Gain Ratio and Decision Tree Classifier for Intrusion Detection

Mabayoje Modinat A.
Department of Computer Science.
University of Ilorin, Ilorin.

Balogun Abdullateef O.
Department of Computer Science.
University of Ilorin, Ilorin

Akintola Abimbola G.
Department of Computer Science.
University of Ilorin, Ilorin.

Ayilara Opeyemi
Department of Computer Science.
University of Ilorin, Ilorin.

ABSTRACT
With the evident need for accuracy in the performance of intrusion detection system, it is expedient that in addition to the algorithms used, more activities should be carried out to improve accuracy and reduce real time used in detection. This paper reviews how data mining relates to IDS, feature selection and classification. This paper proposes architecture of IDS where GainRatio is used for feature selection and decision tree for classification using NSL-KDD99 dataset, It also includes the evaluation of the performance of the Decision tree on the dataset and also on the reduced dataset.

General Terms
Data mining, Intrusion detection, features reduction, and classification algorithms.

Keywords
Decision tree, IDS, Data Mining, Feature selection, data mining, and algorithms.

1. INTRODUCTION
In an age where the use of information is undoubtedly important as it contributes to our daily lives and deeds, data security and management is inevitably important; this undoubted and evident need of accurate information has led to the introduction of various technology in order to manage data efficiently and having secured and accurate.
Intrusion detection is a new field where data mining is currently gaining grounds and proving its potency in classification and detecting anomalies. It is a method of detecting and analyzing the events arising in a computer or network of computers to identify all security problems [15]. In another word, a form of security management system that collects and analyzes information from different areas in a computer or an arbitrary network of such device for identification of possible security breaches which include threats or attacks is regard to as Intrusion Detection. Intrusion Detection is of two types "Anomaly detection and Misuse detection" and there exists different categories of attacks which are Probing, (User to Root) U2R, Denial of Service (DOS) and (Root to User) R2L, each of these attacks have a way of disrupting the accuracy of data in use. In order to ensure accuracy it is necessary and important that the best techniques should be used, in which data mining has proven its potency and authenticity. Data mining involves non-trivial extraction of implicit potentially useful information and previously unknown from data in databases and information repositories [11]. Data mining can be used to build IDS as its techniques can be dissimilated by their different preference criterions and algorithm, model functions and representations [15]. Data mining provides several techniques that could be used for Intrusion Detection which includes Data summarization, Visualization, Clustering, Association, Classification, Prediction and Sequence analysis. A well-known machine learning technique is the decision tree, composing of three basic elements: a decision node specifying a test attributes, an edge or a branch corresponding to the one of the possible attribute values this means one of the test attribute outcomes, a leaf which is also named an answer node contains the class to which the object belongs [10]. Decision tree consists of decision nodes, each node selects the "fitting" test properties, and defines the class label of each leaf [10]. Decision trees as well known have been found to have an accuracy in classification as it uses info gain as an entropy for classification of instances but researchers desire increase in accuracy as the classifying factor (information Gain) used for decision tree has been discovered to be bias to instances with large attribute value which often leads to under-fitting or over-fitting in decision trees. Gain ratio is an advancement of the information gain feature selection technique, which solves the issue of biasness towards features with a larger set of values exhibited by information gain [5]. This paper is further arranged thus; a brief review of related works, the proposed methodology, Evaluation set-up, performance comparisons results, the results and performance analysis of the reduced and full dataset and lastly conclusion.

2. RELATED WORKS
[6] Carried out a comparison among different feature selection methods on KDDCUP'99 benchmark dataset and evaluated their performance in terms of root mean square error, detection rate and computational time. The feature selection methods were combined with search methods as follows Ranker + ChiSquaredAttributeEval, GeneticSearch + CfsSubsetEval, Ranker + GainRatioAttributeEval, GreedyStepwise + CfsSubsetEval, , Ranker + InfoGainAttributeEval and BestFirst + CfsSubsetEval where classification was carried out using Naïve Bayes and C4.5, the author according to the observation opined that Ranker+InfoGainAttributeEval took less computational time among all the feature selection methods while the performance of Ranker+GainRatioAttributeEval is good in

terms of detection rate though it took more testing and training time. [12] applied feature reduction using three standard feature selection methods Information Gain (IG), Correlation-based Feature Selection (CFS), Gain Ratio (GR and) proposed a method. A comparison of feature reduction methods was done by decision tree classifier that shows that the proposed model is more proficient for network intrusion detection. Their experiment pointed out that their method has higher detection rate and lower false alarm rate than that of full dataset and also performed as good as other methods. Another work was carried out where [1] reviews the current state of art data mining techniques. Comparison of various data mining techniques that are used to implement an intrusion detection system such as Decision Trees, Artificial Neural Network, Naïve Bayes, Support Vector Machine and K- Nearest Neighbor Algorithm was carried out. The advantages and disadvantages of each of the techniques were highlighted. An experiment was also carried out on these algorithms **.**From the results of experiments, the decision tree algorithm gave the best detection rate and also had the best kappa statistic making it superlative for real-time classification tasks due to its relatively fast classification speed and high detection rate.

3. RESEARCH METHODOLOGY

3.1 Proposed IDS Architecture

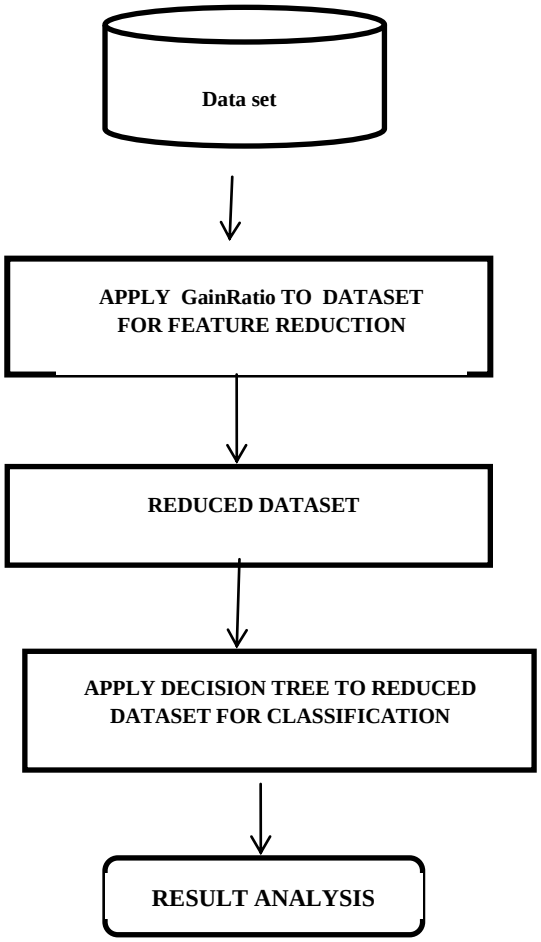

Fig 3.1.1: Proposed IDS architecture

3.2 Evaluation setup

The experiments were carried out on a 64-bitWindows7 operating system with 4GB of RAM and a Pentium (R) Dual –core CPU at 2.20GHz per core using the WEKA tool, developed by the University of Waikato via JAVA programming language, WEKA is a data mining system implementing various machine learning algorithms. For the purpose of this research the datasets that will be used are the U2R, DoS, NORMAL, PROBING AND R2L dataset in the KDD'99 dataset. To access the effectiveness of the algorithm, it was trained and tested using the KDD dataset with a 10-fold cross validation in Weka Environment. This method divides the dataset into 10 subsets, one of the 10 subsets is used as the test set while the remaining k-1 is used as training set, then the performance statistics are calculated across all the 10 subset which in turn provides a good clue of how sound the classifier works. The performance of the algorithm was evaluated on the feature reduced dataset and the original dataset.

3.3 Performance Evaluation

For the purpose of this research, the performance of the classification algorithms used will be evaluated via correctly and Incorrectly Classified Instance, Kappa Statistics, Mean Absolute Error, Root Mean Squared Error and Relative Absolute Error and also it will be measured using True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN), Recall, Accuracy, and Precision. A very high accuracy can be achieved easily by carefully selecting the sample size. Using accuracy as a measure for testing the performance of the system can be biased; However, precision and recall are not dependent on the size of the training and the test samples.

They are defined as follows:

(1) Precision $= \frac{TP}{TP + FP}$ ……(1)

(2) Recall $= \frac{TP}{TP + FN}$ …….. (2)

And also the Training Time (TT): which is the time taken to build the model is an important criteria for measuring the performance of an algorithm.

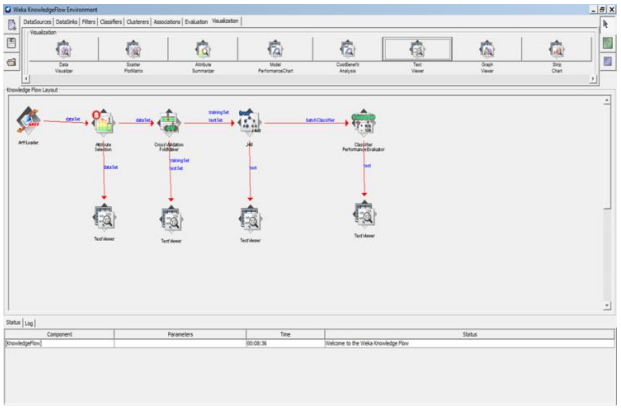

Fig 3.1.2: Knowledge flow in weka.

4. RESULT AND DISCUSSION

From the experiment carried out, it was discovered that there was an improvement in the performance of the decision tree classifier in some categories of attack i.e. Remote to Local: R2L (98.31% for reduced dataset over 98% for full data set)

and User to Root: U2R (76.92% for reduced dataset over 75% for full data set)), in the case of Denial of Service: DoS and Normal categories, both methods gave same result (100% for both full and reduced data sets) and also there was some demeaning results in the category of Probing attack (97.78% for reduced dataset over 99.49% for full data set). However, the time taken to build the model is of importance as it is a vital point in deploying software in a real time environment. The time taken to build the reduced data set is less than the time taken for the full data set which makes the reduced dataset better.

Table 4.1: Performance evaluation of Decision Tree on the reduced dataset

PARAMETERS	DOS	NORMAL	PROBING	R2L	U2R
CORRECTLY CLASSIFIED INSTANCES (%)	100	100	99.4887	98.0462	75
INCORRECTLY CLASSIFIED INSTANCES (%)	0	0	0.5113	1.9538	25
KAPPA STATISTICS	1	1	0.9926	0.8874	0.5937
MEAN ABSOLUTE ERROR	0.00	0.00	0.0005	0.0018	0.0224
ROOT MEAN SQUARED ERROR	0.00	0.000	0.0205	0.0367	0.1294
RELATIVE ABSOLUTE ERROR (%)	0.00	0.00	0.8252	10.9148	35.888
ROOT RELATIVE SQUARED ERROR (%)	0.00	0.00	11.839	41.8572	77.7255

Table 4.2: Performance measurement of Decision tree on the reduced dataset.

PARAMETERS	DOS	NORMAL	PROBING	R2L	U2R
TP RATE	1	1	0.978	0.983	0.769
FP RATE	0	0	0.007	0.06	0.201
PRECISION	1	1	0.977	0.981	0.753
RECALL	1	1	0.978	0.983	0.769
F-MEASURE	1	1	0.978	0.981	0.75
ROC AREA	1	0	0.996	0.949	0.791
Training Time	1.48secs	0.02secs	0.11secs	0.03secs	0secs

Table 4.3 Performance evaluation of Decision Tree on full dataset

PARAMETERS	DOS	NORMAL	PROBING	R2L	U2R
CORRECTLY CLASSIFIED INSTANCES (%)	100	100	97.7843	98.3126	76.9231
INCORRECTLY CLASSIFIED INSTANCES (%)	0	0	2.2157	1.6874	23.0769
KAPPA STATISTICS	1	1	0.9679	0.9014	0.5815
MEAN ABSOLUTE ERROR	0.00	0.00	0.0024	0.0021	0.0237
ROOT MEAN SQUARED ERROR	0.00	0.000	0.0362	0.0357	0.1216
RELATIVE ABSOLUTE ERROR (%)	0.00	0.00	3.9425	12.3839	38.059
ROOT RELATIVE SQUARED ERROR (%)	0.00	0.00	20.8699	40.7015	73.0542

Table 4.4: Performance measurement of Decision tree on full dataset.

PARAMETERS	DOS	NORMAL	PROBING	R2L	U2R
TP RATE	1	1	0.995	0.98	0.75
FP RATE	0	0	0.002	0.043	0.92
PRECISION	1	1	0.995	0.976	0.781
RECALL	1	1	0.995	0.98	0.75
F-MEASURE	1	1	0.995	0.978	0.761
ROC AREA	1	0	0.998	0.978	0.853
Training Time	12.82secs	0.05secs	0.25secs	0.09secs	0.02secs

Table 4.5: Table showing the representation of accuracy and time differences of both performances

CLCLASSIFIER	ATTACK TYPES				
	DOS	NORMAL	PROBING	R2L	U2R
DTDT(with reduced dataset) %	100	100	97.7843	98.3126	76.9231
Time taken to build model	1.48secs	0.02secs	0.11secs	0.03secs	0secs
DTDT (with full dataset) %	100	100	99.4887	98.0462	75.00
Time taken to build model	12.82secs	0.05secs	0.25secs	0.09secs	0.2secs

5. CONCLUSION

Conclusion can be made from the results that the influence of gain ratio technique of feature selection is expedient in the classification of attack by decision tree classifier as it drastically reduced the time taken to build the model knowing full well that decision tree classifier is a also a fast learner and it might take a great deal of time to build if fed with a high dimensional dataset. Also, the influence of feature selection shows improvement in the classification of attack especially in denial of service with respect to the time i.e 1.48secs for reduces dataset and 11.58secs for full dataset, as this category of attack is the most common type of attack found in the KDD'99 dataset. For future research, another type of feature selection technique can be used for filtering the dataset and ensemble method can also be applied.

6. REFERENCES

[1] Ajayi, A, Idowu, S.A., Anyaehie A., (2013).Comparative study of selected data mining algorithms used for intrusion detection. International Journal of Soft computing Engineering (IJSCE).

[2] Eitel J.M. Lauria and Giri K. Tayi, (2008) "A Comparative Study Of Data Mining Algorithms For Network Intrusion Detection In The Presence Of Poor Quality Data". ICIQ-03, 2008.

[3] G.V.Nadiammai, S.Krishaveni, M.Hemalatha (2011). "A comprehensive Analysis and study in intrusion detection system using data mining Techniques". IJCA, Volume 35 –No.8, December 2011.

[4] Heba Ezzat Ibrahim, Sherif M. Badr, and Mohamed A. Shaheen, (2012) "Adaptive Layered Approach using Machine Learning Techniques with Gain Ratio for Intrusion Detection Systems". IJCA, Volume 56 – No.7, October 2012

[5] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann, 2001.

[6] Megha Aggarwal, Amrita ," Performance Analysis Of Different Feature Selection Methods In Intrusion Detection ",International Journal Of Scientific & Technology research volume 2, issue 6, June 2013 ISSN 2277-8616 225 IJSTR©2013 www.ijstr.org

[7] Mitchell D'silva, Deepali Vora (2013) "Comparative Study of Data Mining Techniques to Enhance Intrusion Detection". IJERA, Vol. 3, Issue 1, January – February 2013.

[8] Mohammed J. Zaki, Wagner Meira JR.,"Data mining and analysis : Fundamental Concepts and Algorithms"

[9] Neha Maharaj and Pooja Khanna (2014), "A Comparative Analysis of Different Classification Techniques for Intrusion Detection System", IJCA, Vol. 95 –No.17, June 2014.

[10] Patel Hemant, Bharat Sarkhedi, and Hiren Vaghamshi (2013) "Intrusion Detection in Data Mining with Classification Algorithm". IJAREEIE, Vol. 2, Issue7, July 2013.

[11] Reema Patel, Amit Thakkar, Amit Ganatra (2012) " A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems".IJSCE, Volume-2, Issue-1, March 2012.

[12] Sang-Hyun C. and Hee-Su C. (2014). Feature Selection using Attribute Ratio in NSL-KDD data. International Conference Data Mining, Civil and Mechanical Engineering (ICDMCME'2014), Feb 4-5, 2014 Bali (Indonesia).

[13] Trilok Chand Sharma and Manoj Jain (2013), "WEKA Approach for Comparative Study of Classification Algorithm", IJARCCE, Vol. 2, Issue 4, April 2013.

[14] V. Jaiganesh, Dr. P. Sumathi, and A.Vinitha "Classification Algorithms in Intrusion Detection System: A Survey" IJCTA, Vol 4(5), September – October, 2013.

[15] V. Jaiganesh, S. Mangayarkarasi, and Dr. P. Sumathi (2013). "Intrusion Detection Systems: A Survey and Analysis of Classification Techniques". IJARCCE, Vol. 2, Issue 4, April 2013.

[16] Yogendra Kumar Jain and Upendra (2012), "An efficient Intrusion Detection Based on Decision Tree Classifier Using Feature Reduction". IJSRP, Vol. 2, Issue 1, January 2012.