Comparative Analysis of Naïve Bayes and J48 Algorithms on Intrusion Detection System (IDS)

Priyanshi Kotlia
Graphic Era Hill University

Janmejay Pant
Department of Computer Science and Engineering
Graphic Era Hill University

ABSTRACT
Intrusion detection system (IDS) is the science of detecting malicious activity across the computer network, as it is expanding, there is a challenge to compete with the malicious users or intruders who can easily break into the system .This paper put a light on performance analysis by J48 and Naïve Bayes algorithm to detect the error in intrusion detection system (IDS). Where Naïve bayes algorithm is based on probability and j48 algorithm is based on decision tree. The paper set out to make the comparative study of algorithms Naïve Bayes and J48 in the IDS data set to maximize the True Positive Rate and minimize the False Positive Rate using the WEKA Tool.

The paper is showing the experimental result about classification of accuracy, sensitivity and specificity on data set of IDS and also shows that J48 algorithm is much better than that of Naïve Bayes algorithm in terms of precision and accuracy.

Keywords
Naïve Bayes, J48 decision tree, NSL-KDD, confusion matrix, true positive rate, false positive rate and ROC Curve.

1. INTRODUCTION
Intrusion detection is used to classify normal and anomaly activities in which machine learning can play a vital role. Machine learning based intrusion detection approaches can detect both misuse and anomaly that's why machine learning-based intrusion detection approaches have been subjected to extensive researches [7].

As intrusion detection system has became an important tool for security in many computer network field , because rather than providing security to the network they monitor and analysis each and every packet over the network .To check either those packets are attempt to compromise with the confidentiality , availability and integrity of the data which were send over the network or not[1] .So, basically IDS is a system which protect or secure a network system via monitoring the network and automatically detect the malicious activity over the network or internet.

There were basically 2 types of IDS-Misuse Detection and Anomaly Detection.

1.1 Misuse Detection
This detection system firstly monitor the network for the packets which were send over the network and generate the pattern for the malicious behavior and then identify the intrusion based on those patterns[2].

So, basically this detection is useful because of its higher detection rate for all the attacks which were known but it cannot detect that intrusion for which patterns were not known.

1.2 Anomaly Detection
This detection system have in advance the expected behavior of all the packets and then they monitor each packet over the internet and map with their original one and if both are match it means its a normal packet otherwise there will some chances of attack but unlike of previous detection system they can even identify the unknown attack too but they have low detection rate [2].

2. DATA CLASSIFIERS
Actually classification can be done or performed on either structured data or unstructured data or both. In this, we categorize the data into number of classes or category and then identify which class or category the new data will fall under. And classifier is defined as the algorithm which maps the input data to a specific category

In this, paper we are using 2 classifier i.e. the naïve bayes and J48 algorithm for comparison .The comparison among these algorithm is for accuracy, specificity and sensitivity using true positive rate and false positive rate which is generated by confusion matrix of respective algorithm [3].

2.1 Naïve Bayes Algorithm
The naïve bayes algorithm is based on bayes theorem and that's why it is also called as conditional theorem. This algorithm is called Naïve because it make an assumption that occurrence of a particular feature is totally independent of the occurrence of other features. The motive of using this algorithm is that it require small amount of training data to estimate the necessary parameters .This classifier algorithm works well in real world situations as well.

Algorithm [6]

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

When P (A|B) is called Posterior probability which represent the degree to which we believe that a given model accurately describes the situation and all our prior information. P(B|A) is called likelihood which describes how well the model predict the data, P(A) is called Prior probability which describes the degree to which we believe the model accurately describes reality based on all of our prior information and P(B) is called Normalizing constant , the constant that makes the posterior density integrate to 1.

2.2 J48 Decision Tree Algorithm
J48 is an algorithm used to generate a decision tree which is generated by C4.5, and C4.5 algorithm is an algorithm which is used in data mining as a Decision tree classifier algorithm which can be employed to generate a decision, based on a certain sample of data and J48 is an extension of ID3

(Iterative Dichotomiser 3) which is developed by WEKA Tool.

The feature of J48 algorithm over the Decision Tree algorithm is that J48 accounting for decision tree pruning , for missing values , continuous attribute value ranges, etc and in this algorithm the classification is done recursively until each and every single leaf node is pure and its main aim is to provide flexibility and accuracy over the decision tree.

Algorithm [3]

Input:
D //Training Data Set
Output:
T //Decision Tree
DTBUILD (*D)
{
T=ϕ;
T=Create root node and label it with splitting attribute;
T=Add arc to root node for each split predicate and label;
For each arc do
D=Database created by applying splitting predicate to D;
If stopping point is reached for this path then
T'=Create leaf node and label with appropriate class;
Else
T'=DTBUILD (D)
T=add T' to arc;
}

3. MEASURING PERFORMANCE

By evaluating the accuracy of classification algorithm, the performance of classification algorithm is examined. The main difference between this and the traditional approach is that, the traditional approach evaluates the time and space overhead which is secondary.

Accuracy can be defined by the % of correct prediction which can be explained as follows-

$$\textbf{Accuracy} = \frac{\textbf{No. of correct Prediction}}{\textbf{Total no. of Prediction}}$$

To visualize the performance of classification problem, we use AUC (Area under Curve) ROC curve which is one of the most important evaluation metrics for checking any classification model's performance and therefore can be written as AUROC curve .Which shows the relationship between TPR against FPR at various threshold setting. ROC is the probability curve and AUC represent the degree of separability. It tells how much model is capable of distinguishing between classes. Higher will be the AUC, better the model is predicting 0's as 0's and 1's as 1's [4].

3.1 Confusion Matrix

A confusion matrix is table which is used to describe the performance of classification algorithm on a set of test data for which the true value is known [5] .It shows the ways in which our classification algorithm is confused when it makes predictions, and it is extremely useful for measuring recall, precision, specificity, accuracy and most importantly AUC – ROC curve.

Some standard and terms

1. True positive- here observation is positive and predicted to be positive.

2. False positive-here observation is negative and predicted to be positive.

3. Precision –it is ratio of the predicted correctly over all the classes, i.e. TP over (TP+FP).

4. Recall- it is the ratio of the predicted correctly over the entire positive classes i.e. TP over (TP+FN).

Precision quality can be seen as a measure of exactness where as recall is a measure of completeness or quality. Recall is nothing but TPR (true positive rate) for the class.

In this paper, we have used WEKA (Waikato environment for knowledge analysis) tool for comparison of Naïve Bayes and J48 algorithm and calculating efficiency based on accuracy regarding correct and incorrect instances generated by confusion matrix [3].

4. EXPERIMENTAL WORK AND RESULT

Our experiment is based on NSL-KDD data set of intrusion detection [8]. NSL-KDD data set is used to perform the experiments through the WEKA. It consists of a good and reasonable proportion of various types of records [8].

We have performed classification using Naïve bayes algorithm and J48 algorithm on IDS data set in WEKA tool. WEKA tool provide inbuilt algorithm for Naïve bayes and J48.

A. Results of J48 Algorithm for classifier-

We have applied J48 algorithm in IDS data set to generate the confusion matrix for the intrusion attribute having 2 possible values i.e. normal or anomaly.

Confusion Matrix

A	B
12704	**34**
64	**11064**

Classified as a – Normal
b – Anomaly

from above matrix we conclude that for class **a** i.e. for normal True Positive is **12704** while False Positive is **34** whereas for class **b** i.e. for anomaly True Positive is **11064** while False Positive is **64.**

The diagonal element of matrix represents the correct instances i.e. **12704+11064=237.**and correct instances were 64+34=98.

TPR for class a = 12704 / (12704+34) = 0.9973
FPR for class a = 64 / (64+11064) = 0.0057
TPR for class b = 11064 / (64+11064) = 0.9942
TPR for class b = 34 / (34+12704) = 0.0026
Average of TPR = 0.9957
Average of FPR = 0.0041
Precision for class a = 12704 / (12704+64) = 0.9949
Precision for class b = 11064 / (11064+34) = 0.9969
F measure for class a = 2*0.9949*0.9973/ (0.9949+0.9973) = 0.9960
F measure for class b = 2*0.9969*0.9942/ (0.9969+0.9942) = 0.9955

Analysis Threshold Curve

In a ROC curve the TPR (sensitivity) is plotted against the FPR (specificity) for different cutoff points of a parameter .Each point on a ROC curve represent a sensitivity / specificity pair correspondingly to a particular decision threshold. The area under the ROC curve that is a AUC is a measure of how well a parameter can distinguish between normal and anomaly class.

1. Threshold curve for normal

By applying j48 algorithm in IDS data set we get to know that area under ROC curve for normal classes is about 99. 67%.

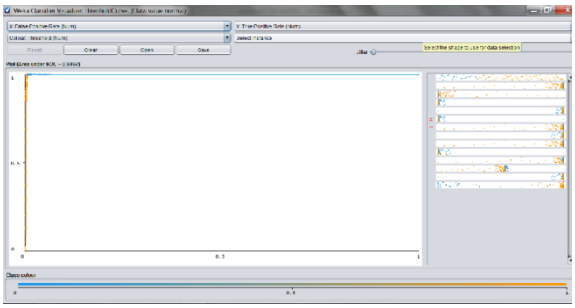

Fig 1: Shows the ROC Curve for Class Normal

2. Threshold curve for anomaly

By applying j48 algorithm in IDS data set we get to know that area under the ROC curve for anomaly classes is about 99.67%.

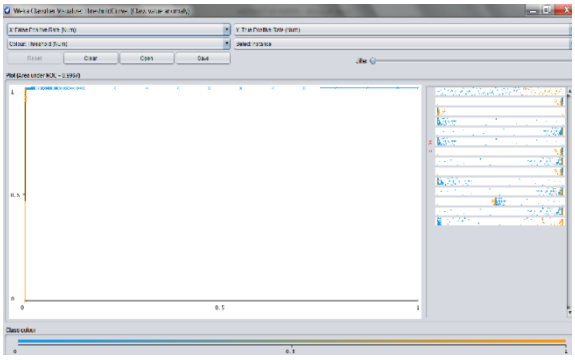

Fig 2: Shows the ROC Curve for Class Anomaly

B. Result of Naïve bayes algorithm for classifier

Now we have applied naïve bayes algorithm in IDS data set to generate the confusion matrix for the intrusion attribute having 2 possible vales i.e. Normal or Anomaly

Confusion matrix

A	B
11875	**863**
1481	**9647**

Classified as a = normal
b = anomaly

From above matrix we conclude that for class **a** i.e. for normal True Positive is **11875** while False Positive is **863** whereas for class **b** i.e. for anomaly True Positive is **9647** while False Positive is **1481.**

The diagonal element of matrix represents the correct instances i.e. **1185+9647=10832**and correct instances were **863+1481=2344.**

TPR for class a = 11875 / (11875+863) = 5.7983
FPR for class a = 1481 / (1481+9647) = 0.1330
TPR for class b = 9647 / (1481+9647) = 0.8669
TPR for class b = 863 / (11875+863) = 0.0677
Average of TPR = 3.3326
Average of FPR = 0.10035
Precision for class a = 11875 / (1481+11875) = 0.8891
Precision for class b = 9647 / (9647+863) = 0.91788
F measure for class a = 2*0.8891*5.7983/ (0.8891+5.7983) = 1.54178
F measure for class b = 2*0.9178*0.8669/ (0.9178+0.8669) = 0.89169

Analysis of Threshold Curve

1. Threshold curve for normal

By applying naïve Bayes algorithm in IDS data set we get to know that area under ROC curve for normal classes is about 96.29%.

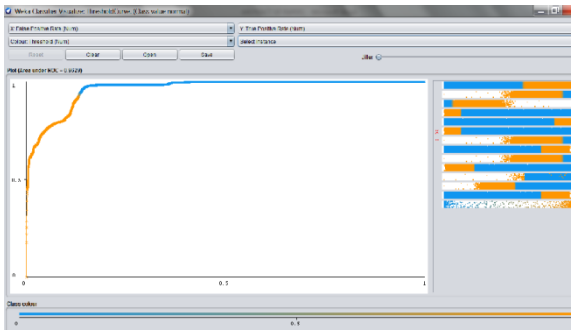

Fig 3: Shows the ROC Curve for Class Normal

2. Threshold curve for anomaly

By applying naïve bayes algorithm in IDS data set we get to know that area under ROC curve for anomaly classes is about 95.88%.

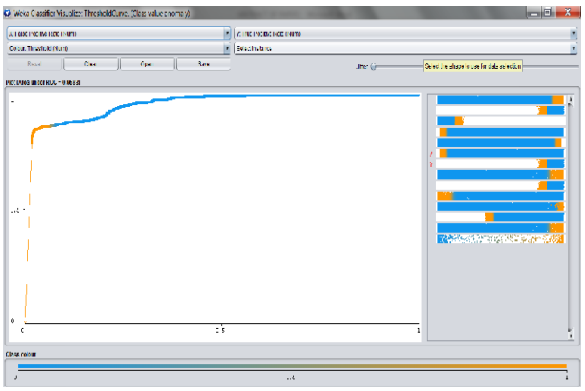

Fig 4: Shows the ROC Curve for Class Anomaly

5. CONCLUSION

From above experimental work we can conclude that by applying the Naïve bayes and J48 algorithm to the IDS data, set the data accuracy is more in J48 rather than naïve bayes for normal classes i.e. for J48 accuracy is 99.67% and in naïve bayes the accuracy is 96.29% as well as for anomaly classes the accuracy is again more in J48 algorithm rather than naïve bayes i.e. in J48 accuracy is 99.67% and in naïve bayes the accuracy is 95.88%.

CLASSIFICATION ACCURACY

	NAÏVE BAYES	J48
Normal	96.29%	99.67%
Anomaly	95.88%	99.67%
PRECISION		
	NAÏVE BAYES	J48
Normal	88.91%	99.49%
Anomaly	91.78%	99.69%

This proves that the, J48 is a simple classifier technique to make a decision tree. Efficient result has been taken from Intrusion Detection dataset using weka tool in the experiment. Naïve Bayes classifier also showing good results. The experiments results shown in the study is about classification accuracy and Precision. J48 gives more classification accuracy and Precision for class in IDS dataset having two classes Normal and Anomaly.

6. REFERENCES

[1] M.Revathi and T. Ramesh,"Network Intrusion Detection System", Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2, No.1 ISSN: 0976-5166.

[2] Kapil Wankhade, Sadia Patka and Ravindra Thool "An efficient approach for intrusion detection using data mining methods," 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI) , 978-1-4673-6217-7/13/$31.00 c 2013 IEEE.

[3] Tina R. Patil and Mrs. S. S. Sherekar,"Performance analysis of Naïve Bays and J48 Classification Algorithm for Data classification", International Journal of Computer Science and Applications Vol. 6, No.2, Apr 2013 ISSN: 0974-1011.

[4] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[5] https://www.dataschool.io/simple-guide-to-confusion-matrix-terminologys

[6] https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[7] Janmejay Pant, Kamlesh Padaliya and , Himanshu Pant," Rough Set Approach for Feature Selection in IDS," International Journal of Innovations & Advancement in Computer Science (IJIACS), ISSN 2347 – 8616 Volume 4, Special Issue September 2015.

[8] Janmejay Pant, Bhaskar Pant and Amit Juyal," Comparative Study of Different Models before Feature Selection and AFTER Feature Selection for Intrusion Detection," International Journal of Computer Applications (0975 – 8887), Volume 98– No.14, July 2014