

# A COMPARATIVE STUDY OF CLASSIFICATION TECHNIQUES FOR INTRUSION DETECTION USING NSL-KDD DATA SETS

**Dr.K.Arunesh<sup>1</sup>, M. Manoj Kumar<sup>2</sup>**

<sup>1</sup>Associate Professor, Department of Computer Science, Sri S. R. N. M. C College, Sattur (India)

<sup>2</sup>M.Phil Scholar, Department of Computer Science, Sri S. R. N. M. C College, Sattur(India)

## ABSTRACT

Data Mining is a technique to drilling the database for giving meaning to the approachable data. It involves systematic analysis of large data sets. And the classification is used to manage data, sometimes tree modeling of data helps to make predictions about new data. Recently, we have increasing in the number of cyber-attacks, detecting the intrusion in networks become a very tough job. In Network Intrusion Detection System (NIDS), many data mining and machine learning techniques are used. However, for evaluation, most of the researchers used data set DARPA 2000, which has widely criticized not suitable for current network situation. We have labeled a network dataset and also an improved version of KDD Cup datasets, called NSL-KDD dataset. In NSL-KDD data set, every instant is labeled as normal (no attack), attack (Dos, U2R, R2L, and Probe). In NSL-KDD dataset we have only a selected dataset to provide a good analysis on various machine learning techniques for intrusion detection. This analysis explains discussion of Random Forest, J48, ZeroR, and Naïve Bayes. Among them we get best classification algorithm for the given dataset.

**Keywords:** Data mining, Intrusion Detection., J48, Naïve Bayes , Random Forest, ZeroR,

## I INTRODUCTION

Intrusion detection is a type of security management system for computers and networks. An ID system gathers and analyzes information from various areas within a computer or a network to identify possible security breaches, which include both intrusions i.e., the attacks from outside the organization and misuse attacks from within the organization. KDD process is used to denote the process of extracting useful knowledge from large dataset. Data Mining is the process of discovering knowledge from the large amount of dataset. Data source can include databases, data warehouses, the web, and any other repositories. Data Mining is the most vital step in the NSL-KDD process and it applies data mining to extract patterns from the data.

Data Mining was generally refers to the process of automatically extracting the models from large stores of data. The recent development in data mining has made available a wide variety of algorithms, drawn from the fields of statics, pattern recognition, machine learning and database. We have more chances of data loss; hacking and

intrusion have been increased with the growth and popularity of the Internet. When continuously growing Internet attacks suppose severe challenges to develop a flexible and adaptive security oriented methods. An intrusion can be defined as a series of actions that compromises the integrity, confidentiality or availability of a computer resource.

**Intrusion Detection System (IDS)** is one of the most important components being used to detect the Internet attacks that can be either host based or network based. Intrusion detection is the process of monitoring and analyzing the activities occurring in a computer system or in a network in order to detect signs of security problems. IDS needs only to detect threats and as such is placed out-of-band on the network infrastructure, meaning that it is not in the true real-time communication path between the sender and receiver of information. In some cases the IDS may also respond to anomalous or malicious traffic by taking action such as blocking the user from accessing the network.

In this paper we have a study on comparing classification algorithms. However, most of these studies have been limited to only a very few classification algorithms. The theme of my thesis is to compare and better understand the prevalent classification algorithms, by evaluating the performance of four different classification algorithms on real network datasets.

## II LITERATURE SURVEY

Many data mining techniques have been used for comparative study of classification algorithm. In 2012, Sunitha B.Aher and Lobo L.M.R.J, compare five algorithms ADTree, Simple Cart, J48, and ZeroR & Naïve Bayes algorithm for course Recommender system. ADTree classification algorithm works better for this dataset. In 2013, S.Revathi and A. Malathi gave an explanation on classification algorithm like Random forest, J48, SVM, CART, and Naïve Bayes for intrusion detection. In 2013, Delveen Luqman Abd Al-Nabi and Shereen Shukri Ahmed gave a discussion about the survey on classification algorithm for data mining. In this survey, the CART decision tree algorithm is one of the best algorithms for classification of data. In 2013, G.Kesavaraj and S.Sukumaran presents a study on performance analysis of data mining algorithm in classification and provides a result as that Random Forest algorithm has very less error rate, when comparing to other classification algorithm. In 2016, Amit Gupta, Ali syed, Azeem Mohammad and Malka N.Halgamuge discuss a comparative study on classification algorithm using data mining used for crime & Accident in Denver city the USA. In this paper, the classification algorithm used in this study is to access trends and patterns that are assessed by BayesNet, Naïve Bayes, J48, JRip, OneR and Decision Table. In this analysis JRip and Decision Table classified the most number of correct incidents and Naïve Bayes Model Builds the Quickest time with 0.57 sec.

## III CLASSIFICATION ALGORITHM

Classification is one of the Data Mining techniques, it is mainly used to analyze a given data set and takes each instance of it and assigns this instance to a particular class such that classification error will be least. It is used to extract models that accurately define important data classes within the given data set. Classification is a two-step process. During first step the model is created by applying classification algorithm on training data set then in

second step the extracted model is tested against a predefined test data set to measure the model trained performance and accuracy. So classification is the process to assign class label from data set whose class label is unknown.

### 3.1 Random Forest Classification Algorithm

A Random forest is a new approach to data exploration, data analysis and predictive modeling. The first algorithm of Random forest was created by Tin Kam Ho by using the random subspace method and the extension of the random forest algorithm was developed by Leo Breiman, who was the father of CART (R). The random forest is a collection of CART – like trees. The Random Forests grows an ensemble of decision tree. The random forest algorithm uses the bagging technique for building an ensemble of decision trees. Bagging is also known to reduce the variance of algorithm. The ensembles are more effective when the individual models that comprise them are uncorrelated. In traditional bagging with decision trees, the constituent decision trees may end up to be very correlated because the same features will tend to be used repeatedly to split the bootstrap samples. By restricting each split-test to a small, random sample of features, we can decrease the correlation between trees in the ensemble.

### 3.2 J48 Classification Algorithm

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan (known in Weka as J48 J for Java). By default J48 creates decision trees of any depth. The decision needs generated by C4.5. The decision tree generated by C4.5 can be used for classification and also the referred to as statistical classifier. The C4.5 algorithm builds a decision trees from a set of training data in the same way as ID3, using the concept of information entropy.

### 3.3 Naïve Bayes Classification Algorithm

Naïve Bayes is a classification algorithm to represent a binary (two-class) and multi-class classification problems. This technique is easiest to understand when describing using binary or categorical input values. It is called as naïve bayes or idiot bayes because the calculations of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value  $P(d_1, d_2, d_3|h)$ , they are assumed to be conditionally independent given the target value and calculated as  $P(d_1|h) * P(d_2|h)$  and so on. This is very strong assumption that is most unlikely in real data, i.e., that the attributes do not interact.

### 3.4 ZeroR Classification Algorithm

ZeroR is the simplest classification method which relies on the target and ignores all predictors and otherwise called 0-R or ZeroR in Weka. ZeroR classifier simply predicts the majority category (class). The 0-R (zero rules) classifier takes a look at the target attribute and its possible values. It will always output the value that is most commonly found for the target attribute in the given dataset. 0-R as its names suggests; it does not include any rule that works on the non-target attributes although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods.

#### IV DATASET

NSL-KDD is a dataset for network-based intrusion detection systems. For experimental study we have used NSL-KDD data set which is an improved version of KDD Cup99 data set and consists of selected records of the complete KDD Cup99 data set. It contains essential records of the complete KDD99 Cup data set. It is the new version of KDD Cup99 dataset. The testing dataset used for experimental purposes has 22 different attacks out of total 37 present in the dataset. The training dataset used for experimental purposes has 23 different attacks out of total 37 present in the dataset. NSL KDD data set is made up of 41 different attributes as shown in Table I and there are five attack classes one of which is normal and other four are different types of attack. These attack types are grouped into four categories as shown in Table II shows different instances of data set present in training and testing data set of NSL KDD data set.

Duration	Is_guest login
Protocol_type	Count
Service	Srv_count
Flag	Serror_rate
Src_bytes	Srv_serror_rate
Dst_bytes	Same_srv_rate
Land	Diff_srv_rate
Wrong_fragment	Srv_diff_host_rate
Urgent	Dst_host_count
Hot	Dst_host_srv_count
Num_failed_logins	Dst_host_Same_srv_rate
Logged_in	Dst_host_diff_srv_rate
Num_compromised	Dst_host_same_src_port_rate
Root_shell	Dst_host_srv_diff_host_rate
Su_attempted	Dst_host_serror_rate
Num_root	Dst_host_srv_serror_rate
Num_file_creations	Dst_host_rerror_rate
Num_shells	Dst_host_srv_rerror_rate
Num_access_files	Class
Num_outbound_cmds	diffic
Is_host_login	

**TABLE I. NSL-KDD datasets Features**

Attack Types	Attack Names
Normal	normal
Probe	Portssweep, Satan, nmap, ipsweep,
Dos	Back, teardrop, murf, pod, Neptune, land, udpstorm, worm, apache2, processtable.
r2l	Warezclient, named, ware master, spy, phf, multihop, imap, guess_passwd, ftp_write, xclock, xsnoop, snmpguess, snmpgetattack, httptunnel, endmail,.
u2r	Root kit, Perl, load module, buffer_overflow, sqlattack, xterm, ps.

**TABLE 2: Attack Types in NSL-KDD Datasets**

NSL-KDD is a dataset proposed by Tavallaee et al. NSL-KDD dataset is a refined version of the original KDD Cup99 dataset. NSL-KDD consists of the same features as KDD Cup99. The NSL - KDD datasets consists of 41 features and one as Class attribute. The Class attribute has 37 different attacks that fall under four types of attacks: Probe attacks, User to Root (U2R) attacks, Remote to Local (R2L) attacks and Denial of Service (DoS) attacks.

## V EXPERIMENTAL AND RESULTS

For the experiments we use very popular data mining tool, WEKA and effectiveness of the classification algorithms in classifying the NSL-KDD data set is analyzed. The data in the NSL-KDD dataset is either labeled as normal or as one of the 37 different kinds of attack. These 37 attacks can be grouped into four classes: Probe, DoS, R2L, and U2R.

WEKA Tool is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

WEKA Tool consists of four applications namely Explorer, Experimenter, Knowledge flow, Simple Command Line Interface and also Java interface. The experimental steps are as follows

1. Select and preprocess the dataset.
2. Run the classifier algorithm.
3. Compare the classifier result.

Before applying any classification techniques to the NSL-KDD dataset, we have to perform discretization as preprocess. Discretization is the process of turning numeric attributes into nominal attributes. The main benefit is that some classifiers can only take nominal attributes as input, not numeric attributes. Another advantage is that some classifiers that can take numeric attributes can achieve improved accuracy if the data is discretized prior to learning.

This experiment is performed using test, train datasets and also the datasets with the attacks in NSL – KDD datasets. The classification Accuracy rate of the NSL-KDD datasets are given as correctly classified instances and incorrectly classified instances is mentioned in the Table.

In J48 algorithm, an output we got a decision tree. Fig. 1, it shows the tree visualization was generated by WEKA..

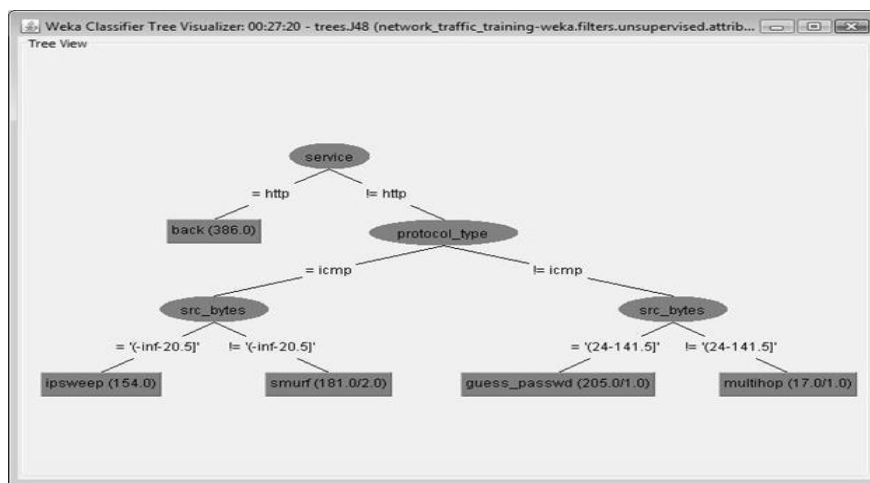
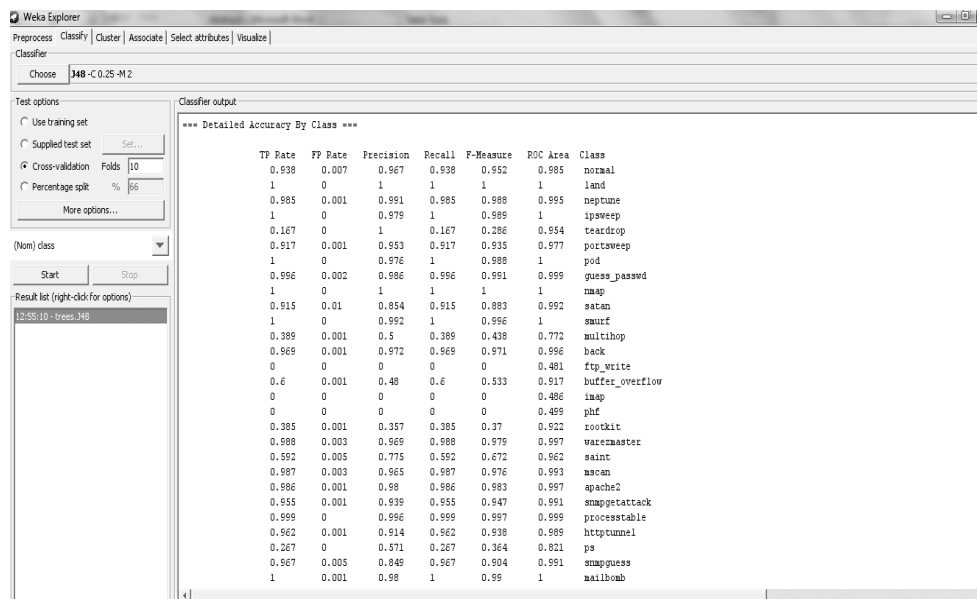


FIGURE 3: Decision Tree generated by J48 algorithm.



TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.938	0.007	0.967	0.938	0.952	0.985	normal
1	0	1	1	1	1	land
0.965	0.001	0.991	0.965	0.989	0.995	neptune
1	0	0.979	1	0.989	1	ipsweep
0.167	0	1	0.167	0.286	0.954	teardrop
0.917	0.001	0.953	0.917	0.935	0.977	postswEEP
1	0	0.976	1	0.988	1	pod
0.996	0.002	0.986	0.996	0.991	0.999	guess_passwd
1	0	1	1	1	1	map
0.915	0.01	0.854	0.915	0.883	0.992	satAn
1	0	0.992	1	0.996	1	smurf
0.389	0.001	0.5	0.389	0.438	0.772	multihop
0.969	0.001	0.972	0.969	0.971	0.996	back
0	0	0	0	0	0.481	ftp_write
0.6	0.001	0.48	0.6	0.533	0.917	buffer_overflow
0	0	0	0	0	0.486	imap
0	0	0	0	0	0.489	phf
0.385	0.001	0.357	0.385	0.37	0.922	rootkit
0.988	0.003	0.969	0.988	0.979	0.997	warezmaster
0.992	0.005	0.775	0.992	0.672	0.962	saint
0.997	0.003	0.965	0.997	0.976	0.993	masen
0.986	0.001	0.98	0.986	0.983	0.997	apache2
0.955	0.001	0.939	0.955	0.947	0.991	smappgetattack
0.999	0	0.996	0.999	0.997	0.999	processstable
0.962	0.001	0.914	0.962	0.938	0.989	htptunnel
0.267	0	0.571	0.267	0.364	0.821	ps
0.967	0.005	0.849	0.967	0.904	0.991	smaguess
1	0.001	0.98	1	0.99	1	mailbomb

Figure 4: Detailed Accuracy by Class



## VI CONCLUSION

In this paper, we use NSL – KDD datasets for classification, it concludes datasets named as KDDTest+, KDDTrain+, KDDTest + Attacks and KDDTrain + Attacks datasets. We applied various Classification algorithms to detect intrusion detection in a NSL- KDD datasets. These classification algorithms are implemented on NSL – KDD datasets to detect the network intrusion detection in the datasets. Based on the Accuracy rate of the correctly classified instances and incorrectly classified instances of the datasets, we desire that which is the best Classification algorithm for the NSL – KDD datasets. Based on this accuracy rate Random Forest Classification Algorithm may get a high accuracy rate when comparing to other classification algorithms. As a result from the NSL- KDD datasets, the correctly classified instances of Random Forest Classification Algorithm are **99.4879%**, **96.2785%**, **97.9729%** .From these accuracy rates, we can decided that Random Forest Classification Algorithm is the best classification algorithm for the NSL – KDD datasets.

Classification	KDD Test+		KDD Train+		KDD Test 21+	
	Total Instances - 11850		Total Instances - 25192		Total Instances - 22544	
	Correctly Classified Instances	Incorrectly Classified Instances	Correctly Classified Instances	Incorrectly Classified Instances	Correctly Classified Instances	Incorrectly Classified Instances
Random Forest	25063	129	11409	441	22037	457
	99.4879%	0.512%	96.2785%	3.7215%	97.9729%	2.0271%
J48	24976	216	11322	528	21979	565
	99.1426%	0.8574%	95.5443%	4.455%	97.4963%	2.5062%
NaiveBayes	23585	1607	10408	1442	22087	457
	93.621%	6.379%	87.8312%	12.1688%	91.8648%	2.0271%
ZeroR	10408	1442	13449	11743	20710	1834
	87.8312%	12.1688%	53.368%	46.614%	91.8648%	8.1352%

**TABLE III. Accuracy Rate of Classification Algorithm**

## REFERENCES

### Journal Papers

- [1] Sakshi and Prof. Sunil Khare,” A Comparative Analysis of Classification Techniques on Categorical Data in Data Mining” International Journal on Recent and Innovation Trends in Computing and Communication, *IJRITCC* | August 2015, *Volume: 3 Issue: 8*, ISSN: 2321-8169, 5142 – 5147.
- [2] Delveen Luqman Abd AL-Nabi and Shereen Shukri Ahmed “Survey on Classification Algorithms for Data Mining :( Comparison and Evaluation) “Computer Engineering and Intelligent Systems, ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online), Vol.4, No.8, 2013.

- [3] Sunita B. Aher and Lobo L.M.R.J.”COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS” *International Journal of Information Technology and Knowledge Management*, July-December 2012, *Volume 5, No. 2, pp. 307-310.*
- [4] Abdelaziz Araar, Rami Bouslama “A comparative study of classification models For detection in ip networks intrusions” *Journal of Theoretical and Applied Information Technology* 10<sup>th</sup> June 2014. *Vol. 64 No.1*© 2005 - 2014 JATIT & LLS.
- [5] G.Kesavaraj and Dr.S.Sukumaran“A Comparison Study on Performance Analysis of Data Mining Algorithms in Classification of Local Area News Dataset using WEKA Tool “*International Journal of Engineering Sciences & Research Technology*, October, 2013, ISSN: 2277-9655, Impact Factor: 1.852.
- [6] Amit Gupta, Azeem Mohammad, Ali Syed and Malka N. Halgamuge “A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA” (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, *Vol. 7, No. 7, 2016.*
- [7] P.Amudha, S.Karthik and S.Sivakumari” Classification Techniques for Intrusion Detection – An Overview” *International Journal of Computer Applications (0975 – 8887) Volume 76– No.16, August 2013.*
- [8] L.Dhanabal, Dr. S.P. Shantharajah “A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms” *International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, June 2015 Copyright to IJARCCCE DOI 10.17148/IJARCCCE.2015.4696.*
- [9] Rajesh Wankhede, India Vikrant Chole” *Intrusion Detection System using Classification Technique* ” *International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016, 25.*

### **Websites**

- [1] [https://en.wikipedia.org/wiki/Intrusion\\_detection\\_](https://en.wikipedia.org/wiki/Intrusion_detection_)
- [2] WEKA Tool downloading from <http://www.cs.waikato.ac.nz/ml/weka/> .
- [3] NSL-KDD datasets available at: <http://www.unb.ca/civ/research/datasets/nsl.html>