International Journal of Economic Research
Indicator 2008-2015 Value
SJR 0.12
Cites per doc 0.04
Total cites 4
www.scimagojr.com

International Journal of Economic Research

ISSN : 0972-9380

available at http: www.serialsjournals.com

© Serials Publications Pvt. Ltd.

Volume 14 • Number 16 (Part 3) • 2017

Detection of Attacks Using Machine Learning techniques

Udayakumar N[a] Siddharth Pukhraj Lakhara[a] and T. Subbulakshmi[a]

[a]*SCSE, VIT University Chennai, Tamilnadu, India*

E-mail: udayakumar.n2014@vit.ac.in, spukhraj.lakhara2014@vit.ac.in, subbulakshmi.t@vit.ac.in

Abstract: The target of this work is to search out the most effective among the ten classification algorithms thought of to classify the association records into normal or abnormal within the KDDCup20% training data set with weka tool. During this work, the experiment is done by applying of ten classification algorithms on the KDD Cup 20% training dataset comprising of 25192 instances through an experiment kind of 10-fold cross validation. The Comparison fields Percent_correct, fmeasure, recall, precision and ROC (area under roc) were taken for analysis. Tests were additionally performed for ranking and outline. As per the results obtained by the weka Experimenter with the 10 classifiers on the KDD 20% training dataset, it's been analysed that Random forest classifier works best with the comparison fields percent_correct, fmeasure and ROC (Area underneath ROC). Simplecart classifier ranks next to Randomforest Classifier with the comparison fields percent_correct and measure. Simplecart classifier outperforms all alternative classifiers with reference to the comparison field precision. ZeroR is found to be the worst classifier in terms of all the comparison fields except recall. Therefore it's been found that with the dataset that's taken for experiment, additional elaborated study could be restricted only with the 5 classifiers particularly Random Forest, simple cart, J48, bagging and IBk. This may positively reduce process time and increase the potency of classification of the KDDCup20% data set.

Keywords : *Classifier, KDDCup, ROC, ZeroR, SimpleCart, J48, Bagging, IBk.*

1. INTRODUCTION

Intrusion is outlined as any set of action that may compromise the integrity, confidentiality and availableness of system resources. There are two kinds of intrusion detection specifically, misuse detection and anomaly detection. Misuse detection refers to the identification of the already best-known intrusion patterns within the dataset. better-known attack patterns are simply known victimization their signatures in misuse detection models. they're additionally known as as Signature based Intrusion Detection Systems (IDS). Signature based IDS are unable to find unknown and latest attacks since signature information needs to be manually revised for any new attack. Another type named Anomaly detection refers to the detection of novel intrusion patterns in information[2][3]. This could be utilized for determine best-known and unknown attacks[4].

There are many papers that handle Intrusion detection in numerous angles. Intrusions are usually tough to spot since there are numerous threats that don't seem to be real intrusions. Sometimes, user might not be able to determine original intrusion[4-6]

2. INTRUSION DATASET

The KDD Cup dataset[7] is taken into account to be the benchmark data in Intrusion detection. The dataset was a set of simulated raw transmission control protocol dump data over a time of 9 weeks on LAN. The well-known attack types are those present within the training dataset whereas the novel attacks are the additional attacks that aren't present within the training dataset. There are numerous attacks like Buffer overflow, Perl, Port sweep, Neptune, Smurf, Teardrop, Guess password, ip Sweep etc. The training dataset consists of 4,94,021 records. The testing dataset consists of 3,11,029 records. In every association record there are forty-one attributes describing totally different options of the connection. within the training dataset, alongside the forty-one attributes a category attribute is additionally given. In our study, we've taken KDD Cup 20% training dataset for experimenting with multiple classifiers. Different types of attacks present in KDD Cup 20% data and its count has been showed in the Table 1, Protocol distribution has been showed in Table 2 and in Figure 1.

Table 1
Attack Distribution in KDDCUP 20% dataset

S.No	*Attack type*	*Count*
1	Back	2203
2	teardrop	979
3	loadmodule	9
4	neptune	107201
5	Rootkit	10
6	Phf	4
7	Satan	1589
8	buffer_overflow	30
9	ftp_write	8
10	Land	21
11	Spy	2
12	ipsweep	1247
13	multihop	7
14	Smurf	280790
15	Pod	264
16	Perl	3
17	warezclient	1020
18	Nmap	231
19	Imap	12
20	warezmaster	20
21	portsweep	1040
22	Normal	97277
23	guess_passwd	53

Table 2
Protocol Distribution in KDD Cup 20% dataset

Protocol	*Count*
TCP	190064
UDP	20354
ICMP	283602

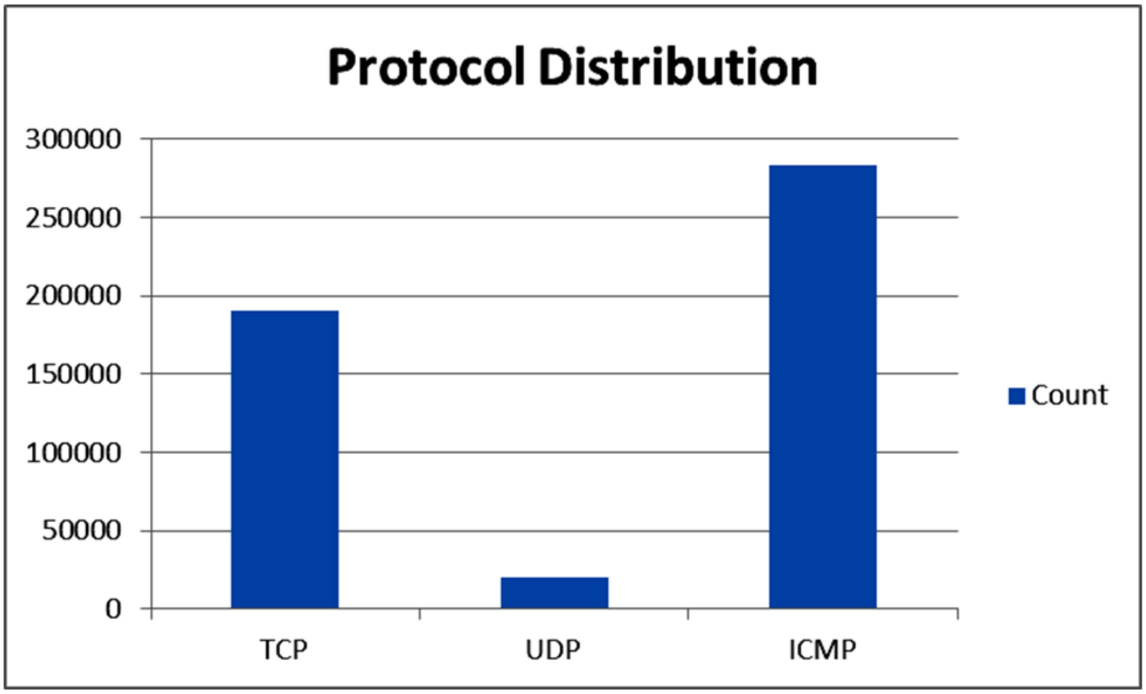

Figure 1: Protocols in KDD Cup 20% data set

3. USE OF DATA MINING IN CLASSIFICATION

Data mining could be a finding method of serious non-intuitive correlations and patterns, creating attainable to urge high level information data from low level information. data processing is additionally data Discovery in information. It is the non-trivial method of distinctive valid and novel helpful and perceivable patterns of information. data processing is more than collection of information. It involves analysis and predictions. Classification may be a data processing task that maps the information into predefined groups and categories. it's conjointly referred as supervised learning. It consists of two steps. initiative is that the model construction that consists of set of preset categories. every tuple is assumed to belong to a predefined category. The set of tuple used for model construction is coaching set. The model is portrayed as classification rules, call trees, or mathematical formulae. Second step is model usage that is employed for classifying future or unknown objects. The well-known label of check sample is compared with the classified result from the model[8][9].

4. ALGORITHMS USED FOR CLASSIFICATION.

Classifiers like OneR, ZeroR, BayesNet, NaiveBayes, IBk, Adaboost, Meta bagging, J48, Random forest and Simple cart are utilized in this paper. a quick note regarding the various classifiers utilized in this paper is given below:

4.1. OneR Classifier

OneR classifier short for 'One Rule' could be a straightforward, nevertheless correct, classification algorithmic rule. It generates one rule for each predictor within the data, and so selects the rule with the tiniest total error as its "one rule". OneR produces rules only slightly less correct than state of the art classification algorithms however produces rules that are easy for humans to interpret.

4.2. Zero R

ZeroR is that the simplest classification technique that depends on the target and ignores all predictors. ZeroR classifier simply predicts the majority class (class). Although there is no certainty power in ZeroR, it's helpful for determining a baseline performance as a benchmark for other classification ways.

4.3. Bayesian Network

A Bayesian Network, Bayes Network, Bayesian Model or Probabilistic directed acyclic graphical model could be a probabilistic graphical model that represents a collection of random variables and their conditional dependencies via a Directed Acyclic Graph (DAG).

4.4. Naive Bayesian

The Naive Bayesian classifier relies on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is simple to create, with no difficult reiterative parameter estimation that makes it significantly helpful for terribly large datasets. Despite its simplicity, the Naive Bayesian classifier usually will astonishingly well and is wide used as a result of it usual outperforms a lot of refined classification ways[9].

4.5. Ibk

The *k*-Nearest Neighbour (*k*-NN) may be a methodology for classification of objects supported the nearest training examples within the feature area. *k*-NN may be a kind of instance primarily based learning or lazy learning. The *k*-NN is one among the best of machine learning algorithms. an object is assessed by a majority vote of its neighbours, with the item being allotted to the category most typical amongst its *k* nearest neighbours (*k* is a positive number, generally small). If $k = 1$, then the item is solely allotted to the category of its nearest neighbour.

4.6. Adaboost Classifier

Bagging and Boosting are Meta algorithms that pool choices from multiple classifiers. This algorithmic rule iteratively learns from weak classifiers. The ultimate result's the weighted total of the results of weak classifiers.

4.7. MetaBagging classifier

Bagging generates bootstrap samples of the training data. Then it trains a classifier or a regression operation using every bootstrap sample. For classification purpose, the bulk vote on the classification results is taken. the common on the expected values is taken for regression. The advantage of bagging is that it reduces variation and it improves performance for unstable classifiers that vary considerably with little changes within the dataset.

4.8. J48 Classifier

J48 is slightly changed C4.5 in WEKA. The C4.5 algorithmic rule generates a classification-decision tree for the given data-set by recursive partitioning of information. the choice is fully grown using Depth-first strategy. The algorithmic rule considers all the attainable checks that may split the info} set and selects a test that provides the most effective information gain. for every separate attribute, one check with outcomes as several because the range of distinct values of the attribute is taken into account. for every continuous attribute, binary tests involving each distinct values of the attribute are thought of. so as to collect the entropy gain of all these binary tests with efficiency, the training data set belonging to the node in consideration is sorted for the values of the continual attribute and therefore the entropy gains of the binary cut supported every distinct values are calculated in one scan of the sorted data. This method is continual for every continuous attributes.

4.9. Random Forest Classifier

Random forests are an ensemble learning technique for classification, regression and different tasks that operate by constructing a large number of call trees at training time and outputting the category that's the mode of the categories (classification) or mean prediction (regression) of the individual trees.

4.10. Simple cart Classifier

Simple Cart (Classification and Regression tree) could be a classification technique that generates the binary call tree. Since output could be a binary tree, it generates solely two children. Entropy is employed to decide on the most effective splitting attribute. simple Cart handles the missing knowledge by ignoring that record.

5. INTRODUCTION TO WEKA

WEKA stands for Waikato environment for Knowledge Learning. it had been developed by the University of Waikato, New Zealand. weka could be a collection of machine learning algorithms for data processing tasks. The algorithms will either be directly applied to the dataset or referred to as from java code. weka contains tools for knowledge pre-processing, classification, regression, clustering, association rules and visualization[10]. it's well matched for developing new machine learning schemes. The dataset employed in rail is to be within the .ARFF format. this kind of file consists of a header which describes the attribute varieties and an information section that may be a comma separated list of data.

WEKA tool contains of four buttons specifically,Explorer, Experimenter, knowledge Flow and simple CLI. Explorer is an environment for exploring knowledge with weka. Experimenter is an environment for performing experiments and conducting statistical tests between learning schemes. Knowledge Flow is an environment that supports basically a similar functions because the explorer however with a drag-and-drop interface. One advantage is that it supports progressive learning. easy command line interface Provides an easy command-line interface that enables direct execution of weka commands for operating systems that don't offer their own command interface[11]

6. EXPERIMENTS DONE

The above mentioned ten classifiers are applied to the KDD Cup 20% training dataset comprising of 25192 instances through an experiment kind of 10-fold cross validation. The experiment specifically took six hours to finish. The tests were organized with Paired T Tester (corrected) and also the test of significance was taken as zero.05. The comparison fields Percent_correct, fmeasure, recall, precision and ROC(area under roc) were taken for analysis.

6.1. Analysis of ZeroR Classifier

The results show that all the classifiers are statistically better than the baseline classifier ZeroR at the significance level specified 0.05. It is also found that all the classifiers are better than ZeroR once and never equivalent to or worse than ZeroR (1/0/0) with difference comparison fields is shown in Table 3

Table 3
ZeroR Classifier Results with different comparison fields

Comparison Field	*Value*
tp rate	0.534
fp rate	0.534
Precision	0.285
Recall	0.534
F-Measure	0.372
ROC Area	0.5

6.2. Analysis of OneR Classifier

The results show that the classifiers IBk, Bagging, J48, Random Forest and SimpleCart are statistically works well than the baseline classifier OneR at the connotation level fixed 0.05. The classifiers ZeroR, NaiveBayes, AdaBoost are statistically worse than OneR classifier. it's additionally determined that there's no statistical distinction between OneR and BayesNet classifier. it's additionally shown from the results that the classifiers IBk, Bagging, J48, Random Forest and SimpleCart are good than OneR once and never similar to or worse than OneR. (1/0/0). The classifiers ZeroR, NaiveBayes and AdaBoost are not better than OneR. (0/0/1). OneR with different comparison fields is shown in Table 4.

Table 4
OneR Classifier Results with different comparison fields

Comparison Field	*Value*
tp rate	0.963
fp rate	0.034
Precision	0.964
Recall	0.963
F-Measure	0.963
ROC Area	0.964

6.3. Analysis of Bayesnet classifier

Table 5
Bayesnet Classifier Results with different comparison fields

Comparison Field	*Value*
tp rate	0.966
fp rate	0.038
Precision	0.967
Recall	0.966
F-Measure	0.966
ROC Area	0.996

The results show that the classifiers IBk, Bagging, J48,Random Forest and SimpleCart are statistically provides the better performance than the baseline classifier BayesNet at understand level specified 0.05.The classifiers ZeroR, NaiveBayes, AdaBoost are statistically worse than BayesNet classifier. it's also determined that there's no statistical contrast between BayesNet and OneR classifier. it's further shown from the results that the classifiers IBk, Bagging, J48, Random Forest and SimpleCart are improved than BayesNet once and never similar to or worse than BayesNet. (1/0/0). The classifiers ZeroR, NaiveBayes and AdaBoost don't seem to be superior than BayesNet. (0/0/1). BayesNet with completely different comparison fields is shown in Table 5.

6.4. Analysis of Naïve bayes classifier

The results show all the classifiers except ZeroR are statistically works well than the baseline classifer NaiveBayes at connotation level given 0.05. it's additionally found that all the classifiers except ZeroR are works well than NaiveBayes once and never similar to or worse than NaiveBayes. (1/0/0) ZeroR isn't good than NaiveBayes. (0/0/1). NaiveBayes with completely different comparison fields is shown in Table 6 .

Table 6
Bayesnet Classifier Results with different comparison fields

Comparison Field	*Value*
tp rate	0.896
fp rate	0.106
Precision	0.896
Recall	0.896
F-Measure	0.896
ROC Area	0.966

6.5. IBk Classifier

The results show that the classifiers bagging, J48, Random Forest and SimpleCart are statistically performs well than the baseline classifier IBk at the connotation level given 0.05. The classifiers ZeroR, OneR, BayesNet, NaiveBayes and AdaBoost are statistically worse than IBk classifier. it's additionally shown from the results that the classifiers bagging, J48, Random Forest and simple Cart are better than IBk once and never adore or worse than IBk. (1/0/0). The classifiers ZeroR, OneR, BayesNet, NaiveBayes and AdaBoost don't seem to be better perform than BayesNet. (0/0/1). IBk with completely different comparison fields is shown in Table7 .

Table 7
IBk Classifier Results with different comparison fields

Comparison Field	*Value*
tp rate	0.994
fp rate	0.006
Precision	0.994
Recall	0.994
F-Measure	0.994
ROC Area	0.994

6.6. Adaboost classifier

The results show that the classifiers OneR, BayesNet, IBk, Bagging, J48, Random Forest and SimpleCart ar statistically works well than the baseline classifier AdaBoost at the connotation level fixed 0.05. The classifiers ZeroR and NaiveBayes are statistically worse than AdaBoost classifier. it's conjointly shown from the results that the classifiers OneR, BayesNet, IBk, Bagging, J48, Random Forest and simple Cart are better than AdaBoost once and never like or worse than AdaBoost. (1/0/0). The classifiers ZeroR and NaiveBayes don't seem to be better performing than AdaBoost. (0/0/1). AdaBoost with totally different comparison fields is shown in Table 8

Table 8
Adaboost Classifier Results with different comparison fields

Comparison Field	*Value*
tp rate	0.944
fp rate	0.059
Precision	0.944
Recall	0.944
F-Measure	0.944
ROC Area	0.988

6.7. Bagging classifier

The results show that the classifier Random Forest is statistically higher than the baseline classifier bagging at the importance level given 0.05. The classifiers ZeroR, OneR, BayesNet, NaiveBayes, IBk, AdaBoost are statistically worse than bagging classifier. it's conjointly ascertained that there's no statistical distinction between J48 and SimpleCart and bagging classifier. it's conjointly shown from the results that the classifier Random Forest is best than bagging once and never similar to or worse than bagging. (1/0/0). The classifiers ZeroR, OneR, BayesNet, NaiveBayes, IBk, AdaBoost aren't better than bagging. (0/0/1). bagging with totally different comparison fields is shown in Table 9.

Table 9
Bagging Classifier Results with different comparison fields

Comparison Field	*Value*
tp rate	0.996
fp rate	0.004
Precision	0.996
Recall	0.996
F-Measure	0.996
ROC Area	0.999

6.8. J48 Classifier

The results show that the classifier Random Forest is statistically good than the baseline classifier J48 at the connotation level such that 0.05. The classifiers ZeroR, OneR, BayesNet, NaiveBayes, IBk, AdaBoost are statistically worse than J48 classifier. It is additionally discovered that there's no statistical distinction between bagging and simple Cart compared to the baseline classifier J48. it's conjointly shown from the results that the classifier Random Forest is best than J48 once and never like or worse than J48. (1/0/0). The classifiers ZeroR, OneR, BayesNet, NaiveBayes, IBk, AdaBoost don't seem to be good than J48. (0/0/1). J48 with completely different comparison fields is shown in Table 10.

Table 10
J48 Classifier Results with different comparison fields

Comparison Field	*Value*
tp rate	0.996
fp rate	0.004
Precision	0.996
Recall	0.996
F-Measure	0.996
ROC Area	0.998

6.9. Random forest classifier

The results show that none of the classifiers area unit statistically excellent than the baseline classifier Random Forest at drift level such 0.05. The classifiers ZeroR, OneR, BayesNet, NaiveBayes, IBk, AdaBoost, Bagging, J48 are statistically worse than Random Forest classifier. it's conjointly determined that there's no statistical distinction between SimpleCart and Random Forest. it's conjointly shown from the results that the classifiers ZeroR, OneR, BayesNet, NaiveBayes, IBk, AdaBoost, bagging and J48 don't seem to be excellent than Random Forest. (0/0/1). Random forest with totally different comparison fields is shown in Table 11.

Table 11
Random forest Classifier Results with different comparison fields

Comparison Field	*Value*
tp rate	0.998
fp rate	0.002
Precision	0.998
Recall	0.998
F-Measure	0.998
ROC Area	1

6.10. Simple cart classifier

The results show that none of the classifiers are statistically excellent than the baseline classifier simple Cart at connotation level stated 0.05. The classifiers ZeroR, OneR, BayesNet, NaiveBayes, IBk and AdaBoost are statistically worse than simple Cart classifier. it's additionally ascertained that there's no statistical distinction exists between bagging, J48, RandomForest and SimpleCart classifier. it's additionally shown from the results that the classifiers ZeroR, OneR, BayesNet, NaiveBayes, IBk and AdaBoost aren't excellent than simple Cart. (0/0/1). SimpleCart with totally different comparison fields is shown in Table 12.

Table 12
Random forest Classifier Results with different comparison fields

Comparison Field	*Value*
tp rate	0.997
fp rate	1
Precision	0.004
Recall	0.997
F-Measure	0.997
ROC Area	0.997

7. COMPARATIVE RESULTS

7.1. TP Rate

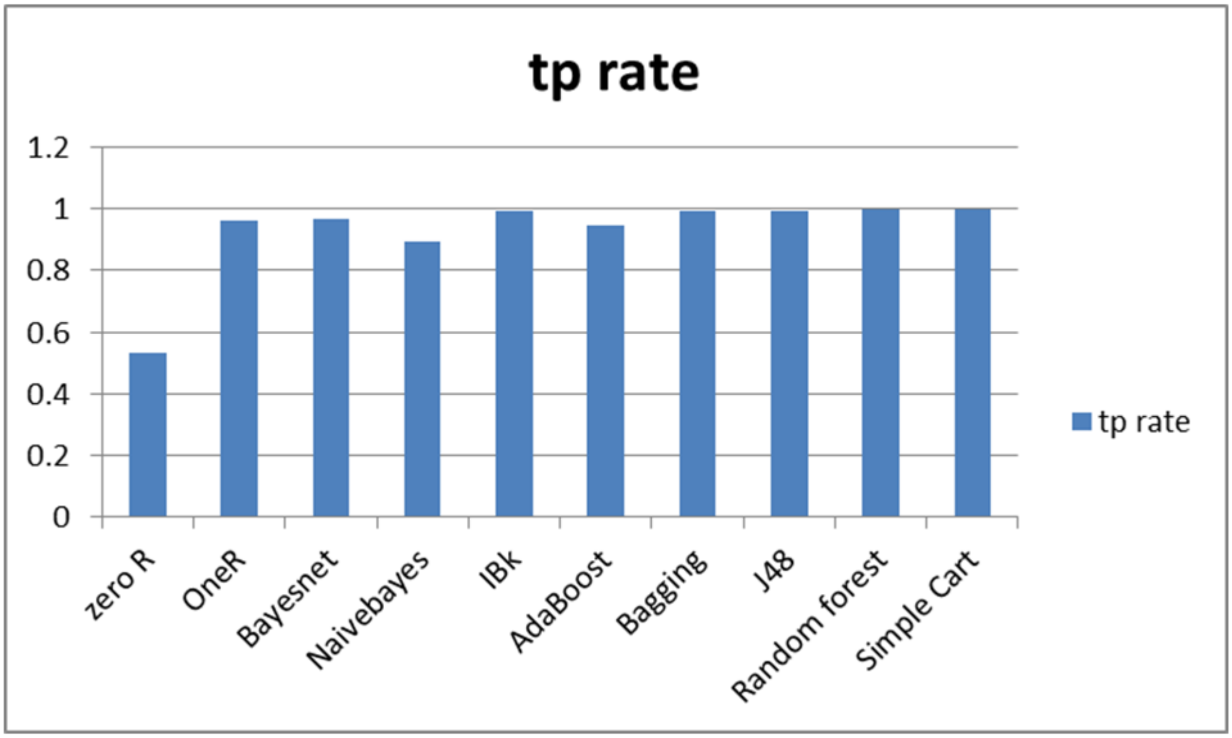

Figure 2: TP Rate result of different classifiers

Figure 2 shows that Zero R Classifiers produce the lowest true positive rate of 0.534 and the random forest provides the better True positive rate of 0.998. Classifiers simple cart , j48 and bagging provides the next better result.

7.2. FP Rate

Figure 3 Provides false positive rate results in the form of graphical representation for various classifiers used for analysis, among them Simple cart provides the value of 1 and Random forest provides the best result with the value of 0.002.

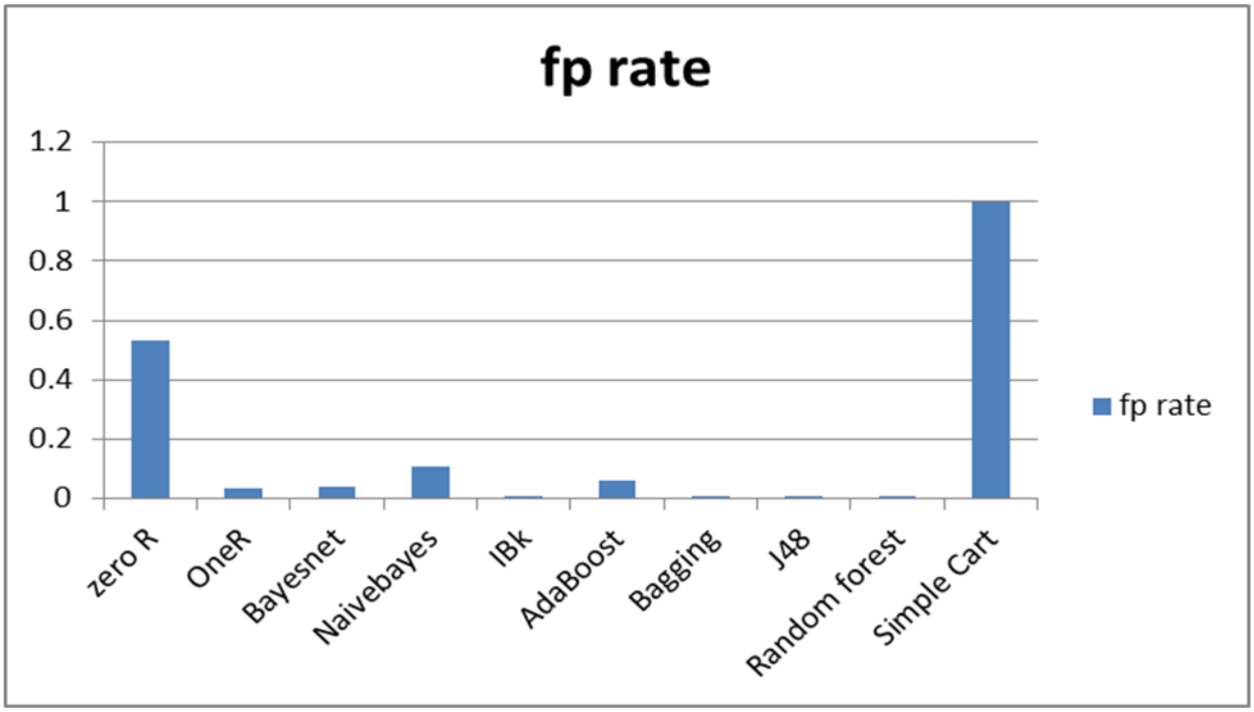

Figure 3: FP Rate result of different classifiers

7.3. Precision

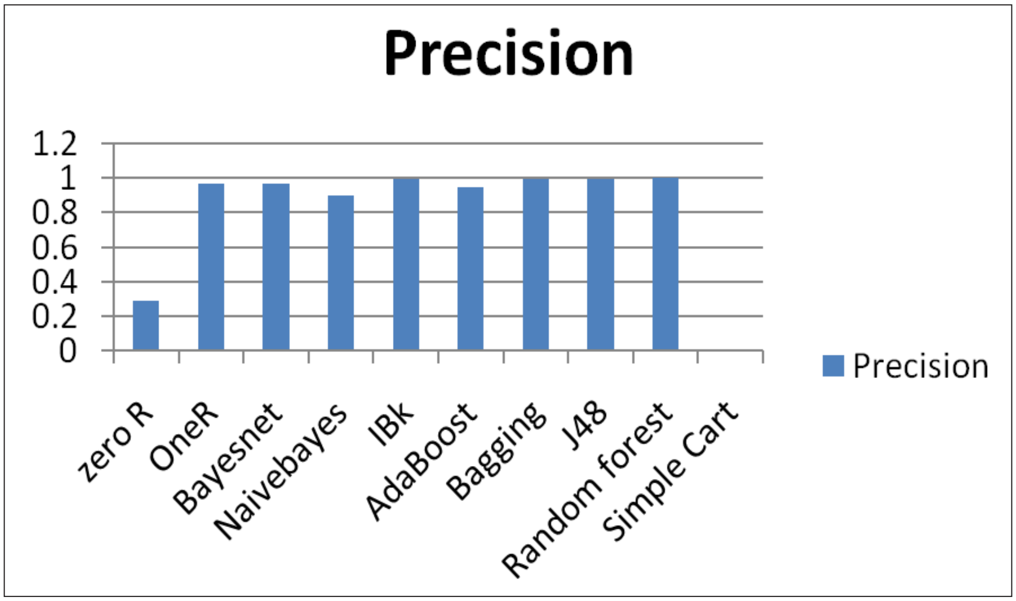

Figure 4: Precision Rate result of different classifiers

Figure 4 Provides precision rate results in the graphical representation for various classifiers used for analysis, among them Simple cart provides the lower value of 0.004 and Random forest provides the highest value of 0.998.

7.4. Recall

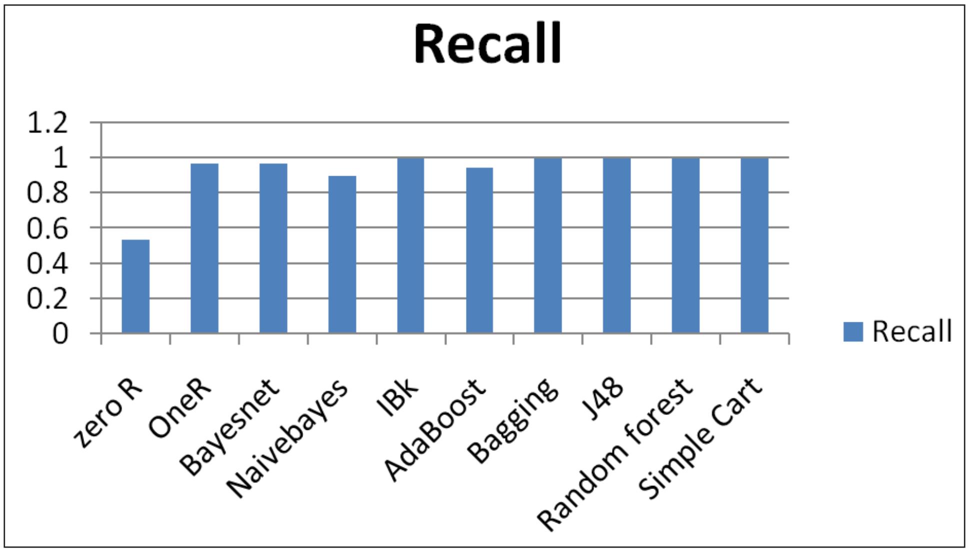

Figure 5: TP Rate result of different classifiers

Figure 5 shows the Recall result comparison of classifiers used for analysis, among all zeroR provides the lowest value of 0.534 and the Random forest provides the highest value of 0.998

7.5. F-Measure

F-measure value depends on the value of Precision (Positive predictive value) and the recall value also said to be as sensitivity. Its the average of precision and recall when they are close and generally said to be as harmonic mean.

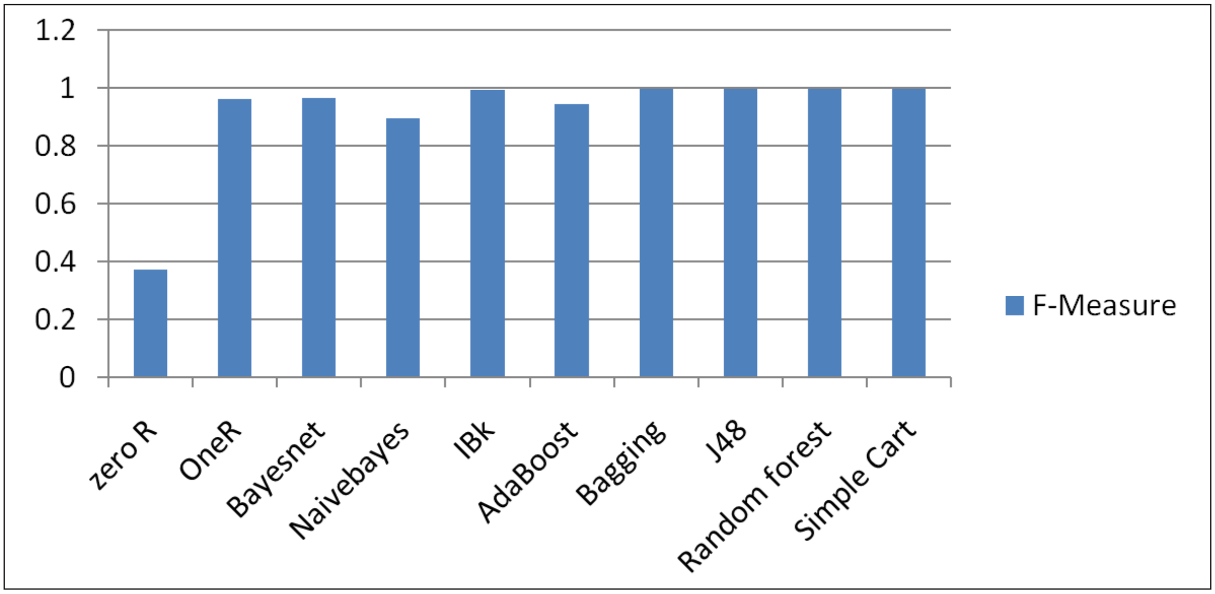

Figure 6: F-Measure result of different classifiers

F-measure comparative results used for analysis has shown in Figure 6. among the classifiers used for analysis Random forest provides the highest value of 0.998 and the ZeroR provides the lowest value of 0.372.

8. CONCLUSION

As per the results obtained by the weka Experimenter with the ten classifiers on the KDD 20% training dataset, it's been analysed that Random forest classifier works best with the comparison fields tp rate, fp rate, percent_ correct, fmeasure and ROC (Area under ROC). Simple cart classifier ranks next to Random forest classifier with the comparison fields percent_correct and fmeasure. Simple cart classifier outperforms all different classifiers with regard to the comparison field precision. ZeroR is found to be the worst classifier in terms of all the comparison fields except recall. With recall because the comparison field, ZeroR ranks 1st when put next to any or all another classifiers. J48 classifier stands next to simple cart classifier with the comparison field's percent_correct, fmeasure and precision. During this analysis work, the 10 types of classifiers are applied on the KDD20% training dataset with distinct comparison fields using the weka tool experimenter.

Thus it's been found that with the dataset that's taken for experiment, more elaborated study may be restricted alone with the 5 classifiers Random Forest, Simplecart, J48, bagging and IBk. this may positively reduce process time and increase the potency of classification of data set. Additionally the explanation for the ZeroR classifier's performance with recall comparison field needs to be studied.

REFERENCES

[1] Adetunmbi AO, Falaki SO, Adewale OS, Alese BK. Network Intrusion Detection based on Rough Set and k-Nearest Neighbour. International Journal of Computing and ICT Research. 2008; 2(1):60–6. Available from: http://www.ijcir.org/volume1number2/article7.pdf

[2] Ranjan R, Sahoo G. A new clustering approach for anomaly intrusion detection. International Journal of Data Mining and Knowledge Management Process. 2014 Mar; 4(2):29–38.

[3] Azad C, Jha VK. Data Mining based Hybrid Intrusion Detection System. Indian Journal of Science and Technology. 2014 Jun; 7(6):781–9.

[4] Khor K-C, Ting C-Y, Amnuaisuk S-P. From Feature Selection to Building of Bayesian Classifiers: A Network Intrusion Detection Perspective. American Journal of Applied Sciences 2009; 6(11):1948–59.

[5] Lee W, Stolfo SJ, Mok KW. Algorithms for Mining System Audit Data. Proc KDD; 1999

[6] Ghali NI. Feature Selection for Effective Anomaly Based Intrusion Detection. International Journal of Computer Science and Network Security. 2009 Mar; 9(3):285–9.

[7] KDD CUP 1999 DATASET: Available from: http://kdd.ics.uci.edu/databases/kddcup99/

[8] SANS Institute InfoSec Reading Room. Understanding Intrusion Detection Systems; 2001

[9] Wu X, Kumar V, Ross Quinlan RJ, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D. Top 10 algorithms in data mining. London: Springer-Verlag; 2008. p. 1–3. DOI: 10.1007/s10115-007-0114-2

[10] Weka Manual. Available from: http://www.ittc. ku.edu/~nivisid/WEKA_MANUAL.pdf