ISSN 2286-4822
www.euacademic.org

EUROPEAN ACADEMIC RESEARCH
Vol. VII, Issue 2/ May 2019

Impact Factor: 3.4546 (UIF)
DRJI Value: 5.9 (B+)

A Comparative Study on Machine Learning Tools Using WEKA and Rapid Miner with Classifier Algorithms C4.5 and Decision Stump for Network Intrusion Detection

WATHQ AHMED ALI SAEED KAWELAH
PhD Research Scholar, Department of Information Systems
Faculty of Science and Technology
Omdurman Islamic University, Sudan
AHMED SALAH ELDIN ABDALA
Professor, Department of Computer Science
Faculty of Computer Science and Information Technology
Open University of Sudan, Sudan

Abstract

Intrusion detection system dealing with the huge amount of data that include repeated irrelevant cause slow process of testing, training and higher learning resource consumption as well as the vulnerability of the detection rate. Data mining techniques are being applied in the construction of intrusion detection systems to protect computing resources against unauthorized access. In this paper, we have done a comparative study on machine learning tools using WEKA and Rapid Miner with two classifier algorithms C4.5 and Decision Stump for Network Intrusion Detection to measure the accuracy, sensitivity and precision. The results of the experiments using the KDD' 99 attack dataset and select seven features, The results show the best tools Rapid Miner for the accuracy and precision, while the best algorithms is C4.5.

Keywords: Data Mining Tools; WEKA; Rapid Miner; C4.5 and Decision Stump

I. INTRODUCTION

We live in the information age, the modern communication revolution where most things are handled automatically by computers. Information can be accessed and processed online. However, the growth of information technology has also led to an increase in the number of cyber-attacks. The recently distributed attack faces a denial of service (Dos) by DYN when 100,000 robots are hit with Mirai malware [1].

II. OBJECTIVES OF THE STUDY

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

A. A comparison of the results of all the algorithms C4.5 and Decision Stump with machine learning tools.

B. Determine the best C4.5 and Decision Stump algorithms for the WEKA and Rapid Miner tools and the determine the best machine learning tools for work.

The main focus of this paper is to apply of classifier algorithms C4.5 and Decision Stump in two tools WEKA and Rapid Miner to measure accuracy, sensitivity and precision by the case of network intrusion detection.

III. METHODOLOGY

We performed a performance analysis of two different tools to measure accuracy, sensitivity and accuracy for C4.5 and Decision Stump algorithms so we can analyze the use in different aspects. All experiments were performed in a computer with the configurations Intel(R) Core(TM) i5 CPU 2.50GHz, 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 [2] and used two classifier algorithms C4.5 and Decision Stump provided by the tools. Our motivation is to analyze the performance of these classifiers using the tools listed above. 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 1), The seventh feature contains data records of two types, normal and anomaly (attacks: "Probe, Dos, U2R and R2L")[2].

Table-1 Describes the KDD'99 Data set Seven Features

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")

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 the steps[3]:

1. Collecting Data (The KDD'99 Data set).
2. Data Cleansing (Training 494,021, Testing 311,029):
    a) Missing data handling.

   b) Removing or estimating missing values in the data.
   c) Database balancing.
   d) Correcting imbalances in the target field.
   e) Removing repeated records.
3. Data Preprocessing (Training 494,021, Testing 311,029)
   a) Data Entry.
   b) Converting data from type to other (single valued attributes)
4. Data Analyzing Classifier (Training 49,388, Testing 27,688):
   Selected algorithms (C4.5 and Decision Stump).
5. 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 C4.5 and Decision Stump. 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[4]:

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

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[4]:

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

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[4]:

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

IV. DATA MINING TOOLS AND ALGORITHMS

Data mining tools supports different machine learning algorithms that are very useful in intrusion detection applications. There is a data extraction tools appropriate for various skilled users of various types of data formats. Comparative knowledge of data mining companies can help users choose a particular tool. Data mining includes various processes such as extracting, conversion, loading, data management, etc. Data mining tools and algorithms have 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 a collection of various machine learning algorithms which can be used with data mining [6]. Algorithms can be applied directly to a data set 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 an open source under the GNU Public License [5]. As an independent platform because the program written in Java™ and features a graphical user interface in the interaction with the data files visible 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. Rapid Miner

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

C. C4.5

J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple.[10]

While building a tree, J48 ignores the missing values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic idea is to divide the data into range based on the attribute values for that item that are found in the training sample. J48 allows classification via either decision trees or rules generated from them.[11]

D. Decision Stump

A Decision Stump is a machine learning model consisting of a one-level decision tree.[12] That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rules.[13]

Depending on the type of the input feature, several variations are possible. For nominal features, one may build a stump which contains a leaf for each possible feature value[14] or a stump with the two leaves, one of which corresponds to some chosen category, and the other leaf to all the other categories. For binary features these two schemes are identical. A missing value may be treated as a yet another category[14].

V. RESULT ANALYSIS

We performed experimental analysis on WEKA and Rapid Miner tools using two classifier algorithms C4.5 and Decision Stump.

Summary of overall performance results for all two tools using C4.5 and Decision Stump (see Table 2). Rapid Miner 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.

Table-2 Comparative Results of Tools

Algorithm	Accuracy		Sensitivity		Precision	
	WEKA	Rapid Miner	WEKA	Rapid Miner	WEKA	Rapid Miner
C4.5	96.47%	99.19%	72.55%	86.19%	97.20%	97.55%
Decision Stump	93.87%	96.88%	74.40%	42.34%	74.20%	91.69%

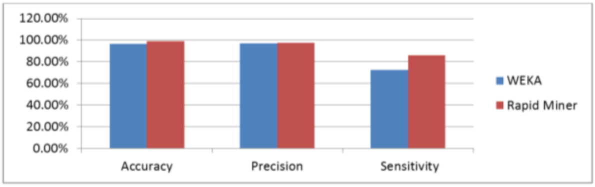

Fig.1.Comparative Results of Tools using C4.5.

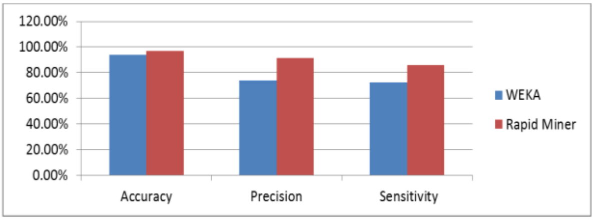

Fig.2.Comparative Results of Tools using Decision Stump.

In the result analysis over C4.5, We can see that most of the Rapid Miner accuracy of 99.19% which means that WEKA performs better with C4.5 classifier. Rapid Miner also provides the best sensitivity up to 86.19% which means it can correctly identify the positive results from samples than the WEKA tool. We can also observe that Rapid Miner have the most precision of 97.55% which means that Rapid Miner categorized positive predictive value.

VI. CONCLUSION

The main goal of this paper is to apply of classifier algorithms C4.5 and Decision Stump in two tools WEKA and Rapid Miner by the case of network intrusion detection. We summarized the experiments conducted using KDD' 99 data set and result is explored based on the accuracy, sensitivity and precision .

We compared the performance of WEKA and Rapid Miner on C4.5 and Decision Stump approaches of network intrusion detection. The results of our experimental study shows that the best tools Rapid Miner, while the best algorithms is C4.5.

REFERENCES

1. S. Hilton. Dyn ddos attack analysis summary. [Online]. Available: https://dyn.com/blog/dyn-analysis -summary-of -friday -october21-attack/.
2. KDD99, KDD Cup 1999 Data , 1999. [Online]. Available http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.
3. Han J, Kamber M. Data mining: concepts and techniques. 2006**.**
4. Confusion matrix - Wikipedia [Online]. Available: https://en.wikipedia.org/wiki/Confusion_matrix
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. Tina R. Patil, Mrs. S. S. Sherekar: Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification.

11. [Online].Available : http://stackoverflow.com/questions/10317885/decision-tree-vs-naive-bayes-classifier//.
12. Iba, Wayne; and Langley, Pat (1992); Induction of One-Level Decision Trees, in ML92: Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, Scotland, 1–3 July 1992, San Francisco, CA: Morgan Kaufmann, pp. 233–240
13. Holte, Robert C. (1993). "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets". CiteSeerX 10.1.1.67.2711. Missing or empty |url= (help **(**
14. Loper, Edward L.; Bird,, Steven; Klein, Ewan (2009). Natural language processing with Python. Sebastopol, CA: O'Reilly. ISBN 0-596-51649-5.