed entry. The guard can then check each incoming person with the database and find out if the person is one of those culprits. If so, the guard prevents the culprit from passing through the entrance. The problem here is that a culprit whose photograph is not in the database will be permitted entry. This method corresponds to the Misuse Detection approach.

Another approach that the guard may follow is to maintain a database of photographs of all the valid persons to be allowed entry. The guard allows entry to the entering person, only if his picture is stored in the database. This way, all persons whose photographs are not found in the database are identified as culprits and not permitted entry. This approach corresponds to the Anomaly detection method.

Data mining from anomaly detection point of view is the search of malicious (in case of misuse detection) activity patterns or normal activity patterns (in the case of anomaly detection) from the large amount of data traveling through the network or stored in system logs. One of the important steps in the data mining is to describe the data by summarizing its statistical attributes. The selection of the useful attributes holds the key to the success of the data mining system. This selection is done at the pre-processing stage of any data mining process. The use of extra features or using fewer features may drive the data mining system in a wrong way.

Data mining is used to classify attack because of following reasons
data mining is used to solve network attack problem because of the following reasons [14-16]:

- Data mining algorithm can handle large volumes of data.
- Data mining can effectively find unseen information from huge volumes of data.

Data mining algorithms are used to implement data summarization and visualization that help the security analysis in several research areas [17].

In this paper, dimensionality reduction of data set is performed on the basis of their gain values and proposed methods are tested using kddcup99 data set. The KDD 99 intrusion detection datasets are based on the 1998 DARPA initiative, which provides designers of intrusion detection systems (IDS) with a benchmark on which to evaluate different methodologies [12-13]. An optimization technique is also used to improve the detection rate and reduce false positive rate during training and testing of data set.

The remainder of this paper is organized as follows. The next section deliberated related work. Section 3 describes proposed algorithm in detail. An implementation and evaluation is described in Section 4. The conclusion is in Section 5.

II. RELATED WORK

In the following we summarize some of the recent research works in the area of intrusion detection Ming Xue and changjun Zhu [3] give the idea of a data mining algorithm for intrusion detection. They researched and give two key algorithms namely the pattern comparison and clustering algorithm. In pattern comparison, they first collect a normal behaviour pattern under association rules then they differentiate normal behaviour and intrusion behaviour. The basic idea of clustering analysis creates in the difference between intrusion and normal pattern and in the fact that the number of normal patterns should exceed that of intrusion pattern, so that apply data sets into different categories and detect intrusion by distinguish normal and abnormal behaviours. These techniques suggested was good but some points like correct rate of intrusion detection, control the rate of false alarm in intrusion detection provide me the redirection for searching effective data mining algorithm.

Mohammadreza Ektefa et al. [4] use C4.5 and SVM (support vector machine) for detecting attacks. They calculate the detection rate (percentage of detecting attacks among all attack data) and false alarm rate (percentage of normal data which is wrongly recognized as an attack) and compare both algorithm result and find C4.5 has better performance than SVM in both detection and false alarm rate. The data used by the author is KDD cup99 dataset. The first stage is pre-processing. Data in this phase partition into training and testing. In the next step, they applied C4.5 and SVM on the training dataset in order to build and train the models. Finally, trained models are examined on the testing dataset to see the efficiency of them. In future authors will be examined more powerful techniques for detecting attacks in the network. Furthermore, data mining techniques can be applied in order to improve the quality of data.

Shu Wu and Shengrui Wang [5] discuss about outlier detection can usually considered as pre-processing for discovery new or rare attacks. Authors are investigating outlier detection for categorical data sets. In this paper author proposed optimization model of attack detection via a new concept of hole entropy. They proposed two parameters named ITB-SS and ITB-SP. Experimental result shows that these parameters are more effective and efficient than mainstream methods and can be used with large and high dimensional data sets where existing algorithms fail. They apply the greedy approach to develop two efficient algorithms, ITB-SS and ITB-SP, which provide practical solutions to the optimization problem for outlier detection. They also estimate an upper bound for the number of outliers and an anomaly candidate set. This bound, obtained under a very reasonable hypothesis on the number of possible outliers, allows us to further reduce the search cost. The proposed algorithms have been evaluated on real and synthetic data sets, and compared with different mainstream algorithms.

Ashish Kumar et al. [6] says that effect of internet in our daily life is increased day by day. Security becomes major problem within and outside the organizations. They also discuss about the common steps that organizations take to secure their computers. Different types of attacks are increasing rapidly.

Yuh-Jye Lee et al.[10] discuss about intrusion or credit card fraud detection that require an effective and efficient framework to identify deviated data instances. They propose an online oversampling principal component analysis (osPCA) algorithm to solve detecting the presence of intrusion from a huge amount of data via an online updating technique.

A.M. Chandrasekhar and K. Raghuveer [11] says that the intrusion detection is one of the software to resolve the problem of network security. They discuss many researchers are using data mining techniques to build intrusion detection system. They propose a new approach by using neuro fuzzy and support vector machine. They use K-mean clustering to generate training various subsets to trained neuro fuzzy models and SVM classification is used. After that they used radial SVM to detect intrusions. They used KDDCUP99 data sets.

Previous research has been carried out with many Intrusion detection methods for detecting various types of attack categories. Each Algorithm is giving the different result with the some parameters like false alarm rate, detection rate, accuracy, precision, and recall, f-measure of all attacks and overall accuracy of complete method.

III. PROPOSED WORK

In this paper we have proposed a methodology in which feature selection is performed using information gain, classification is done using C4.5 Decision tree algorithm and random forest is used to optimize the result of attack class. The proposed model is described below.

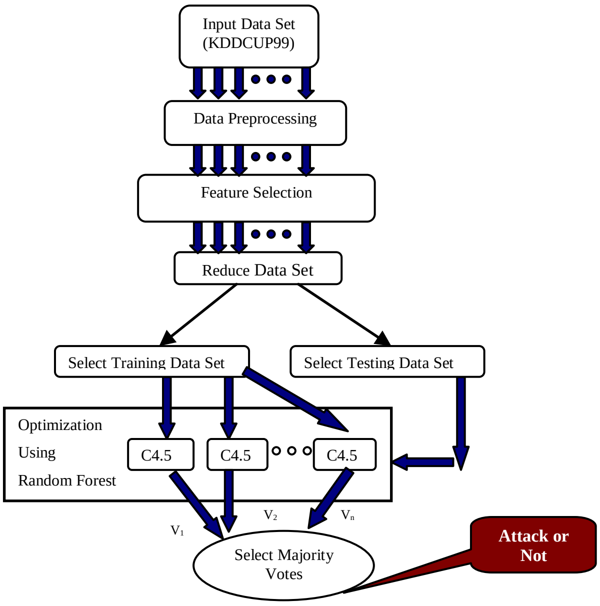

Fig.1 Proposed Model

Following are the steps used in proposed methodology.

Step 1 Select data set with 25000 instances from 10% Kddcup99 dataset. Convert this dataset into Attribute-Relation File Format (arff).

Step 2 apply Information gain for feature selection. Steps are given below

1. Calculate estimated information required to categorize a given instance

$$I(d_1, d_2, ...d_m) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Where d_i is the no of instances belong to class level C_i

$$p_i = \frac{d_i}{D}$$

p_i is probability of attribute instance related to class C_i

2. Compute the entropy that is related to the partitioning into subsets by attribute X.

$$E(X) = -\sum_{j=1}^{v} |D_v| / |D| I(d_{1j}, d_{2j}, ...d_{mj})$$

Where $|D_v| = (d_{1j} + d_{2j} + ... + d_{mj})$

| D | = Total no of instances

Where $p_{ij} = d_{ij} / |D_j|$ is the probability that a instance in D_j belongs to class C_i.

3. Compute information that would be gained by dividing on X is

Gain (X) = I (d_1, d_2, ……., d_m) – E(X)

Gain (X) is the expected drop in entropy produced by the value of Attribute A.

4. Recap step 2 to 3 and Calculate the information gain of each attribute
5. Arrange these value larger to smaller and select gain value >0.1

Step 3 after feature selection apply Dataset D to C4.5

Decision Tree algorithm

Tree = Null
if *D* is "pure" OR other preventing conditions seen
then
stop
end if
for all attribute *x* € *D* **do**
Compute information-theoretic criteria if split on *x*
end for
x_{best} = Best attribute giving to above calculated Conditions
Tree = Build a decision node that tests x_{best} in the root
D_v = Generated sub-datasets from *D* based on x_{best}
for all D_v **do**
$Tree_v$ = C4.5(D_v)
Attach $Tree_v$ to the matching branch of Tree
end for Loop
return Decision Tree

Where Information Theoretic criteria for selecting best attribute to create decision tree can be calculated as follows:

1. Compute expected information needed to categorize a given instance is given by

$$\mathbf{I(d_1, d_2, ..., d_m) = -\sum_{i=1}^{m} p_i \log_2 (p_i)}$$

Where $\mathbf{d_i}$ be the number of instances of **D** in class $\mathbf{C_i}$.

$$\mathbf{p_i = d_i / D}$$

$\mathbf{p_i}$ is the probability that an attribute instance belongs to class $\mathbf{C_i}$.

2. Calculate the entropy belong to the subsets of attribute **X**.

$$\mathbf{E(X) = -\sum_{j=1}^{v} |D_v| / |D| \quad I(d_{1j}, d_{2j}, ..., d_{mj})}$$

Where $\mathbf{| D_v | = (d_{1j} + d_{2j} + ... + d_{mj})}$

| D | = Total no of samples

Where $\mathbf{p_{ij} = d_{ij} / | D_j |}$ is the probability that a instance in $\mathbf{D_j}$ related to class $\mathbf{C_i}$

3. Calculate the expected drop in entropy

Gain (X) = I (d_1, d_2, ..., d_m) – E(X)

Gain (X) is the expected drop in entropy Produced by knowing the value of Attribute **X.**

4. Calculate gain ratio for attribute A

Gain ratio (X) =Gain (X) / Entropy (X)

5. Compute the gain ratio of each attribute. The attribute with the highest gain ration is chosen as the test attribute for the given set D. A node is created and labelled with the attribute, branches are created for each value of the attribute, and the samples are divided accordingly.

Step 4 Optimization is performed by applying Random Forest

Algorithm

1. Build bootstrapped sample S_i from the original dataset D, where $|S_i| = |D|$ and instances are selected at random with replacement from D.
2. Build tree T_i with using S*i* as the training dataset and decision tree algorithm(C4.5 **).** Pruning is not performed with tree.
3. Repeat steps (1) and (2) for i = 1, ... , no. of trees, Forming a jungle of trees T_i resultant from different bootstrap instances.
4. When categorizing an example x, collect all the votes given by different tree. The max no of votes given by tree to a particular class is the result.

Step 5 Exit

IV. IMPLEMENTATION AND EVALUATION

To investigate the effectiveness of the proposed model for attack detection. We perform some experimental task, all these tasks perform in jdk1.6 software and well famous intrusion data set kddcup99 provided by DARPA agency.

To evaluate the effectiveness of both the classification algorithm over the DARPA test data, it describes the results using Detection Rate (DR), False Positive Rate (FPR), and Accuracy (ACC). Each metric is defined below.

Detection of attack can be measured by following metrics:

- False positive (FP): it is the total no of attack detected but they are normal.
- False negative (FN): it is the total no of normal instances detected but they are actually attack instances.
- True positive (TP): it is the total no of attack instances detected and they are actually attack instances.
- True negative (TN): it is the total no of normal instances detected and they are actually normal instances.

1. Detection Rate (DR): Detection rate is the ratio of correctly classified intrusive examples to the total number of intrusive examples.

$$DetectionRate = \frac{TP \times 100}{TP + TN}$$

2. False Positive Rate (FPR): False positive rate is the ratio of incorrectly classified normal examples (false alarms) to the total number of normal examples.

$$FalseAlarmRate = \frac{FP \times 100}{FP + FN}$$

3. Accuracy (ACC): Accuracy is the ratio of correctly classified examples to the total number of classified examples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100$$

4. Recall: - what percentage of the positive cases did you catch?

$$Recall = \frac{TP}{TP + FN}$$

5. Precision:-what percentage of positive predictions was correct?

$$Precision = \frac{TP}{TP + FP}$$

6. F-measure:-

$$F-measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

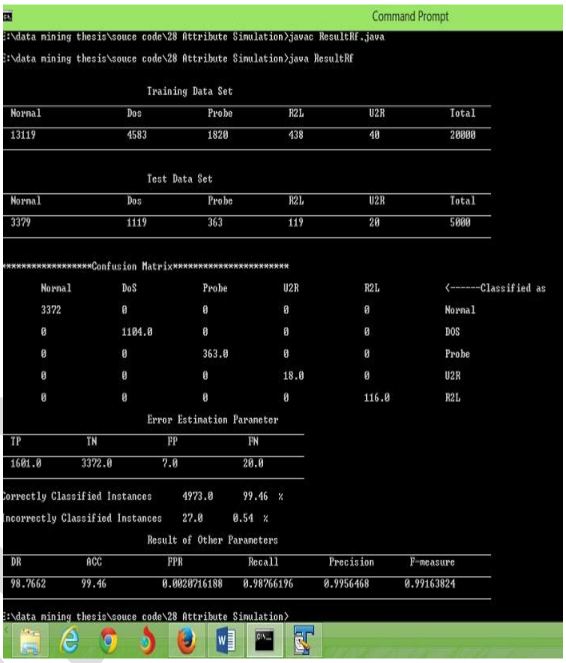

Fig. 2 Java Implementation of proposed work using jdk1.6

TABLE I
DETECTION RATE COMPARISONS OF DIFFERENT ATTACKS THROUGH C4.5 AND PROPOSED METHODOLOGY

Test Set	**Attack Type**	**Classifier**	
		C4.5	**Proposed Methodology**
41 Attribute Set	DoS	97.31	98.21
	Probe	99.72	100.0
	U2R	80.00	85.00
	R2L	40.33	41.17
28 Attribute Set	DoS	97.94	98.65
	Probe	99.44	100.0
	U2R	75.00	90.00
	R2L	75.63	97.47

TABLE II
DETECTION RATE COMPARISON OF ATTACK WITH VARIOUS METHODS

Method/Category of Attack	**Dos**	**U2R**	**Probe**	**R2L**
C4.5 [4]	93.87	33.33	95.38	16.44
Support Vector Machine (SVM) [4]	93.84	66.67	89.09	15.90
C4.5	97.94	75.00	99.44	75.63
Proposed Methodology	98.65	90.00	100	97.47

TABLE III
COMPARISON OF RESULT WITH OTHER METHODS

Algorithm/Parameter	ACC	FPR
Naive BayesNB [8]	92.69	0.077
Bayes Net[8]	99.10	0.001
OneR[8]	92.69	0.002
NB-ACO[9]	97.00	1.91
C4.5	98.18	0.009
Proposed Methodology	99.46	0.002

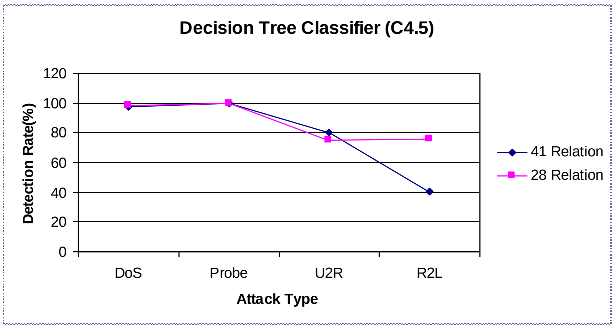

Fig. 3 Attack found in tested data (C4.5 algorithm with 28 and 41 attributes)

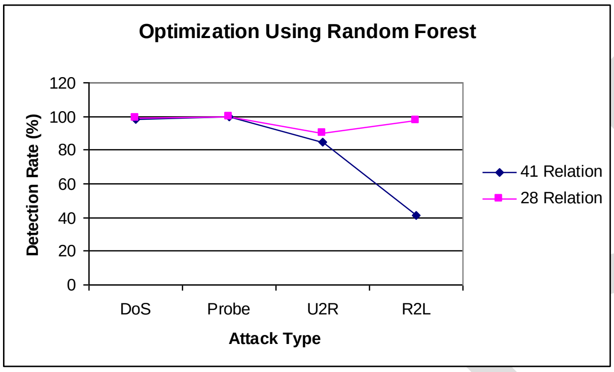

Fig. 4 (RF&C4.5 algo with 28 and 41 attributes)

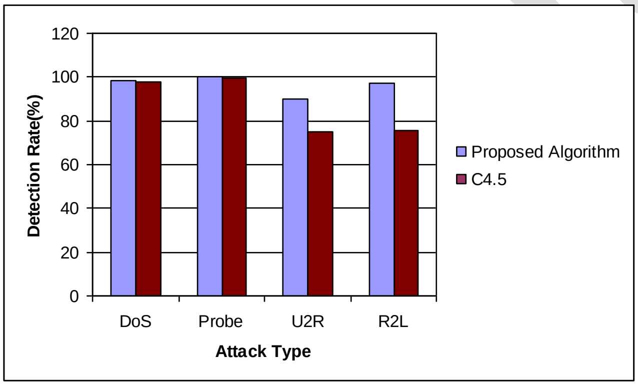

Fig. 5 Comparison of proposed algorithm with C4.5

Our experiment shows the promising result as compare with earlier approaches.

V. CONCLUSION

In this paper, an algorithm based on the Decision Tree Classification for analysing program behaviour in intrusion detection is evaluated by experiments. The Random forest is used for improvement of detection rate during testing of the data set. The preliminary experiments with the 1998 DARPA audit data have shown that this proposed methodology approach is able to effectively detect intrusive program behaviour. The test data contain 4 kinds of different attacks in addition to normal system call. This proposed methodology in Comparison to other classifiers has found it to be comparable in some domains.

To improve the usability of the IDS, we can use supervised and unsupervised learning algorithms. For relevant feature selection a hybridized Rough-PCA Methodology of Attribute Reduction can be used. Through which high dimensionality of the data set into lower dimension with most important attribute set.

REFERENCES

[1.] L. Breiman, "Random Forests", *Machine Learning* 45(1):5–32, 2001.
[2.] T. Bhavani et al., "Data Mining for Security Applications," Proceedings of the 2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing - Volume 02, IEEE Computer Society, 2008.
[3.] Ming Xue and Changjun Zhu "Applied Research on Data Mining Algorithm in Network Intrusion Detection" International joint Conference on Artificial Intelligence ,IEEE,978-0-7695-3615-6/09 © 2009.
[4.] Mohammadreza Ektefa , Sara Memar, Fatimah Sidi ,Lilly Suriani Affendey "Intrusion Detection Using Data Mining Techniques" IEEE 978- 1-4244-5651-2/10 © 2010.
[5.] Shu Wu and Shengrui Wang, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL. 25, NO. 3, MARCH 2013.
[6.] Ashish Kumar, Shrikant Chandak ,Rita Dewanjee, "Recent Advances in Intrusion Detection Systems: An Analytical Evaluation and Comparative Study " ,International Journal of Computer Applications (0975 – 8887) Volume 86 – No 4, January 2014
[7.] Prabhjeet Kaur , Amit Kumar Sharma, Sudesh Kumar Prajapat "MADAM ID FOR INTRUSION DETECTION USING DATA MINING" IJRIM Volume 2, Issue 2 (ISSN 2231-4334) (February 2012).
[8.] Yogendra Kumar jain and Upendra "An Efficient Intrusion Detection Based on Decision Tree Classifier Using Feature Reduction " International Jornal of Scientific and Research Publications, ISSN 2250-3153,Volume 2,Issue 1,January 2012.
[9.] Namita Shrivastava and Vineet Richariya " Ant Colony Optimization with Classification Algorithms used for Intrusion Detection" ,International Journal of computational Engineering & Management Volume 15 Issue 1,January 2012.
[10.] A.M. Chandrasekhar and K. Raghuveer ," Intrusion Detection Technique by using K-means,Fuzzy Neural Network and SVM classifiers", International Conference on Computer Communication and Informatics(ICCCI-2013), IEEE Jan. 04 06, 2013,Coimbatore,INDIA.
[11.] MIT Lincoln Laboratory. DARPA Intrusion Detection Evaluation Data Sets. Available at http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/index.html,1999.
[12.] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.
[13.] T. Lappas and K. P. ,"Data Mining Techniques for (Network) Intrusion Detection System," January 2007.
[14.] S. Sun, Y. Wang, "A Weighted Support Vector Clustering Algorithm and its Application in Network Intrusion Detection," etcs, vol. 1, pp. 352-355, 2009 First International Workshop on Education Technology and Computer Science, 2009.
[15.] S. Wu, E. Yen. "Data mining-based intrusion detectors," Elsevier Computer Network, 2009.
[16.] E. Bloedorn et al, "Data Mining for Network Intrusion Detection: How to Get Started," Technical paper, 2001.