A Comparative Study on Various Data Mining Tools for Intrusion Detection

[1]Prithvi Bisht, [2]Neeraj Negi, [3]Preeti Mishra, [4]Pushpanjali Chauhan
Department of Computer Science and Engineering
Graphic Era University, Dehradun
Email: {[1]prithvisbisht, [2]neeraj.negi174, [3]dr.preetimishranit, [4]pushpanajlichauhan}@gmail.com

Abstract—Internet world is expanding day by day and so are the threats related to it. Nowadays, cyber attacks are happening more frequently than a decade before. Intrusion detection is one of the most popular research area which provides various security tools and techniques to detect cyber attacks. There are many ways to detect anomaly in any system but the most flexible and efficient way is through data mining. Data mining tools provide various machine learning algorithms which are helpful for implementing machine-learning based IDS. In this paper, we have done a comparative study of various state of the art data mining tools such as RapidMiner, WEKA, EOA, Scikit-Learn, Shogun, MATLAB, R, TensorFlow, etc for intrusion detection. The specific characteristics of individual tool are discussed along with its pros & cons. These tools can be used to implement data mining based intrusion detection techniques. A preliminary result analysis of three different data-mining tools is carried out using KDD' 99 attack dataset and results seem to be promising.

Keywords: Data mining, Data mining tools, WEKA, RapidMiner, Orange, KNIME, MOA, ELKI, Shogun, R, Scikit-Learn, Matlab

1. Introduction

We are living in the modern era of Information technology where most of the things have been automated and processed through computers. Information can be shared, accessed & processed across the Internet. However, growth in the information technology has also led to increase in the number of cyber attacks. Recently Distributed Denial of Service (DDos) attack is faced by DYN when 100,000 bots infected with Mirai malware [1]. On Apr, 2016, the global hacking group Anonymous launched a DDos campaign against Donald Trump. Under the name OpTrump, the group sought to take down the billionaire s websites for his hotel chain and presidential campaign, as well as his email servers [2].

There exist some of the defensive system in order to deal with these attacks. An intrusion detection system (IDS) is of the popular defensive system. IDS is a software that checks the system for malicious activities. IDS are of two types: Misuse detection and Anomaly detection. Misuse detection based IDS generally detects attacks by regularly looking at the specific patterns in network traffic. Anomaly detection based IDS classifies regular system activity as normal or anomalous by observing deviation from the normal behavior of the system. Misuse detection can be performed using traditional signature-based IDS or supervised machine learning learning (ML) based IDS. Supervised machine learning algorithms assume to have the knowledge of all attack and normal network connections. Anomaly Detection can be performed using semi-supervised/unsupervised machine learning-based IDS, Finite State Machine based IDS, Statistical learning based IDS, described in detail in our recent survey [3]. These systems assume to have the knowledge of normal behavior of the system or no knowledge about behavior type.

The process of converting raw data into useful information or routine patterns is called data mining [4]. These patterns can help in differentiating between regular activity and malicious activity. Machine-learning based IDS makes use of data mining algorithms to process the data. There exist many data mining tools that provide the implementation of various data mining algorithms. Data mining tools provide various functions to load the data, filter the data, approaches to select features, classification/clustering algorithms and data visualization, etc. The input data is either divided into subgroups by clustering or divided into user created categories through classification. This helps in creating a prediction model or a mathematical function which helps in regression.

The main focus of this paper is to provide a comparative study of various state of the art different types of data mining tools. Some of these tools are WEKA [5], Orange [6], ELKI [7], Rapid Miner [8], Tensor flow [9], R [10], Scikit-learn [11] etc. Some of these tools provide various machine learning algorithms and some support deep learning algorithms. Some of them are open source and some are paid. These tools are coded in different languages and have distinct characteristics. The special key features of each of the tools, along with its advantages and disadvantages are discussed in this paper. This helps in gaining information about various data mining tools which can be applied for intrusion detection application. The main contributions of this paper are as follows:

- To provide a comparative study on various data mining tools based on different parameters.
- To provide performance analysis of the popular three data mining tools to illustrate their capability for detecting attacks.

The paper is organized into 5 sections. Section 2 provides the details about the related work. Section 3 provides the

comparative study of various data mining tools and their key characteristics. We have also discussed the pros and cons of these tools. In Section 4, a performance analysis is carried out using three data mining tools for attack detection. Section 5 concludes the work with future directions.

2. Related Work

Sharma et al. [12] mostly highlighted the features and use of WEKA tool and has given a brief knowledge about other data mining tools such as Orange, Knime, R and Keel. Paper had brief introduction to data mining techniques and parameters. Others tools were briefly described in related work section. Experimental work is also limited to only WEKA tool using different classifiers like Naive Bayes and Classification Tree. No other tool was used in the experimental analysis of data mining tools.

Patel et al. [13] proposed the use of data mining tools to perform Intrusion Detection in WLAN and Anomaly Detection. Author gave some theoretical details about only four tools i.e. WEKA, SPSS, Tanagra and BIDS of MS SQL Server 2008. A short description of each tool was given and an experimental work was also limited to these four tools only. Author performed clustering with these tools and interpreted the results. Some of the major tools like ScikitLearn, Matlab and R were missing in the research of this paper.

Mikut et al. [14] gave a detailed description of various data mining tools and also gave their historical development to their likewise categorization. It presents a complete theoretical guide in data mining tools, but there was no experimental work and tools were just described not compared in any way. It lacked experimental analysis of the tools.

King and Elder [15] have conducted an assessment of various data mining tools. Results have provided a technical report that details the evaluation procedure and the scoring of all component criteria. Authors also showed that the choice of a tool depends on a weighted score of several categories such as software budget and user experience. Finally, authors have showed that the tools' price is related to quality. However, the paper lacks in implementing multiclass classification problem of network intrusion application using data mining tools.

Abbott [16] have deeply compared five of the best data mining tool available at the time. They have selected a two phase selection followed by an in depth evaluation of algorithm used as well as the tool. Author has added screenshots for the validation of the results. Tools were categorized based on different characteristics. For the final evaluation i.e. hands on evaluation author selected 5 best tools with the help of expert evaluators. However, the tools used by authors are rarely used now and the technology has exponentially grown since then, more attacks have been introduced, new algorithms are created.

Wahbeh et al. [4] provides a good classification of tools like Knime, Tanagra, Weka and orange, With a rich comparative study of these tools they also propose the comparative study of these tools. They have selected different data sets from the UCI repository to work on. Some of these are Audiology (Standardized), Breast Cancer Wisconsin (Original), Car Evaluation. Author has also used different algorithm for classification on each tool and given analysis table for above. No performance analysis was done as only algorithms were classified in tools; no comparison was done between the
tools

Rawat [17] gave a perfect introduction to data mining tools and techniques. It highlights the main features of data mining tools and basically categorizes them into different types. The tools taken for study are most commonly used tools. Any kind of differentiation between tools is unavailable & no comparative analysis and no performance analysis is provided. Also some of the mostly used tools such as Scikit-learn and Matlab were not included in the analysis. The research analysis in the paper was short and was mostly theoretical.

Gera et al. [18] also gave a good theoretical description of tools. Comparison between different tools was shown in tabular form. Theoretical comparison is also done but there was no experimental comparison or performance analysis was done between any tools in any way.

Solanki et al. [19] described only three data mining tools and provided various features and functionality of the tools. Comparative analysis of these tools was also given in the paper along with some experimental work on these three tools. Author also described unified data mining theory in the paper. There are many other tools available for data mining. Some them supports deep learning, included in our paper.

In this survey, we have included state of the art eleven most popular different data mining tools. We have provided the key characteristics of all these tools along with their pros and cons. Most popular tools for deep learning application such TensorFlow, MOA, KMINE, RapidMiner etc. have also been included in our paper. A performance analysis is carried for network attack detection using three distinct tools.

3. Data Mining Tools

Data mining tools supports different machine learning algorithms that are very helpful in intrusion detection applications. There are different data mining tools suitable for different skilled users and for different types of data formats. A comparative knowledge of these data mining tools can help the users in selecting particular tool. Data Mining includes different processes like extraction, transformation and loading of data, managing data etc. Various data mining tools along with their pros and cons are described below:

3.1. WEKA

Waikato Environment for Knowledge Analysis was developed in 1992. Weka is collection of different machine learning algorithms which can be used for data mining [20]. It is written in Java and is especially used for educational

research purposes. Weka is a platform independent, open source, easy to use, data processing tool, flexible for scripting experiments and 3 graphical user interface tool. Weka contains different tools and algorithms for regression, classification, pre-processing and clustering. It is supportable on different platforms such as Mac OS, Linux and Windows. When dealing with large data sets, it is best to use a CL based approach as Explorer tries to load the whole data set into the main memory causing performance issues.

3.2. RapidMiner

RapidMiner also called Yet Another Learning Environment), was developed in 2001, written in java by Klinkenberg et al. [8]. RapidMiner provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It is available as both free and commercial editions. It is one of the most used predictive analytic tools. Gartner recognized Rapid Miner and Knife as leaders in the magic quadrant for advanced analytic platforms in 2016. It is available for Mac OS, Linux and Windows. Its user friendly and rich library of data science algorithms and machine learning algorithms makes it first choice for enterprises to implement predictive analysis in their business processes. Its unique features are repeatable work flows, built in templates, visualization and integration with different languages like Weka, SPSS, Python and R which helps in rapid prototyping. RapidMiner is mainly used in educational and research fields for data exploration and visualization, data mining, financial forecasting, segmentation image mining can be integrated with Weka.

3.3. Orange

It has python libraries that benefits python scripts with its affluent collection of machine learning algorithms, data mining and clustering techniques. Orange has many features like Add-ons extension functionality, hands-on training, visual representation and easy to use GUI [6]. It executes simple and complex data analysis, have access to external functions for advance analysis, also help to create graphics. Orange can perform data mining through different visual programming or by python scripting, which makes it easier for naive users. It can be installed on any OS and can be used as a python library. It also uses libraries like numpy, scipy etc. Due to its extended features this tool is best for the naive users.

3.4. KNIME

It is also called as Konstanz Information Miner and is an open source data mining tool. It is written in Java and built upon Eclipse. It is used for data analytics, reporting and integration platform. It has modular data pipelining concept [21]. It can also be used in different areas like business intelligence, financial data analysis and customer data analysis. For different works it is sometime considered as SAS alternative. It supports different languages such as perl, python, java etc. It has different features like seamless integration with different software like R and Weka, platform flexibility and openness.

3.5. MOA

Massive Online Analysis (MOA) is also known for Big Data Stream Analytics Framework is well suited for handling large volume of real-time data streams at a very high speed. MOA is distributed under GNU GPL, and can be used via the command line, GUI or Java API. It's a framework providing storage setting for data streams, run the experiments in data stream mining context [22]. Here, we can use WEKA classifiers from Massive online analysis (MOA) and vice versa also possible. Stream clustering and classification both are possible in MOA. It has change and outlier detection quality, and also support concept drift. In many cases, this tool is very similar to WEKA and includes different online as well as offline tools for evaluate Hoeffding Trees with and without Nave Bayes classifiers. It is possible to use WEKA classifiers from MOA, and MOA classifiers and streams from WEKA.

3.6. ELKI

A open source software for developing KDDApplication [7] and offers data indexing structures and provides major performance gain. It is written in Java. ELKI offers data index structures such as the R*-tree that can provide major performance gains. ELKI [23] provides a large assortment of algorithms. It is written in java. Algorithms can be accelerated using appropriate index structure. Every algorithm, index, visualization and even every distance function and data type in ELKI is an extensible. The main focus of ELKI is in algorithm research, while giving more importance to unsupervised methods in cluster analysis.

3.7. Shogun

It is a free and open source machine learning software library written in C++. It has a vast ranges of machine learning techniques [24]. It offers numerous data structures and algos for ML(machine learning) problem. This tool mostly attract by SVMs (kernel machine) for classification and regression. It is capable of processing huge datasets. Platforms Supported: Shogun supports GNU/Linux, MacOS, FreeBSD, and Windows. It has an easy combination of multiple data representations, and efficient tools and algorithms that supports rapid prototyping of data pipelines. It is free of cost which is community-based and also supports machine.

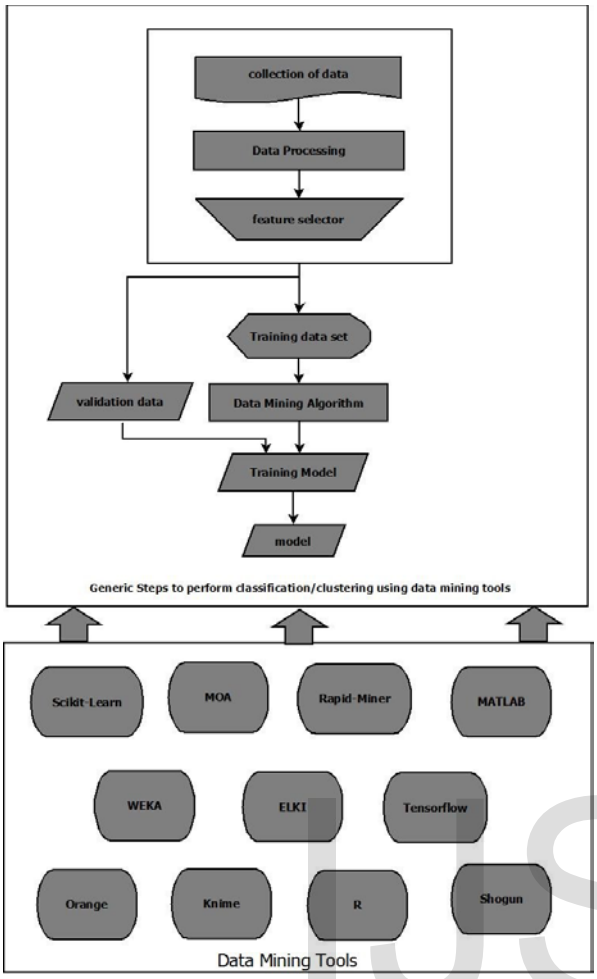

Figure 1. Generic Execution Flow of Data Mining Tools

learning educations. It Supports Dimensionality reduction algorithms, Support Vector Machines, Online learning algorithms, Clustering, Regression, and KNN. It offers interfaces for Octave, Python, R, Java, Lua, and Ruby and C# using SWIG.

3.8. Scikit-learn

Scikit-learn [11] is a free of cost library which supports Python.It set in such a way so that it can easily work on Python numerical and support scientific libraries like numpy, scipy etc. Scikit-learn support both supervised and unsupervised learning. Clustering, Cross validation, Datasets, Dimensionality Reduction, Ensemble Methods, Feature extraction, Feature selection, Parameter Tuning, Manifold Learning and Supervised models [26]. Scikit learn can be used in any OS that supports python. Input Datasets Format consist of svmlight/libsvm, pandas.io provides tools to read CSV, Excel, JSON and SQL, scipy.io provides tools to read in binary formats such as .mat and .arff, images, videos and audios. Scikit is very easy to use and quite easy to install and can even work on large datasets. Commercially Used by: Spotify, Inria, betaworks, Evernote and many more. It is distributed under BSD license. Though it have some functions to load data in a streamline manner so that larger datasets can be loaded. Pre-requisite: NumPy, SciPy, Matplotlib, Pandas, SymPy.

3.9. R

R [10] is an open source programming language and environment for numerical computation. R language is used worldwide and a common language for data mining and developing numerical software. The R software is mainly written in C, FORTRAN and R. It is supported in Mac OS, Windows and Linux. Rs objects can be manipulated with C, C++, java, .net, Python. R studio is the mainly used GUI for R. To visualize or mine data in R, some packages may need to be installed like RODBC, Gmodels, class, tm, e107i, neural net, kernlab, rpart etc. Installing all these packages may take extra space in host system. R has the most packages available in all data mining tools available at CRAN and github, which make it diverse than others. Almost all machine learning algorithms are available in R.
It is mainly used in finance, genetics, machine learning 3.10.

3.10. Matlab (MATrix LABoratory)

MATLAB [25] is a high-performance multi-paradigm environment for numerical computing. It is developed by Math works organization and first released in 1984. It mixes computation, visualization and programming in a very easy way. Its' programming environment is very easy to use where problems and solutions are represented in easy mathematical form. When doing data mining, a large part of the work is to manipulate data. The part of coding the algorithm can be quite short since Matlab has a lot of powerful toolboxes for data mining. And when manipulating data, Matlab is definitely better. It is normal since it is done to work with matrices (MATrix LABoratory). Matlab is available for all OS. It is primarily written in c, c++ and java and is mainly used for numerical computing. Typical uses include: Data analysis, exploration, and visualization.

3.11. Tensor Flow

Tensor Flow [9] is an open source Deep Learning library developed by Google Brain Team which was released in 9 November 2015. It is written in Python, C++ and CUDA and supports all major platforms such as Linux, Mac OS, Windows and also android. Tensor Flow can be used to make machine learning applications for android as well which is its unique feature. It was developed to replace its predecessor, Dist Belief. It was released under the Apache
2.0 open source license. Googles Rank Brain is also backed

TABLE 1. MY CAPTION

Tool Name	Matlab [25]	R [10]	Scikit-Learn [11]	WEKA [20]	Orange [6]	Knime [21]	RapidMiner [8]	ELKI [7]	Tensorflow [9]
Initial Release	1984	1993	June, 2007	1993	1997	2004	2006	2008	2015
Availability	Not open source Trial period	Open Source	Open Source	Open Source	Open Source	Open Source	Not open source Trial period	NA	Open Source
Current Version	R2017b	3.4.1	0.19.0	3.8.1	3.5.0	3.4.1	7.5	0.7.1	1.3.0
License	Proprietary commercial software	GNU GPL v2	BSD License	GNU General Public License	GNU General Public License	GNU General Public License	Proprietary	AGPL	Apache 2.0 open source lic.
Platform	Windows, Mac OS and Linux	Windows Linux	Windows, Mac OS and Linux	Windows, OS X, Linux	Windows, Mac OS , Linux	Windows, Mac OS , Linux	Windows, Mac OS , Linux	Windows, Mac OS , Linux	Windows, Mac os, Linux
Written in	C, C++ and java	C and Fortran	Python, Cython , C and C++	Java	Python, Cython, C++, C	Java	NA	Java	Python,c++, CUDA
Language Supported	Matlab	R	Python	Java	Python	NA	NA	Java	Python
User Groups	EXPERT	AVERAGE	AVERAGE	NAIVE	AVERAGE	EXPERT	EXPERT	Average	EXPERT
System requirement	High	Average	Average	Average	High	NA	high	NA	Average
Graphical Representation	yes	yes	yes	yes	yes	NA	NA	Na	yes
Ease of Learning	Hard	Easy	Average	Easy	Hard		Hard		Average
Interface	GUI	Both CLI and GUI	CLI	GUI	Both GUI and CLI	GUI	GUI	GUI	CLI
Type	Computational Analysis	Statistical computing	ML-based	ML-based	ML, Data mining, data visualization	Enterprise reporting, Business Intelligence	Statistical analysis, predictive analysis	cluster analysis and outlier detection	Machine Learning Library, Deep Learning
Deep Learning Support	yes	Yes	No	No	No	Yes	Yes	No	Yes
Y									

by Tensor Flow. Tensor Flow is mostly used for deep learning problems such as automated image captioning, neural networks, pattern recognitions etc.

Discussion: By doing an exhaustive survey on data mining tools, it can be inferred that some of these tools are easy to install and user friendly; while some may be more sophisticated to use. For a naive user, WEKA can be the easiest tool to start with due to its user friendly GUI. It supports many machine learning algorithms hence it is well suited for educational purposes and researches that requires data mining. For users with intermediate knowledge of Python language, Scikit-learn is the best tool for data mining purpose. Scikit-learn contains large amount of machine learning libraries and can be combined with python numerical and scientific calculations libraries like numpy, scipy, etc.

It is easy to install and is one of the most useful data mining tool for attack detection application. For users with good programming background, R and Matlab can also be one of the options. Both R and Matlab have numerous packages to use in data mining. Both are comparatively moderate to install and use. With less lines of code, users can perform many data mining tasks. Matlab requires more memory space and can be processor heavy. R requires more memory space for package installation but it is easy on processors.

For a naive users, Orange can also be an alternate but on using Orange few complications were found. Its GUI is easy to use. While doing hands-on analysis, it is found that it requires some more parameters in dataset unlike other data

mining tools. It can also be used as CLI just like Scikitlearn, but Scikit-learn has greater advantages over Orange like more data mining functionalities than Orange. MOA is mostly used for large datasets or data streams hence for ordinary users it is not much of use but for industries it is a great option to consider. For programmers, Shogun is also a good option as it supports multiple languages and mainly focuses on kernel machine like support vector machines etc. Most of the above tools provide graphical representation of data and results. Scikit-learn use matplotlib library to provide options for many graphical representation of data.

R and Matlab have packages that can provide the graphical representation as well. Other tools can provide graphical representations as well but Scikit-learn, R and Matlab are better than the other tools. Tensor Flow is a deep learning library made by Google hence it can be easily used for image recognition, neural networks, pattern deciphering and correlations etc. It is mostly meant for deep learning problems. Tensor Flow can be easily used thorough Anaconda tool or it can be used with the help of Docker images.

Tensor Flow is also easy to learn and use. It supports Python Language hence a user with programming background can easily learn to use it. MOA is generally used to handle large volumes of real-time data streams hence it may not be the perfect option for a general purpose user but may be of significant use to enterprises and industries that require processing large chunks of data. Rapid Miner is a simple and fast tool for anyone who knows the steps of data mining but do not know how to code for it. With Rapid Miner anyone can perform data mining without writing a single line of code hence it has an advantage over other CLI based data mining tool that everyone may not be able to use. This advantage makes tools like Orange and Rapid Miner more popular to naive users.

We have considered three different tools such as Scikitlearn, R and Weka for experimentation over attack dataset which are best suited for universal use. These all are easy to install, learn and use. With some little knowledge of the respective language, any user can easily perform data mining and analyses using these three tools. Learning resources are also easily available for these three tools. These tools have many flexible uses and are well suited for educational research, industrial use and scientific purposes.

4. Experiment

We carried out performance analysis of three tools which are different from each other in terms of user interface so that we can analyze the usage in different aspects. We use Scikit-learn (CLI), Weka (GUI) and R (both CLI and GUI).

4.1. Data Set

For the performance analysis, we have considered KDD'99 data set [27] and used Decision Tree Classifier provided by the tools. Our motive is to analyze the performance of these three tools using above mentioned classifiers. The KDD'99 data set [27] is a large data set for intrusion detection which has 4,94,021 records and 42 attributes. It contains data records of two types, normal and anomaly. Anomaly data contain 21 kinds of attacks and we categorized these 21 attacks into four categories namely DOS, Probe, R2L and U2R since we only need to analyze the performance of the tools here.

4.2. Performance Metrics:

We compared performance of all the three tools using two Decision Tree classifier. The performance parameters that we considered are Accuracy, Sensitivity, Specificity, Precision and False positive rate.

Accuracy: It is used to statistically measure the performance of the classifier. It tells how well classifier correctly identifies a instance of the dataset. It can be calculated using formula:

$Accuracy = TN + TP/TN + TP + FN + FP$

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 formula:

$Sensitivity = TP/TP + FN$

Specificity: It measures how accurately the model will identify true negatives. It can be calculated using formula:

$Specificity = TN/TN + FP$

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 formula:

$Precision = TP/TP + FP$

It measures the ratio of negatives which are wrongly identified as positive to total number of actual negatives.

$FalsePositiveRate = FP/FP + TN$

4.3. Result Analysis

We performed experimental analysis on three different tools using two different classifier algorithms. Currently, we have considered three popular tools for result analysis: Weka [20], R Studio [10] and Scikit Learn [11] due to page limitations. Decision Tree (DT) and Nave Bayes (NB) are considered for implementation.

The overall performance results of all three tools using Naive Bayes classifier and Decision Tree Classifier is summarized in Table 2. R provides the best result in all three tools The detailed attack-wise results using R are shown in Table 3. A graph is also included showing the comparison of these 3 tools over different properties as shown in Figure2.

In the result analysis over Nave Bayes, we can see that R is providing the most accuracy of 95.80% which means that R performs better with Naive Bayes classifier. R also provides the best sensitivity upto 75.26% which means it can correctly identify the positive results from samples than the other two tools. R also has highest specificity of

TABLE 2. COMPARATIVE ANALYSIS OF RESULTS OF SCIKIT-LEARN, WEKA AND R

Algo	Tool Name	Acc.	Sensitivity	Specificity	Precision	FAR
DT	Scikit-learn	99.93	86.42	99.94	87.16	0.06
DT	WEKA	99.87	96.59	99.84	76.92	0.16
DT	R	99.98	93.23	99.98	92.07	0.02
NB	Scikit-learn	91.88	43.52	62.06	89.29	10.70
NB	WEKA	93.59	50.43	95.25	79.70	4.78
NB	R	95.80	75.26	96.22	49.95	3.78

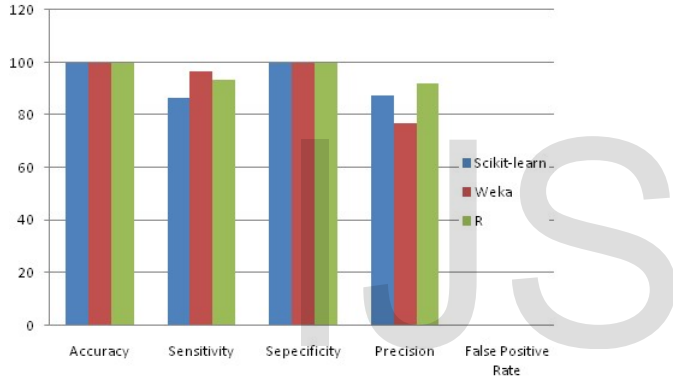

Figure 2. Comparative Results of Tools using Decision Tree

96.22% i.e. it can correctly classify the negatives. We can also observe that Scikit-learn have the most false positive value of 10.70% which means that Scikit-learn categorized negative instances as positive the most.

We perform the same process using Decision Tree algorithm. Results are improved in all three tools. We can see that here also R performs better than others with the accuracy score of 99.98% which clearly means that R is better tool than the other two over considered dataset. In case of decision tree, WEKA has the most false positive rate of 0.16% which means Weka is considering normal processes as attack more than the other two tools. WEKA has the highest sensitivity value of 96.59% which means it correctly classified the positives better than other two tools. The experimental table of R shows that DOS attack, Probe attack and User to Root attacks have highest accuracy of 99.99%. DOS attacks have highest sensitivity of 99.99% and DOS attacks also have the highest false positive rate of .03%. The overall accuracy of R remains the highest 99.98%.

The above performed experimental work and its analysis

TABLE 3. ATTACK-WISE RESULTS FOR R USING DECISION TREE ALGORITHM

Class	Accuracy	Sensitivity	Specificity	Precision	FPR
normal	99.97	99.92	99.98	99.93	0.03
DOS	99.99	99.99	99.97	99.99	0.03
Probe	99.99	99.39	99.99	99.51	0.01
R2L	99.98	96.88	99.99	97.32	0.01
U2R	99.99	70.00	99.99	63.63	0.01
Overall	99.98	93.23	99.98	92.07	0.02

is done from a beginners perspective i.e. the outcome of a particular model either Nave Bayes or Decision Tree is done on default parameters of their respective models. However, results can be further improved. Since each tool has its own function and default parameters to use an algorithm; the result of every tool is different for common algorithms.

In some tools like Scikit learn and R, results can be further improved by parameter tuning of the classifier. For example, in decision tree classifier, we can limit/increase the Min-sample-per-leaf node to a certain value (which is 1 by default) of the data points to stop the tree from prematurely classifying the outliers. We can also decrease the maximum depth of the tree to give all features a chance of becoming a decision node. In Scikit-learn, we can use RandomizedSearchCV function which returns the value of parameters for which the result of any algorithm will be optimum. Many other parameters can be changed to improve the results for both the classifier. On an average, we can say that data mining tools are performing promising for attack detection applications.

5. Conclusion

Data mining tools play an important role in analyzing the behavior of incoming and outgoing traffic from network. Researchers can use them in their security application to provide the support for machine leaning algorithms. In this paper, a comparative study is performed using different types of data mining tools. We have also demonstrated the results of some of these tools. Our work include state of the art data mining tools and also includes the description about some of the deep leaning tools. In future, we would like to extend our work to provide the deep result analysis on all possible data mining tools. Also, the application of these tools for cloud-based security application is under progress.

References

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

[2] M. Wycislik-Wilson. Anonymous hacks donald trump's voicemail and leaks the messages. [Online]. Available: https://betanews.com/2016/03/05/anonymous-hacks-trump/

[3] P. Mishra, E. S. Pilli, V. Varadharajan, and U. Tupakula, "Intrusion detection techniques in cloud environment: A survey," *Journal of Network and Computer Applications*, vol. 77, pp. 18–47, 2017.

[4] A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi, and E. M. AlShawakfa, "A comparison study between data mining tools over some classification methods."

[5] S. K. David, A. T. Saeb, and K. Al Rubeaan, "Comparative analysis of data mining tools and classification techniques using weka in medical bioinformatics," *Computer Engineering and Intelligent Systems*, vol. 4, no. 13, pp. 28–38, 2013.

[6] U. of Ljubljana. Orange - data mining fruitful fun. [Online]. Available: https://orange.biolab.si/

[7] E. Achtert, T. Bernecker, H.-P. Kriegel, E. Schubert, and A. Zimek, "Elki in time: Elki 0.2 for the performance evaluation of distance measures for time series," *Advances in Spatial and Temporal Databases*, pp. 436–440, 2009.

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

[9] TensorFlow. Tensorflow. [Online]. Available: https://www.tensorflow.org/

[10] T. R. Foundation. R: What is r? [Online]. Available: https://www.rproject.org/about.html

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[12] 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.

[13] 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.

[14] R. Mikut and M. Reischl, "Data mining tools," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 5, pp. 431–443, 2011.

[15] M. A. King, J. Elder, B. Gomolka, E. Schmidt, M. Summers, and K. Toop, "Evaluation of fourteen desktop data mining tools," in *Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on*, vol. 3. IEEE, 1998, pp. 2927–2932.

[16] J. F. Elder and D. W. Abbott, "A comparison of leading data mining tools," in *Fourth International Conference on Knowledge Discovery and Data Mining*, vol. 28, 1998.

[17] K. S. Rawat, "Comparative analysis of data mining techniques, tools and machine learning algorithms for efficient data analytics," *IOS*

[18] *R Journal of Computer Engineering*, vol. 19, no. 1, pp. 56–61, 2017.

[19] M. Gera and S. Goel, "Data mining-techniques, methods and algorithms: A review on tools and their validity," *International Journal of Computer Applications*, vol. 113, no. 18, 2015.

[20] 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.

[21] T. U. of Waikato. Weka 3: Data mining software in java. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/

[22] KNIME. Knime — open for innovation. [Online]. Available: http://www.knime.com

[23] T. U. of Waikato. Moa massive online analysis. [Online]. Available: https://moa.cms.waikato.ac.nz/

[24] T. E. Team. Elki data mining framework. [Online]. Available: https://elki-project.github.io/

[25] Shogun. Shogun machine learning. [Online]. Available: http://www.shogun-toolbox.org/

[26] M. Paluszek and S. Thomas, *MATLAB Machine Learning*. Apress, 2016.

[27] Scikit-learn. Scikit-learn: machine learning in python. [Online]. Available: http://scikit-learn.org/

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