ISSN No:-2456-2165

A Comparative Study for Machine Learning Tools Using WEKA and Rapid Miner with Classifier Algorithms Random Tree and Random Forest for Network Intrusion Detection

[1]Wathq Ahmed Ali Saeed Kawelah
1st Department of Information Systems, Faculty of Science and Technology, Omdurman Islamic University
Khartoum, Sudan

[2]Ahmed Salah Eldin Abdala
2nd Department of Computer Science, Faculty of Computer Science and Information Technology, Open University of Sudan, Khartoum, Sudan

Abstract:- The internet world expands day by day as well as threats related to it. Nowadays, Cyber-attacks often happen more than a decade ago. Intrusion detection is one of the most popular search area that provides various technologies and security techniques for detecting cyber-attacks. Different data extraction tools learn to algorithms that help in the implementation of the Learn to build Identities. In this paper, we have done a comparative study for machine learning tools using WEKA and Rapid Miner with two algorithms Random Tree and Random Forest for network intrusion detection. These can be used to implement intrusion detection techniques based on data mining. Analysis of the initial results of two different machine learning tools WEKA and Rapid Miner is carried out using KDD' 99 attack dataset and results are the best tools is WEKA, while the best algorithms is Random Forest.

Keywords:- *Data Mining Tools; WEKA; Rapid Miner; Random Tree and Random Forest.*

I. INTRODUCTION

We live in the information age, the modern communications revolution where most of the things automatically processed through computers. Information could be accessed and processed through the Internet. However, The growth of information technology has also led to an increase in the number of cyber-attacks. The recently distributed attack is facing denial of service (Dos) by DYN when 100,000 bots infected with Mirai malware [1].

This helps in obtaining information about various data mining tools that can be applied to the intrusion detection application. The main contributions of this paper are as follows:

- Comparison of the results of each classifier algorithms Random Tree and Random Forest with machine learning tools.
- Determine the best classifier algorithms Random Tree and Random Forest for two tools WEKA and Rapid Miner and determine the best machine learning tools for work.

The main focus of this paper is to apply classifier algorithms Random Tree and Random Forest in two tools WEKA and Rapid Miner to measure accuracy, sensitivity and precision by the case of network intrusion detection.

The paper is organized into 6 sections. Section 2 provides details about concepts and related work. Section 3 Provide a comparative study of the various data mining tools basic properties. In Section 4 ,The specified algorithm is described for comparison. In Section 5, Methodology are performed to analyze performance using two data mining tools to detect attack using Random Tree and Random Forest .Section 6 Concludes work.

II. CONCEPTS AND RELATED WORK

Prithvi Bisht et al.[2] have investigate " The Comparative Study on Various Data Mining Tools for Intrusion Detection". In their study, The results of some of these have also been shown. Our work includes the latest technical data mining tools and also includes descriptions of some of the depths that tend to be in depth. In the future, we would like to expand our work to provide an analysis of the profound outcome on all possible data mining tools. The application of these applications to the cloud-based security application is in progress.

Patel et al.[3] Suggested use of data mining tools to perform intrusion detection in WLAN and detect anomalies. The author provided some theoretical details about only four tools, such as WEKA, SPSS, Tanagra, and BIDS from MS SQL Server 2008. A brief description of each tool was provided and the experimental work was also limited to only the four.

Anil Sharma et al.[4] have "A Research Review on Comparative Analysis of Data Mining Tools, Techniques and Parameters", In their paper, Most of the features highlighted the use of the WEKA tool and gave knowledge of other data mining tools. The working paper A Brief Introduction to data mining techniques and parameters. Other instruments referred to briefly in the relevant work. The experimental work is also limited to a WEKA tool that only uses different workbooks such as Naive Bayes and

Classification Tree. No other tool has been used in experimental analysis of data mining tools.

III. DATA MINING TOOLS

Data mining tools supports different machine learning algorithms that are very useful in intrusion detection applications. There are different data mining tools suitable for different skilled users and for different types of data formats. Comparative knowledge of these data mining companies can help users choose a particular tool. Data mining includes various processes such as extracting, converting, uploading data, data management etc. Data mining tools with different advantages and disadvantages as follows:

A. WEKA

Data mining system developed by the University of Waikato in New Zealand in 1992[5]. WEKA is collection of different machine learning algorithms which can be used with data mining [6]. The algorithms can be applied directly to a dataset or from your Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well suited to develop new machine learning schemes[7]. WEKA is open source software issued under the GNU General Public License [5]. It is also an independent platform because the program is written in the Java™ language and contains a graphical user interface to interact with data files and produce visual results (tables and curves thinking). It also contains a generic API, so you can include WEKA, like any other library, in our applications for things like server-side data mining tasks automatically [8].

B. RapidMiner

Rapid Miner is also called another learning environment, developed in 2001, written in java by Klinkenberg et al.[9]. It is used for business purposes and commercial applications as well as for research, education and training. Quick forms. The application development supports all the steps of the data mining process including data preparation, visualization results, model validation and optimization. It is available as free and commercial versions. It is one of the most analytical tools used predictive Gartner rapid recognition of the knife in the leadership of the advanced magic quadrant analytical platforms in 2016[9].

IV. DATA MINING ALGORITHMS

A. Random Tree

Random Tree is a supervised classifier; It is a collective learning algorithm that generates many individual learners. It employs the idea of packing to build a random set of data to build a decision tree. In a standard tree each node is divided by using the best split between all variants. In a random forest, each node is divided by using the best between a subset of the randomly chosen predicates in that node. Random trees were made by Leo Breiman and Adele Cutler. The algorithm can handle classification and regression problems. Random trees are a set (set) of tree indices called forests. The classification mechanisms are as follows: Random trees creators get the input feature, distance vector protocols with a tree in the forest, Random trees are essentially a combination of two algorithms found in automated learning: One model trees are merged with random forest ideas. Model trees are decision trees where each single sheet holds a linear model that has been improved for the local sub-space and explained by this paper. Random forests have shown to improve the performance of trees one decision to a large extent: tree diversity is created by two methods of randomization [10,11].

B. Random Forest

Random Forest is an idea of the general technique of the random decision Forest which is a band learning technique for grading, regression and other tasks, that control by building many decision trees at the time of training and taking out a ten e category that is the mode of classes (classification) or means prediction (regression) of Individual trees. The forest resolution is a random minute to usually the decision trees ' over-installation of their training set. The first algorithm of a random decision Forest was created by the Tin Cam is using a random sub-space method which, in the formulation is a means of implementing the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. The extension algorithm was developed by Leo Breiman [12] and Adele Cutler,[13] and "Random Forests" is their trademark[14]. The extension of the Breiman's "bagging An idea and a random selection of advantages, first introduced by Ho[15] and later independently by Amit and Gemen [16] in order to build a collection of decision trees with controlled contrast.

V. METHODOLOGY

We have performed a performance analysis of two tools that are different from each other to measure accuracy, sensitivity and precision for Random Tree algorithm and Random Forest algorithm so that we can analyze the usage in different aspects, All experiments were performed in a computer with the configurations Intel(R) Core(TM) i5 CPU 2.50 GHz, 12 GB RAM, and the operation system platform is Microsoft Windows 7 Ultimate, We use WEKA and Rapid Miner tools (The version is WEKA 3.6.11 and Rapid Miner 6.5.2) using the following steps:

A. Data Set

For performance analysis, we have considered KDD'99 data set [16] and used two classifier algorithms Random Tree and Random Forest provided by the tools. Our motive is to analyze the performance of these classifiers using the above-mentioned tools. The KDD'99 data set is a large data set for network intrusion detection, We used 10% of the KDD'99 data set (494,021 records) for training while (311,029 records) for test and select seven features from the KDD'99 data set (see Table I), The seventh feature contains data records of two types, normal and anomaly (attacks: "Probe, Dos, U2R and R2L")[16].

NO	Feature Name	Description
1	Duration	Length of the connection in seconds
2	Flag	Status flag of the connection (normal or error)
3	Src_Bytes (Source Bytes)	Number of data bytes from source to destination
4	Dst_Bytes (Destination Bytes)	Number of data bytes from destination to source
5	Dst_Host_Same_Src_Port_Rate	Percentage of connections to the same service for destination host
6	Dst_Host_Srv_Diff_Host_Rate	Percentage of connections to the same service coming from different hosts
7	Label	Type of label (normal or anomaly “attacks: Probe, Dos, U2R and R2L”)

Table 1:- Describes the KDD’99 Data Set Seven Features

B. *Preprocessing The KDD’99 Data Set*

We prepared the KDD’99 data set in the suitable format before starting the experiments this is an analytic experimental method by the following steps:

- *Collecting Data (The KDD’99 Data set).*

- *Data Cleansing (Training 494,021, Testing 311,029):*

- Missing data handling.
- Removing or estimating missing values in the data.
- Database balancing.
- Correcting imbalances in the target field.
- Removing repeated records.

- *Data Preprocessing (Training 494,021, Testing 311,029):*
- Data Entry.
- Converting data from type to other (single valued attributes).

- *Data Analyzing Classifier (Training 49,388, Testing 27,688):*

Selected algorithms (Random Tree and Random Forest).

- *Interpretation and Analysis :*

Measure the performance of each one(accuracy, sensitivity and precision).

The total number of records in the training data set labeled 10% KDD'99 is 494,021, After filtering duplicate records, there were a total of 49,388 records. While the total number of records in the test data set labeled 10% KDD'99 is 311,029, After filtering duplicate records, there was a total of 27,688 records.

C. *Performance Measurement Terms*

We compared the performance of two tools (WEKA and Rapid Miner) using two classifier algorithms (Random Tree and Random Forest). The performance standards we consider are accuracy, sensitivity and precision.

- *Accuracy:*

Used to measure the performance of a workbook statistically. It tells how well classifier correctly identifies an instance of the dataset, Or as a percentage of the total number of predictions that are true. It can be calculated using as Equation 1:

$$\text{Accuracy} = TN + TP/TN + TP + FN + FP \qquad (1)$$

- *Sensitivity:*

It is measures the ratio of true positives with all the positives and also referred as true positive rate or recall. It can be calculated using as Equation 2:

$$\text{Sensitivity} = TP/TP + FN \qquad (2)$$

- *Precision:*

It is also referred as positive predictive value and it is the fraction of relevant instances among the retrieved instances i.e. it gives the detail of correctly identified instances. It can be calculated using as Equation 3 :

$$\text{Precision} = TP/TP + FP \qquad (3)$$

D. *Result Analysis*

We performed experimental analysis on WEKA and Rapid Miner tools using two classifier algorithms (Random Tree and Random Forest).

Summary of overall performance results for all two tools using Random Tree and Random Forest in the Table I. WEKA provides the best result in two tools. A graph is also included showing the comparison between the two different properties tools as shown in Figure1 and Figure2.

Algorithm	**Accuracy**		**Sensitivity**		**Precision**	
	WEKA	***Rapid Miner***	***WEKA***	***Rapid Miner***	***WEKA***	***Rapid Miner***
Random Tree	96.90%	96.22%	76.15%	40.92%	97.35%	72.73%
Random Forest	96.87%	97.38%	75.75%	49.45%	97.60%	98.19%

Table 2:- Comparative Results of Tools.

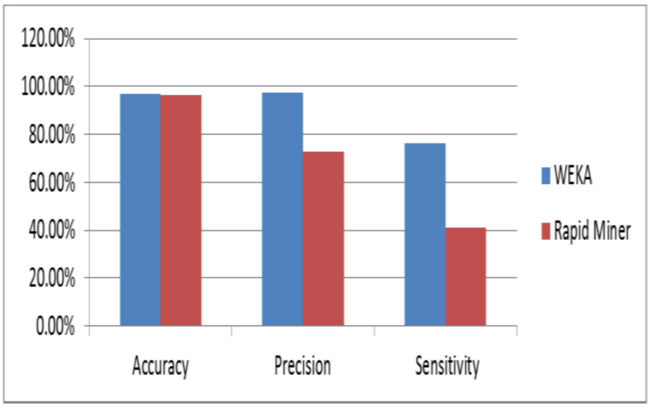

Fig 1:- Comparative Results of Tools using Random Tree.

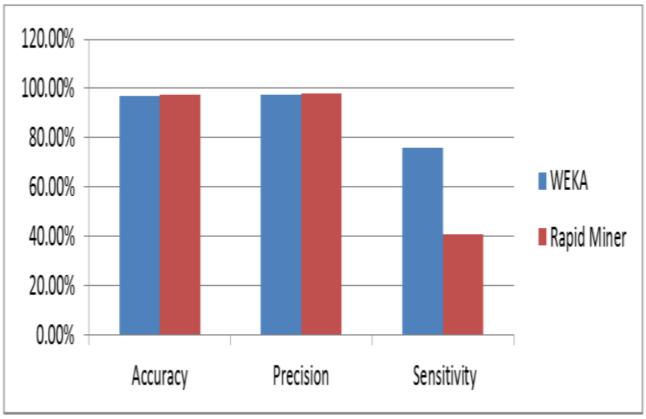

Fig 2:- Comparative Results of Tools using Random Forest.

In the result analysis over Random Forest, We can see that most of the WEKA sensitivity of 76.15% which means that WEKA performs better with Random Tree classifier. Rapid Miner also provides the best accuracy up to 97.38%. We can also observe that Random Forest have the most precision of 98.19% which means that WEKA categorized positive predictive value.

VI. CONCLUSION

Data mining tools play an important role in analyzing inbound and outbound traffic behavior from the network, Researchers can be used in the application of the Security Council to provide the necessary support machine leaning algorithms. In this paper, a comparative study is performed using WEKA and Rapid Miner tools. We have also demonstrated the results of these tools. In the future, we would like to expand our work to provide an analysis of the deep results of possible data mining tools. Also, The application of these applications for cloud-based security is in progress.

REFERENCES

[1]. S. Hilton. Dyn ddos attack analysis summary. [Online]. Available: https://dyn.com/blog/dyn-analysis -summary-of -friday -october21-attack/.

[2]. P. Bisht, N. Eeraj, M. Preeti and C. Pushpanjali “A Comparative Study on Various Data Mining Tools for Intrusion Detection,” International Journal of Scientific & Engineering Research Volume 9, Issue 5,2018.

[3]. A. M. Patel, D. A. Patel, and M. H. R. Patel, "A comparative analysis of data mining tools for performance mapping of wlan data," International Journal of Computer Engineering & Technology (IJCET), vol. 4, no. 2, pp. 241 –251, 2013.

[4]. A. Sharma and B. Kaur, “A research review on comparative analysis of data mining tools, techniques and parameters,” International Journal of Advanced Research in Computer Science , vol. 8, no. 7, 2017.

[5]. Aksenova SS. WEKA Explorer Tutorial. 2004.

[6]. H. Solanki, “Comparative study of data mining tools and analysis with unified data mining theory,”International Journal of Computer Applications, vol. 75, no. 16, 2013.

[7]. Laboratory Module 1 Description of WEKA (Java-implemented machine learning tool).

[8]. Bouckaert RR, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, et al. WEKA Manual for Version 3-7-8. Hamilton, New Zealand. 2013.

[9]. RapidMiner. Gartner magic quadrant for data science platforms. [Online].Available:https://rapidminer.com/resource/gartnermagicquadrant- data -science-platforms//.

[10]. Wikipedia contributors, C4.5_algorithm,‖ Wikipedia, The Free Encyclopedia. Wikimedia Foundation, 28-Jan-2015.

[11]. Wikipedia contributors, Random tree,‖ Wikipedia, The Free Encyclopedia. Wikimedia Foundation, 13-Jul- 2014.

[12]. Breiman Leo (2001). "Random Forests". Machine Learning 45 (1): 5–32.

[13]. Liaw, Andy (16 October 2012). “Documentation for R package random forest”. Retrieved 15 March 2013.

[14]. Ho, Tin Kam (1995). Random Decision Tree (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[15]. U.S. trademark registration number 3185828, registered 2006/12/19.

[16]. KDD99, KDD Cup 1999 Data , 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.